Acta Geodaetica et Geophysica Hungarica

VOLUME 29, NUMBERS 1–2, 1994

Akadémiai Kiadó, Budapest

AGG 29 (1–2) 1–275 (1994) HU ISSN 1217–8977

ACTA GEODAETICA ET GEOPHYSICA HUNGARICA
A Quarterly Journal of the Hungarian Academy of Sciences

Acta Geodaetica et Geophysica Hungarica (AGG) publishes original reports and reviews on geodesy and geophysics in English.

AGG is published in yearly volumes of four issues by

AKADÉMIAI KIADÓ
Publishing House of the Hungarian Academy of Sciences
H-1117 Budapest, Prielle K. u. 19–35

Manuscripts and editorial correspondence should be addressed to

AGG Editorial Office
Geodetical and Geophysical Research Institute
H-9401 Sopron, P.O. Box 5
E-mail: actagg@sun10.ggki.hu

Subscription information

Orders should be addressed to

AKADÉMIAI KIADÓ
H-1519 Budapest, P.O. Box 245

Subscription price for Volume 29 (1994) in 4 issues US$ 80.00, including normal postage, airmail delivery US$ 20.00.

INSTRUCTIONS TO AUTHORS

Manuscripts should be sent to the Editorial Office (address see above). Articles not published or submitted for publication elsewhere are only accepted.

Manuscripts should be typewritten in duplicate or on disk (LATEX .tex files or Word for Windows Documents or text-only ASCII files) with a written copy, or via E-mail (LATEX .tex files or text-only ASCII files).

CONTENTS

Book reviews

Editorial

Acta Geodaetica, Geophysica et Montanistica has been published since 1966 as one of the foreign language journals of Section X of the Hungarian Academy of Sciences, and its title and content have been chosen accordingly. During the 28 years of its existence, the issues within the volumes were differently distributed among the three disciplines: geodesy, geophysics and mining, but none of these distributions proved to be satisfactory. That is why the scope of the journal has been changed from 1994 on, beginning with Volume 29. As papers on mining will be published elsewhere, the title of the journal changes for Acta Geodaetica et Geophysica Hungarica. Nevertheless, this journal is considered as the continuation of its predecessor.

A new Editorial Board was set up corresponding to the reduced field, and a new editor-in-chief was elected. As editor, however, he has acted at the journal for a long time. Both he and the technical editor are from the Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences, as before; manuscripts are to be sent to this institute and the preparation for publication is made there, too.

Eminent foreign scientists who had contacts with the journal were asked to act as members of the Advisory Board in their respective fields of interest. They are listed on the cover; on page 4, the addresses are given, too.

The publication policy remains basically unchanged, there are topics which have preference. These topics partly deal with the area of Hungary and in a wider sense with that of the Carpathian Basin. Another part covers fields in geodesy and geophysics having an important role in Hungary. In geodesy, such fields are geodynamics and robust adjustment, while in geophysics, seismicity, electromagnetic induction, geoelectricity and geomagnetic pulsations. In addition, the journal will publish material presented at symposia and conferences. National reports to scientific unions may also be included.

We invite all past and future authors who wish to publish papers in any of the fields listed above to send us manuscripts in a written form or on a floppy disk. Papers are published after peer-reviewing. We thank all colleagues who helped the journal in past years in any way and we hope that we shall gain new friends for the future volumes of Acta Geodaetica et Geophysica Hungarica.

J Verő

Adresses of the members of the Advisory Board

Acta Geod. Geoph. Hung., Vol. 29(1–2), pp. 5–18 (1994)

EFFECT OF WATER LOADING IN INDUCING SEISMICITY AROUND ASWAN RESERVOIR, EGYPT

ABD EL-MONEM S MOHAMED[1]

[Manuscript received November 16, 1993]

Aswan Lake is the second largest man-made reservoir in the world. Filling started in 1964 and an earthquake of magnitude 5.5 took place in 1981 at the Kalabsha fault area. This earthquake was followed by a tremendous number of smaller events.

The seismicity is clustered in three main zones. The most active zone is directly beneath Gebel Marawa on the Kalabsha fault, at a depth of 15 to 25 km. Activity outside the Marawa area is of shallower origin having depths of 0 to 10 km, no activity below 10 km.

Correlation was computed between seismicity and lake water level. It is found that there is a continuous decrease in the seismicity level while the water level (amount) in the High Dam Lake fluctuated. There is no obvious increase in seismicity related to the seasonal water level maxima during this time period, on the other hand, activity increases during the middle of the year, and near water level minima.

Keywords: data analysis; induced earthquakes; seismic data; water loading

Introduction

Induced seismicity is controlled at some reservoirs by short term changes in the water level in the reservoir. Increased seismicity most often occurs soon after peaks in water level or following abrupt changes in the rate of filling of reservoir.

The Aswan high dam impounds the second largest man-made reservoir in the world. Although the dam is not high by world standard (110 m), the reservoir extends over a large area and its maximum capacity is 160 km^3.

The 1981 Aswan earthquake occurred 17 years after the reservoir began to be filled in 1964. The mainshock and much of the aftershock activities are at depths of 15 to 25 km beneath the northwestern edge of the reservoir. The delay from the start of filling to the onset of seismicity and the great depths of the earthquakes are at odds with the observations at many other sites of induced seismicity. This study indicates the relationship between the seismicity and changes in water level at Aswan reservoir which strengthens the case for a causal relationship between the reservoir and the seismicity.

[1] Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences, H–9401 Sopron, POB 5, Hungary
Permanent address: National Research Institute of Astronomy and Geophysics, Helwan, Cairo, Egypt

1217-8977/94/$ 4.00

Geologic history and tectonics

Along the Nile in the lower reservoir area, from the High Dam to Garf Hussin (Fig. 1), the basement complex forms the rocks of the eastern shore, and comprises several small outcrops on the western bank. To the west of the Nile, beneath the Nubian plain, the irregular surface of the basement dips gently westward, the thickness of the overlying Nubian sandstone is gradually increasing away from the reservoir (Fig. 2). At Wadi Kurkur, a deep borehole reached this basement at about 400 m; thus the dip of the basement surface can be estimated at about 1°. The basement again outcrops at over 100 km west of the Nile (Issawi 1969), but the structure in the basement which allows this resurfacing has not been well determined. If it is due to basement faulting, this faulting must have occurred before the deposition of the post-Nubian formations which are undisturbed. Alternatively, the Late Cretaceous Nubian sands may have been deposited in an existing trough or basin in the basement. The thickness of the Nubian is small. Issawi (1969) indicated an angular unconformity between the Nubian formation and the overlying Maastrichtian through of lower Eocene marine formations. Where the Nubian formation can be seen to contact basement, it is also generally in an angular unconformity with it. At this contact, there is generally an extensive kaolinitic zone of weathered granitic rock.

The Nubian plain predominantly consists of a sandstone outcrop of the Nubia formation, except at Gebel Marawa, where there appears a remnant outcrop of the limestone plateau which has receded to the west and now forms the Sinn El-Kaddab (liars teeth) scarp. Marawa is cut by the east-west trending Kalabsha fault (Fig. 1), and is slightly folded into a syncline, elongate along the fault. The fault forms the south side of an east-trending graben. East-west normal faulting postdates north-south normal faulting; the latter does not generally cut the lower Eocene formations, and is offset by the minor strike-slip which has occurred along the E-W faults. Since the sedimentary section is merely a thin cover over the basement,

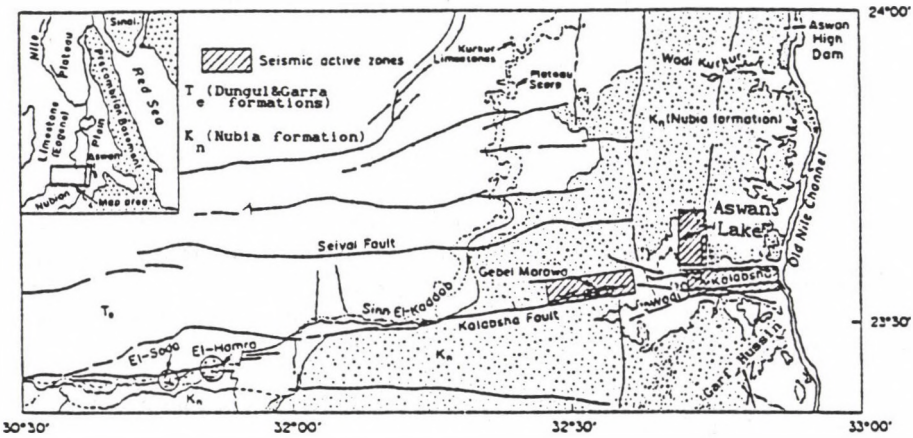

Fig. 1. Geology of the Kalabsha area and seismicity zones (after Issawi 1969)

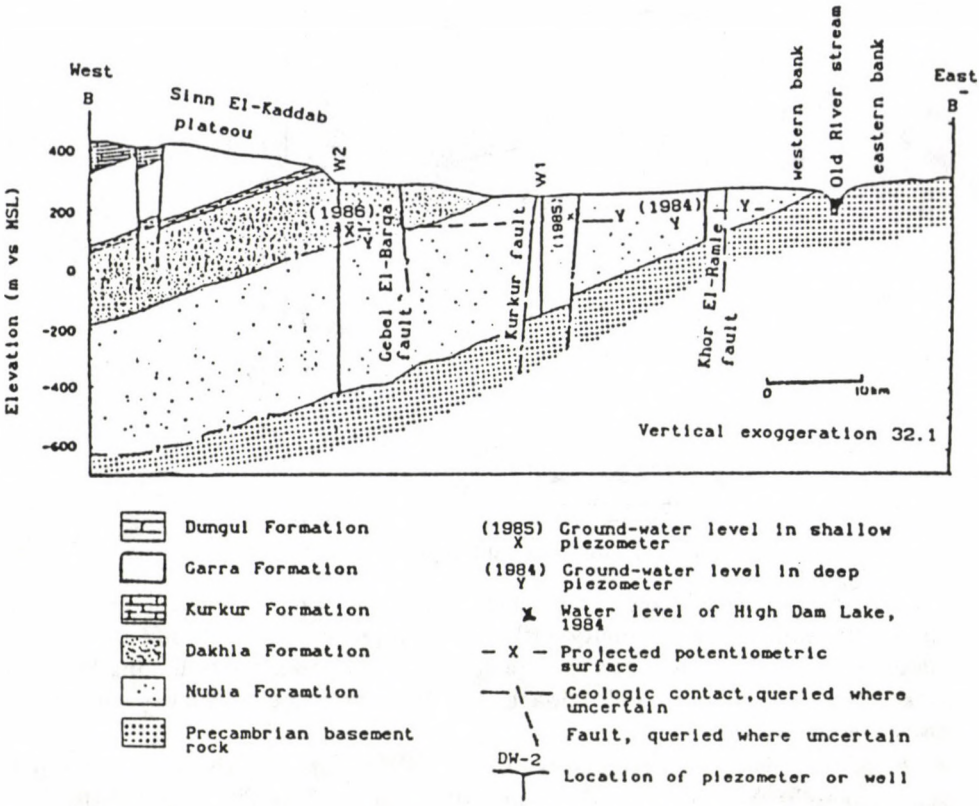

Fig. 2. E-W geologic cross-section representing sedimentary thickness deduced from drillholes (after WCC 1985)

faulting in the sedimentary section essentially reflects basement strain. The time of fault movements can only be constrained to be post-early Eocene.

The tectonic history of the area is complex and is mainly determined by faulting. Said (1962) classifies Egypt into 3 main structural units: the Arabo-Nubian massif, the stable shelf and the unstable belt. The area in and around Aswan Lake is included in Said's stable shelf (Fig. 3). The structural pattern of the area under investigation is controlled primarily by regional faulting and regional uplift of the basement rocks (Issawi 1978). The main structural features of the area are two sets of faults, one trending north-south and the other east-west (Fig. 1).

Historical and recent seismicity

From the point of view of the historical seismicity, Egypt is divided into five seismic zones. These are the coastal zone, the Gulf of Suez, the Western desert, the Northern Red sea, and the Aswan region. The historical earthquake activity within the Aswan region, (within 300 km of the Aswan and High dams) is characterized

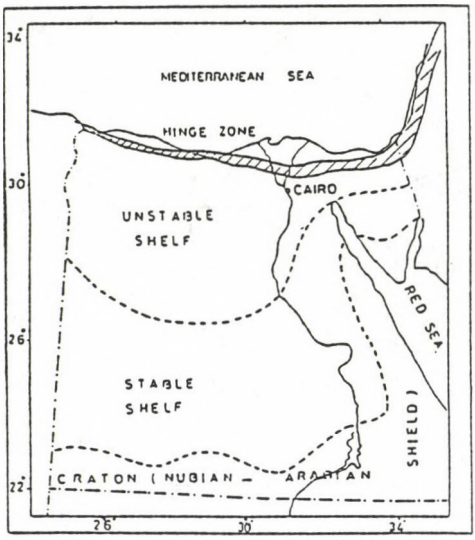

Fig. 3. Tectonic map of Egypt (after Said 1962)

by a low level of seismic activity that is revealed by long recorded history (more than 2,000 years). The seismicity within the Aswan region was found to be lower than expected due to mislocated earthquakes, distant earthquakes having been misinterpreted as local ones, and damage was attributed to earthquakes when other causes are clearly the case.

The largest historical earthquakes likely to have occurred in the Aswan region do not appear to have exceeded a magnitude (MS) of about 7.0. The largest earthquake that may have occurred along the Nile Valley is the 27 B.C. event with a magnitude estimated at 5.5 to 6.0. Somewhat larger earthquakes, including the 1481 events of magnitude approximately 6.5 appear to have occurred to the east in the vicinity of the Red Sea.

The pattern of seismicity in the Aswan region suggests that a transition zone lies between the higher level of seismicity within the Red Sea or the Red Sea border to the lower level of seismicity west of the Nile Valley in the Western desert. There have been possibly eight earthquakes within the Aswan region since about 2000 B.C., with magnitude about 5.5 and larger. Since the event in 27 B.C., these have occurred on the average every 300 years.

Prior to 14 November 1981, no earthquakes had been reported in the Aswan area in the catalog of the International Seismological Center since the ISC's inception in 1920. Because of the lack of continuous and reliable data during the early stages of the filling of the reservoir, it is not possible to determine exactly when low-magnitude activity may have started.

On 14 November 1981, a magnitude 5.5 earthquake occurred 60 km south-west of Aswan high dam, under a large embayment of the lake. This embayment is a structural depression that has been only submerged since 1976. It marks the intersection of two major sets of faults, trending approximately east-west and north-

south. The seismicity was concentrated at Gebel Marawa, near the intersection of the easterly trending Kalabsha fault with a north trending fault.

The November 1981 earthquake was unprecedented in recent history in the Aswan area. It occurred seventeen years after Aswan Lake started to fill. During these years, the lake level rose gradually more than sixty meters submerging ancient tributaries of the Nile.

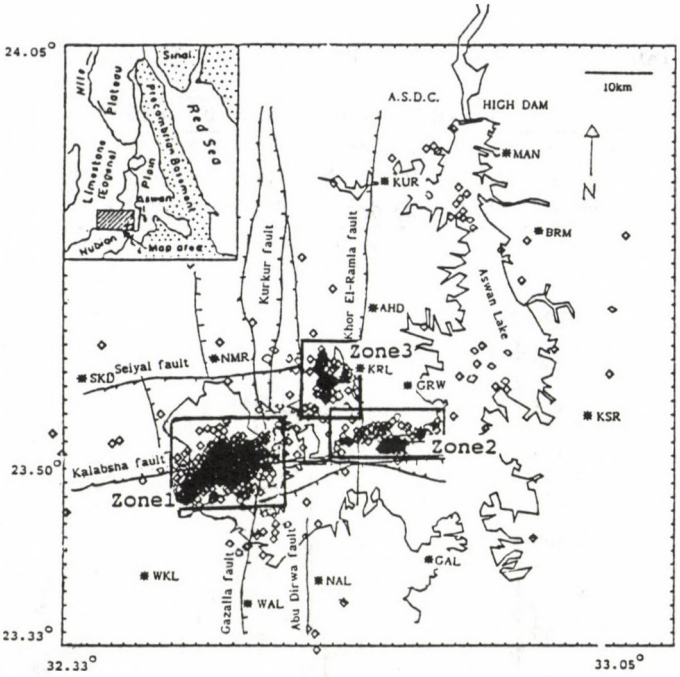

Fig. 4. Microearthquakes recorded by the Aswan seismic network from 1982 to 1993. ⊥⊥⊥ faults (taken from map by Issawi 1969), ◇ epicenters, ∗ seismic station

Seismic activity in Aswan region

In order to monitor continuously the earthquake activity around Aswan Lake and to investigate its relationship with the variation of the physical parameters which occurred in the area as the waterrose in the lake, a radio-telemetry network of 13 seismic stations (Fig. 4) was established around the northern part of the lake and has operated since 1982. The seismic events are recorded by telemetry at the Regional Seismological Center at Aswan. The purpose of this network is to investigate the relationship between the change of water level and the earthquake activity in and around the lake.

The distribution of the hypocenters of earthquakes recorded by the above telemetric network from 1982 to June 1992 has been studied. The total number of earthquakes within the entire Aswan region was about 2429 events occurred over

the entire Aswan region during the 11 years of collected data of magnitudes ranging between 0 to 4.8. However, most of them (about 96 %) are located in the area enclosed by latitudes (23.40° N, 23.80° N) and longitudes (32.50° E, 33.00° E). The final outcome of this research showed that the distribution can be presented in three adjacent active zones 1,2 and 3 whose extent is given in Table I, and whose geographical locations are shown in Fig. 4. From Fig. 4 it can be seen that the most of the active faults are in these regions.

Table I. The extent of the three active zones of the specified area

zone	latitude extent	longitude extent
zone1	23.49° 23.58° N	32.49° 32.61° E
zone2	23.54° 23.60° N	32.68° 32.85° E
zone3	23.62° 23.68° N	32.65° 32.73° E

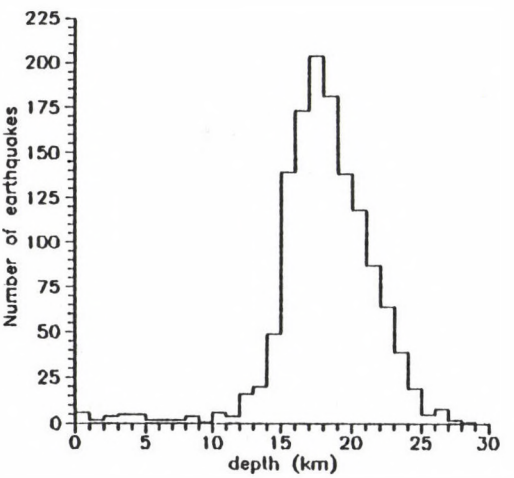

Fig. 5. Frequency of focal depths of earthquakes at zone 1 (Aswan region 1982–1993)

From an analysis of the data it can be noticed that the seismicity is concentrated at Gebel Marawa, near the intersection of the easterly trending Kalabsha fault with a north trending fault. The seismicity is clustered in three main zones (Table I):

1. Gebel Marawa,
2. east of Gebel Marawa, and
3. northeast of Marawa along the Khor El-Ramile N-S fault.

The most active zone (zone 1) is directly beneath Gebel Marawa on the Kalabsha fault, at a depth of 15 to 25 km (Fig. 5). This deeper activity is taking place where the two fault sets intersect beneath Gebel Marawa. All activity outside the Marawa area is shallower, i.e. 0–10 km, with no activity below 10 km. A second zone (zone 2) of much less activity is located farther east along the Kalabsha fault, at a depth

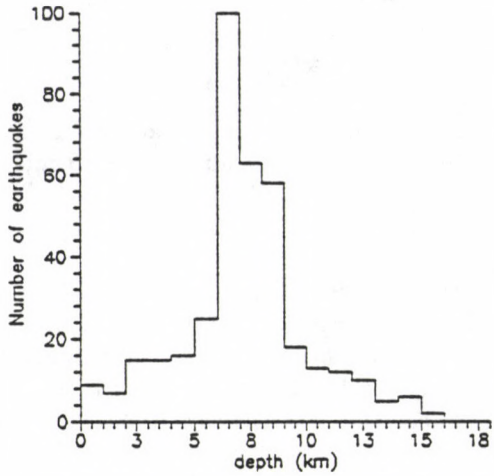

Fig. 6. Frequency of focal depths of earthquakes at zone 2 (Aswan region 1982–1993)

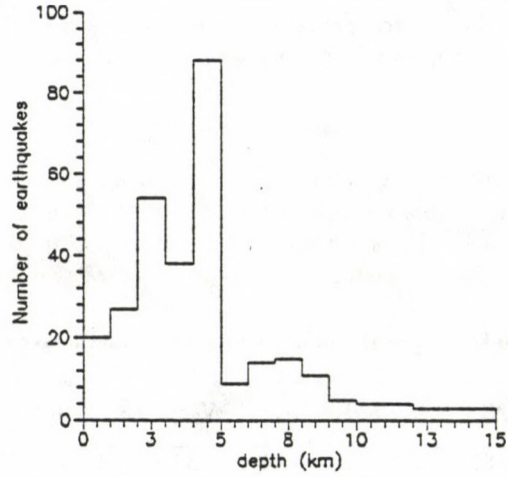

Fig. 7. Frequency of focal depths of earthquakes at zone 3 (Aswan region 1982–1993)

of 5 to 10 km (Fig. 6). The third zone (zone 3) is located north of the Kalabsha fault, and south of Wadi Kurkur, at a depth of 0 to 5 km (Fig. 7). These three seismic zones are located under a major western branch of the lake (Fig. 4). A few epicenters were located in the mainstream of the Nile between the High Dam and Wadi Kalabsha.

Changes in the seismicity level with time

The seismicity level can be defined as the density of earthquake occurrence which can be calculated as the number of earthquake occurrence during a year within a

certain area per each square kilometer of this area (Sadovsky et al. 1972)

$$\vartheta = \frac{N}{\text{year km}^2} .$$

The density estimates are computed from the above formula, and the obtained results are indicated in Table II, in order to indicate the seismic activity of the three active zones 1,2 and 3.
where

N_t = the total number of earthquakes,
N_1 = the total number of earthquakes in the first zone,
N_2 = the total number of earthquakes in the second zone,
N_3 = the total number of earthquakes in the third zone.

Figure 8 shows the change in the seismicity level for the three active zones 1, 2 and 3 during 10 years (1982–1992). Figure 8 shows that a sharp decrease of seismicity can be seen in zone 1 after the strong earthquake 1981, until 1985, and after this time the level of activity remained relatively constant. Zone 2 shows a decrease of seismicity through 1982 to 1989, except for a swarm in June 1987. From 1989 the activity has begun to increase. Zone 3 shows a decrease of the level of activity from 1984 to 1989, and it has begun to increase from 1989 to 1992.

Induced seismicity

One of the more interesting aspects of the induced seismicity at Aswan is the possible role that the Nubian sandstone plays in the control of the earthquake activity. The water level in the region where the earthquakes are occurring is less than 10 m deep. The sandstone surrounding the reservoir, however, is highly

Table II. The earthquake density for the three active zones

year (month)	N_t	N_1	$\vartheta \times 10^{-2}$	N_2	$\vartheta \times 10^{-2}$	N_3	$\vartheta \times 10^{-2}$
1981(1)	83	38	254.40	3	32.30	7	104.10
1982(12)	1155	878	487.80	99	88.40	46	56.80
1983(12)	326	171	95.00	55	49.10	37	45.70
1984(12)	241	70	38.90	54	48.20	64	79.00
1985(9)	81	26	19.30	11	13.10	22	36.20
1986(12)	78	26	14.40	17	15.20	19	23.50
1987(12)	168	21	11.70	118	105.40	13	16.00
1988(12)	71	15	8.30	8	7.14	17	21.00
1989(12)	55	18	10.00	3	2.68	10	12.30
1990(12)	76	22	12.20	8	7.14	38	46.90
1991(12)	69	15	8.30	15	13.40	30	37.00
1992(5)	26	8	10.60	3	6.40	6	17.60

porous (25 %) and relatively permeable compared to the underlying granite. As the reservoir fills, the expanding area of the reservoir and the rising water level are allowing water to seep into the sandstone, thus raising the local water table. Because of the impermeable basement, the water is confined to the westward thickening sandstone lens.

The increased load of the reservoir thus consists not only of the water within the reservoir, but also of a significant amount of water stored in the sandstone. The increased pressure at the base of the sandstone results from the combined influence of both the water in the reservoir and the increased water table in the sandstone. The time required for the water to diffuse into the sandstone may explain the delay between the filling of the reservoir and the start of the seismic activity.

Figure 9 shows water levels at Aswan Lake for the period 1982 through 1992, 10 days changes in the lake level and earthquakes of magnitude ≥ 1. Figure 9 shows a gradual decrease of the average seismicity with exception of some spikes in microearthquakes that follow rapid rate of seasonal discharges indicating that there is some correlation between seismicity and change in water level. It is not, however, clear whether the decrease in seismicity can be attributed to gradual decrease in the lake water level, or to the gradual stabilization of the area after the start of seismicity in 1981.

Figures 10, 11 and 12 show the relation between the average changes of the water level and totals of earthquake frequency for the same period. The data for 11 rainy seasons from 1982 to 1992 provide 11 such examples, where the rainy season extends from August to November in every year.

Figure 10 shows that there is no obvious increase in seismicity related to the seasonal maxima in water level over this time period. A careful examination of Figs 11 and 12 indicate that every year, following the rainy season, the activity slightly

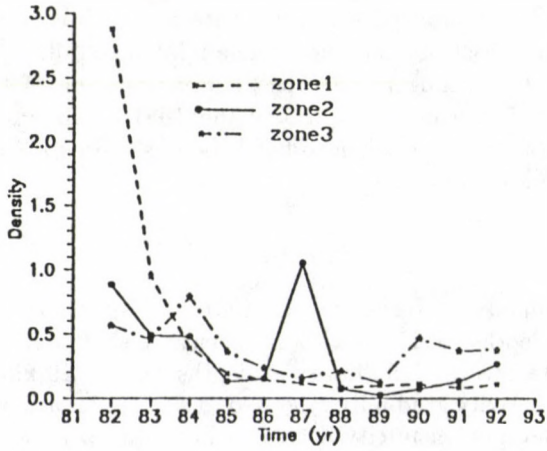

Fig. 8. Secular changes in earthquakes density as defined by earthquake numbers per km^2 year in three zones (Aswan region 1982–1992)

increases. The activity increases during the middle of the year (Fig. 11), and it reaches maximum near the time of local minima in water level (Fig. 12).

Comparison with other reservoirs

At a number of reservoir sites there appears to be a close association of time of increased seismicity with rapid changes in the water level of the reservoir. One of the best documented examples is the Nurek reservoir, where most of the larger earthquakes and swarms of increased activity follow closely rapid decreases in the rate of filling (Keith et al. 1982). A second example is lake Oroville in California where draw-downs of 3 to 5 m resulted in renewed induced seismicity (Toppozada and Morrison 1982) At Koyna (Gupta 1983) the seismicity appears to be triggered if the rate of filling exceeds 12 m/week.

Simpson et al. (1988) include cases in which there is a correlation of increased seismicity with water level change in a "rapid response" classification of induced seismicity. In addition to Nurek, other reservoirs that have shown this rapid response include Koyna (Gupta 1983), Monticello (Lablance and Anglin 1978, Talwani and Acree 1988). Of these, the relationship between changes in water level change and seismicity is most obvious at Nurek and Koyna.

One may separate two classes of reservoirs from the general population of reservoirs that exhibit induced seismicity, depending on their behavior after the initial filling: one that gives rise to seismicity during down-draw, and another that gives rise to it on refilling.

Short-term temporal changes in induced seismicity are suggested to be related to spatial inhomogeneities in rock properties. The latter produce transient and localized increase in pore pressure following rapid changes in surface load. This mechanism may explain the correlation observed at Aswan between the annual peaks in water level and seismicity.

A delayed response in induced seismicity, with a long time between first filling of the reservoir and the onset of seismicity can be related to diffusion of pore pressure from the reservoir to seismogenic depths. An explanation of the delay between the start of filling of Aswan reservoir and the 1981 earthquake can invoke this diffusion process plus the time taken to flood the unsaturated Nubian sandstone in the seismicity active area.

Conclusions

The main earthquake on November 14, 1981, and aftershocks, were in the deep zone (zone 1) at depths of 15 to 25 km beneath Gebel Marawa. All activities outside the Marawa area are shallower, at depths of 0 to 10 km, with no activity below 10 km. The activity gradually migrated to east and northeastward (shallower zones), while the level of seismicity decreased in the deep zone. This seismicity is concentrated along the edge of the Kalabsha embayment.

The reservoir behind the Aswan High Dam gives rise to seismicity during down-draw. Every year following the rainy season, the activity slightly increases, and it

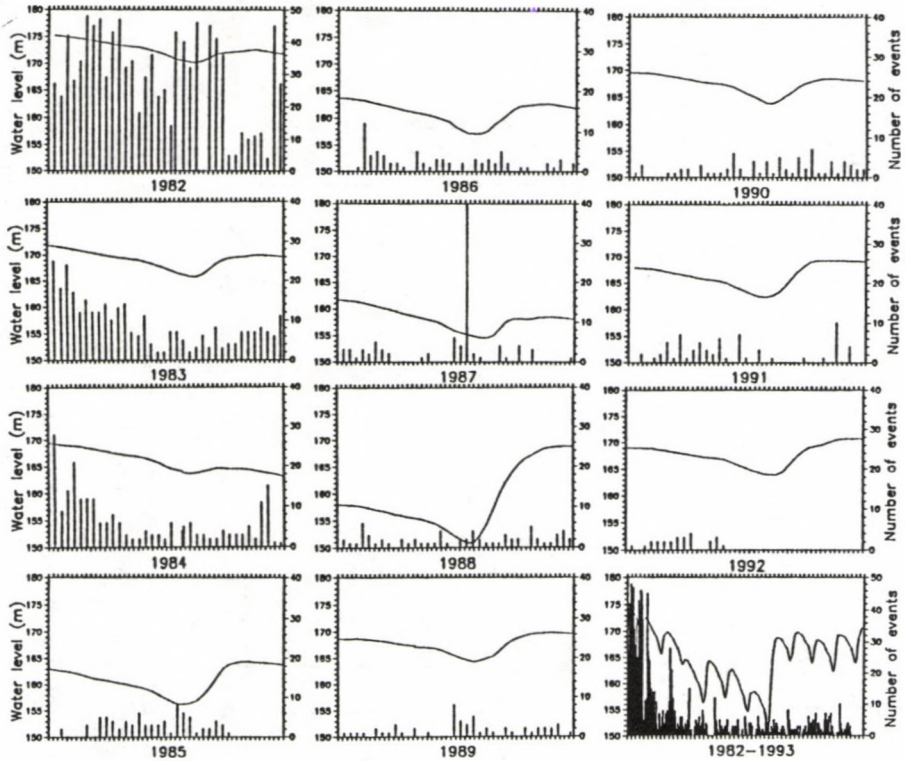

Fig. 9. Water level in the reservoir and number of earthquakes occurred within the considered area for the period 1982 to 1993

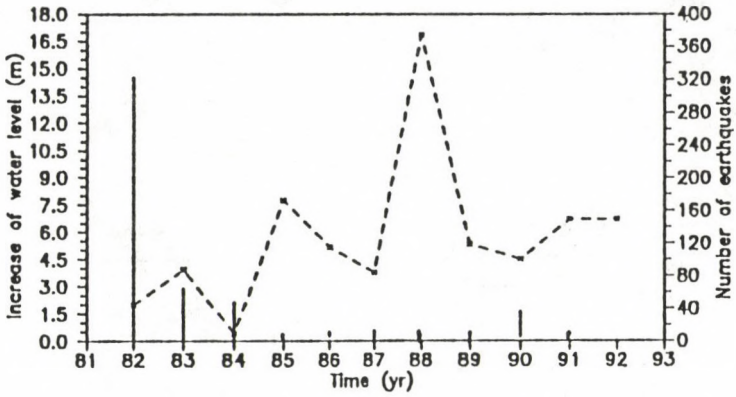

Fig. 10. Relation between the average increase of water level (dashed line) and totals of earthquake frequency (vertical bar) for the same months at Aswan Lake (4 months/year, Aug., Sept., Oct. and Nov.) from 1982–1992

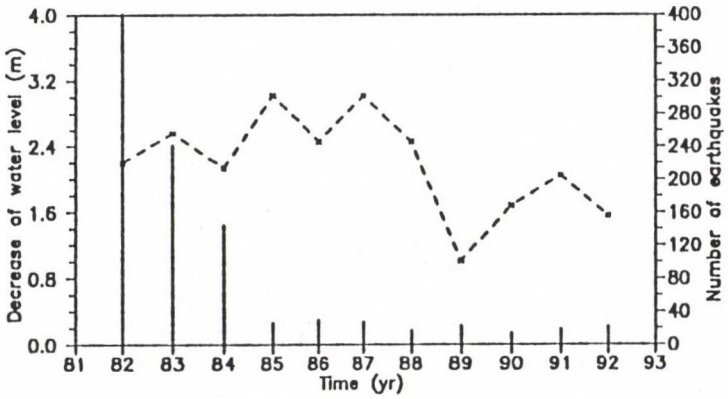

Fig. 11. Relation between the average decrease of water level (dashed line) and totals of earthquake frequency (vertical bar) for the same months at Aswan Lake (4 months/year, Dec., Jan., Feb. and Mar.) from 1982–1992

reaches maximum near the time of local minima in water level.

From the study of the seismicity in the Aswan area, it was found that a continuous decrease in the seismicity level happened between 1982 to 1992, although the water level (amount) in Aswan lake fluctuated during the same period. This means that the gradual stabilization of the area has begun after the start up of seismicity in 1981. Also, the water loading is only one from several factors as an activating medium in triggering earthquakes. The common factors for all cases of induced seismicity seem to be the presence of specific geological conditions, the tectonic setting and water loading.

Any relationship between the water level and seismicity at Aswan is complicated by the details of the interaction between the reservoir and the regional groundwater

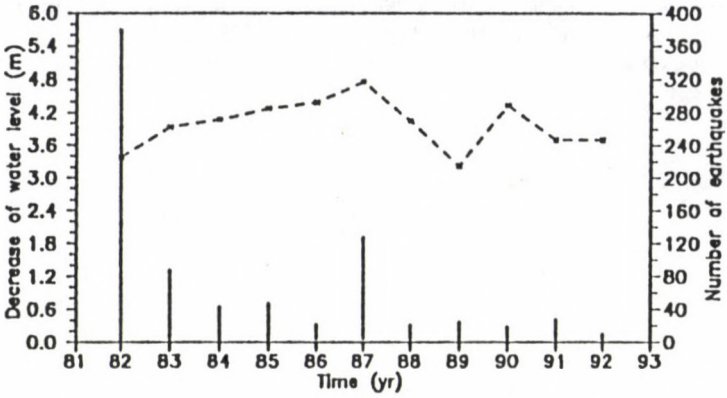

Fig. 12. Relation between the average increasing of water level (dashed line) and totals of earthquake frequency (vertical bar) for the same months at Aswan Lake (4 months/year, Apr., May, Jun. and July) from 1982–1992

surface. Lateral variations in the extent of the reservoir are large in the area where the seismicity is occurring, with the reservoir first entering the active area 6 years before the 1981 earthquake and then completely receding during the recent drought. The process which controls the relationship between the seismicity and the water level is obviously complex and not deterministic. There are, however, sufficient examples of increased seismicity closely associated in time with conspicuous changes in water level to add support to a causal relationship between the reservoir and the Aswan earthquakes. Any increase in the level of seismicity as the reservoir reenters the Marawa area in future years will help to refine our understanding of the mechanism by which the seismicity is related to the reservoir.

Acknowledgment

The author would like to express his gratitude to Dr. Gy Szeidovitz, head of Seismological department, Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences, for his continuous help, for reading the manuscript.

Rerefences

Gupta H K 1983: *Bull. Seismol. Soc. Am.*, 73, 679–682.
Issawi B 1969: *Geological Survey*, 46, 101 pp.
Issawi B 1978: *Annales of the Geological Survey of Egypt*, B, 237–253.
Keith C M, Simpson D W, Soboleva O V 1982: *J. Geoph. Res.*, 17, 4609–4624.
Lablance G, Anglin F 1978: *Bull. Seismol. Soc. Am.*, 68, 1469–1485.
Sadovsky M A, Nersesov I L, Semenov A N 1972: *Tectonophysics*, 14, 295–308.
Said R 1962: The geology of Egypt. Elsevier, Amsterdam, New York
Simpson D W, Leith W S, Scholz C H 1988: *Bull. Seismol. Soc. Am.*, 78, 2024–2040.
Talwani P, Acree S 1988: *Pageoph*, 122, 947–965.
Toppozada T R, Morrison P W 1982: *Calif. Geol.*, 35, 115–118.

WCC (Woodward-Clyde Consultants) 1985: Earthquake activity and Dam stability evaluations for the Aswan High Dam, Egypt. For the High and Aswan Dams Authority, Ministry of Irrigation, Arab Republic of Egypt, Woodward-Clyde Consultants, Walnut Creek, CA (internal report).

Acta Geod. Geoph. Hung., Vol. 29(1–2), pp. 19–38 (1994)

GEOPHYSICAL STUDIES OF THE ANTARCTIC PENINSULA

Procedures and criticism of their use to obtain an acceptable synthesis of results coming from six campaigns on the Antarctic Peninsula using four geophysical methods: magnetotelluric, audiomagnetotelluric, electric and seismic soundings, from 1979 till 1992

H G FOURNIER[1]

[Manuscript received December 6, 1993]

With the results of a total of 15 soundings made in an area of 200 km diameter by 4 methods (9 MTS, 2 VES, 2 AMTS, 2 SS) at the NE end of the Antarctic Peninsula, in a synthesis, we follow: 1) the permafrost; 2) the brine layer below the permafrost; 3) the contact between the Lower and Upper Cretaceous; 4) the sedimentary basement; 5) the top of the intermediate conductive layer (ICL).

Keywords: Antarctic Peninsula; geoelectric sounding; magnetotellurics; permafrost; seismics

Introduction

Figure 1 shows the region studied at the North-East extremity of the Antarctic Peninsula. The map of Fig. 2 gives more details of this region and also the sounding sites: Seymour and James Ross Islands in the northeastern part (Section I), Robertson, Larsen, Pedersen and Fair Weather Island, on the Barrier of Larsen, in the southwestern part (Section II). The distance between the two Sections I and II is about 200 km. Table I gives a detailed list and content of the 6 campaigns made from 1979 till 1992. A general presentation of the geology is given by Grikurov (1978).

Synthesis

Section I.1 Seymour Island: the two MTS and the two VES were made at the Argentine Marambio Scientific Air Base situated in this island. There is no glacier ice. Recent geological developments are given by Del Valle et al. (1983a) and by Leguizamón and Mamani (1993). For Seymour Island iterative soundings were used to find an acceptable coherent section; it will be the base to continue this synthesis. Consequently, we are obliged to conduct this study carefully, with a maximum of field data. By chance, there are many results permitting us to give a detailed description of the Marambio soundings sections.

Field results are obtained:

[1]CRICYT-IIACE-GEOFISICA, Casilla de Correo 131, 5500 Mendoza, Argentina

1217-8977/94/$ 4.00

A) by the MTS method: Fig. 3 shows the MT sounding curves obtained: ALBK for MA1 in 1979 (Fournier et al. 1980) and CMD for MA2 in 1980 (Del Valle et al. 1983b, 1988). The two curves are parallel with a small shifts between them. The two sites were chosen at a distance of 2 km to avoid a hypothetic heterogeneity of local first sedimentary layers. The two teams of geophysicists were completely different and also the equipment, with magnetic variometers (core coiled bars) built using different techniques. We think that the fact of the quasi equality of the two sounding results is very important for the validity of the following synthesis because all the comparisons will be based on the Seymour Island results.

B) by the VES method: Figs. 4 and 5 give the results obtained: two vertical electrical sounding curves with the values of AB/2 between 0.003 and 0.4 km. The VES sites were the same as for the MTS. We remark that the upper parts of the permafrost resistivities are very different for the two sounding sections (Fournier et al. 1990).

What have we done with these results to obtain an acceptable complete section for Seymour Island?

The basic idea is to replace the VES curves by corresponding MT curves. The purpose of this replacement is to obtain a unique curve, e.g. for MA1: EJB, and for MA2: GHB (Benderitter et al. 1978). To obtain this replacement we have done the following: 1. we determined the corresponding VES layer sections by a programme given by the Garchy MT base in France; 2. we calculated with these sections of layers the corresponding MT curves using the formula proposed by Fournier et al. (1963). We have obtained for the "VES transform in MTS" the curves EQ for MA1 and GP for MA2. There is no overlapping interval between the two kinds of curves. It is to be investigated if the nature of the ohmm values used in the two kinds of soundings is the same (VES and MTS) i.e. it permits us to connect the two experimental curves (the transformed VES and MTS) by an adjusting calculated curve. On one hand, it is well known that the earth currents are flowing horizontally

Table I. Studied zones, during fourteen years, on the NE end of the Antarctic Peninsula, with geophysical methods operating on the soil, on the ice and on the Larsen Ice Barrier

Campaign year	Region of the site	Sounding method	Number and name of the soundings	
1979	Seymour Island	MTS	1	MA1
1980	Seymour Island	MTS	1	MA2
	Robertson Island	MTS	1	RO1
1981	Larsen Barrier	MTS	3	LA1, PE1, FA1
1987/1988	Seymour Island	VES	2	MA1, MA2
1987/1988	James Ross Island	SS	1	SBR1
1992	James Ross Island	AMTS	2	BR1, BR2
	James Ross Island	MTS	3	BR1, BR2, HI1

MTS: Magnetotelluric Sounding, VES: Vertical Electrical Sounding,
SS: Seismic Sounding, AMTS: Audio Magnetotelluric Sounding

in the earth, on the other hand man-made electric currents of the VES are crossing the earth partly with a vertical component. Because of this, the electric current crosses and contorts the horizontal micro-resistive particles, as compressed by the terrain above, an elevation of the resistivity is generated in this way (Maillet and Doll 1933; Cagniard 1948).

We know similar cases presented in the literature. We present first a most similar case, published by Fournier et al. (1986). We reproduce in Fig. 3 this case of adjustment. But in this study, the permafrost is a valley covered glacier (fossil glacier) situated near Vallecitos on the Argentine side of the Andes Chain, Province of Mendoza. The SXT curve is the resulting MT curve of the transformation of the VES curve (AB/2 from 0.003 to 0.3 km). The UV curve is the AMTS curve obtained on the same site. We see no overlapping interval but an adjustment is permitted with a small shift of the joining calculated curve SXUV near point T. (See the Vallecitos sounding section in Table II.) The topographic conditions of the site were very uncomfortable due to a very strong and irregular slope of the glacier. This was not in favour of a general harmony of the results. In conclusion, Table II shows us that the successive layers are very similar for the two studied cases Marambio and Vallecitos, particularly for the conductive layer below the permafrost. There is no overlapping interval of periods, but only an adjustment from one curve to the other. The XT curve is a deviation generated by the effect of the narrowness of the shoulders of the valley, affecting the distribution of the electric currents of the VES emission in the earth, lateral rocks having a great resistivity. Figure 3 shows that only a combination of adjustment layers for MA1 and MA2 is possible

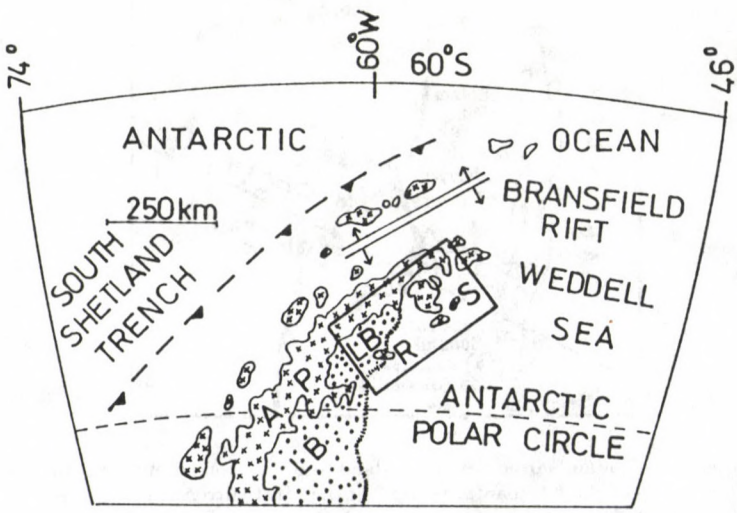

Fig. 1. Position of the studied region, a rectangle, in the north end of the Antarctic Peninsula (after a map of the "Dirección Nacional del Antártico", Buenos Aires, and after Dalziel 1983). AP: Antarctic Peninsula. LB: Larsen Barrier; R: Robertson Island; S: Seymour Island

that obey at point C the law of continuity for the resulting MT curve (Fournier et al. 1987). We see that the combination of layers was chosen so to cross point C (Fig. 3) continuously with curve CMD. There is a small interval of variants without important change in the layer combination to obtain the joining of the two curves of different origin. These connections QJC and PHC are made intercalating a brine layer (0.018 km at 0.6 ohmm) (McGinnis et al. 1973). A final adjustment layer

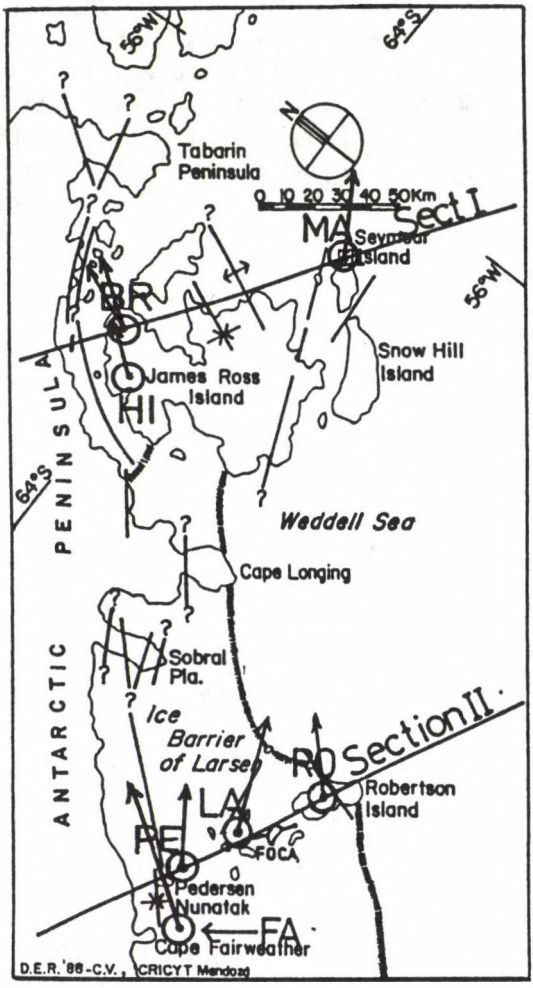

Fig. 2. Studied region, mainly from Seymour Island to Robertson Island, with the general tectonic features (after a map of the " Geoantar Group" of the "Instituto Antártico Argentino", and after Del Valle et al. 1982). Localization of the sites: MA (Marambio Base on Seymour Island) for MA1 and MA2 soundings; BR (Brandy sites on James Ross Island) for BR1 and BR2; HI (Hidden Lake site on James Ross Island) for HI1; RO (on the center of Robertson Island) for RO1; LA (near Larsen nunatak) for LA1; PE (near Pedersen nunatak) for PE1 and FA (near Cape Fair Wether) for FA1. Arrows give the strike direction (tectonic angle) See Table IV

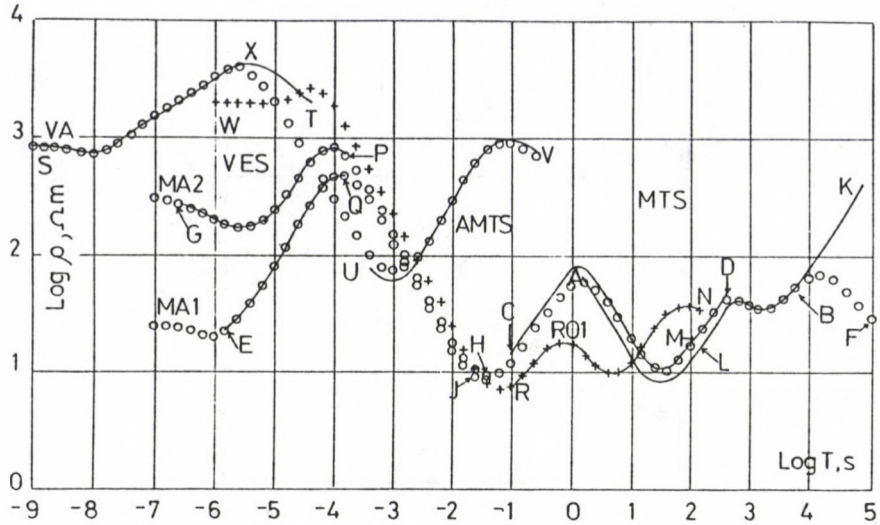

Fig. 3. Tensorial interpretative unidimensional sounding curves

for Seymour Island
- for MA1 full line
  - from VES origin : EQ
  - from MTS origin : ALBK
  - A)
- for MA2 full line
  - from VES origin : GP
  - from MTS origin : CMD
  - B)

A): EQJALBF is the calculated unidimensional theoretical curve for MA1 dots curve.

B): GPHCMD is the same for MA2 dots curve.
For Robertson Island (RO1): from MTS origin: RN full line; WRN is the calculated unidimensional theoretical curve for RO1 dots curve.
For Vallecitos (VA): from VES origin: full line SXT.
For Vallecitos (VA): from AMTS origin: full line UV.
SXUV is the whole corresponding calculated curve for (VA) full line.

Remark: we see a great difference between the two curves BK and BF: BK is considered erroneous by one electrode effect at very long periods (a very small difference in temperature between the two electrodes). BF is an approximate trace to avoid the eventual sandwich or mutual effect (Duhau et al. 1988).

follows having a thickness of 4.25 km at 260 ohmm, that is the first layer of the MT curve CMD.

Now, we shall see a case of linkage for terrain in normal conditions (no permafrost). Figure 6 shows the curve "MTS from VES-MTS": Tittarelli site, with 5 km of sediments, near Mendoza, Argentine. There is a perfect overlapping CB of the two curves AB and CD between 0.002 s (50 m depth) and 0.013 s (150 m depth). AB is given by the electric sounding and CD by the magnetotelluric sounding.

In conclusion, we think that it is allowed us to connect the two kinds of curves from the MA1 and MA2 sites on Seymour Island: MTS from VES with MTS as we did.

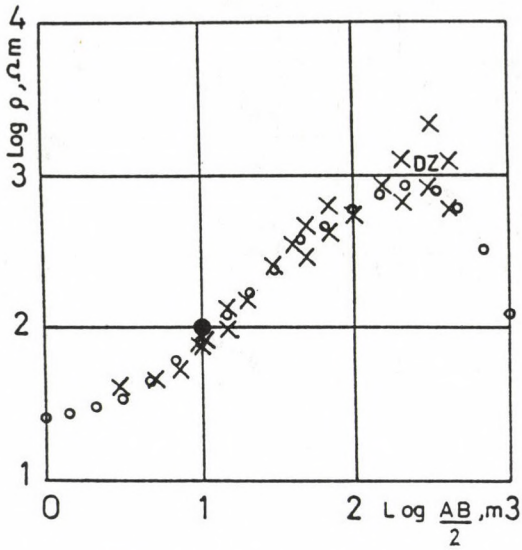

Fig. 4. VES results for MA1 sounding of Seymour Island — the X-es are the measuring values for AB/2 from 0.003 to 0.3 km, the great circle is the result obtained in 1979 (Fournier et al. 1980), with DZ a dispersion zone, because the measuring apparatus is at its sensibility limit. The circles give the calculated corresponding MT curve used to adjust the MTS results

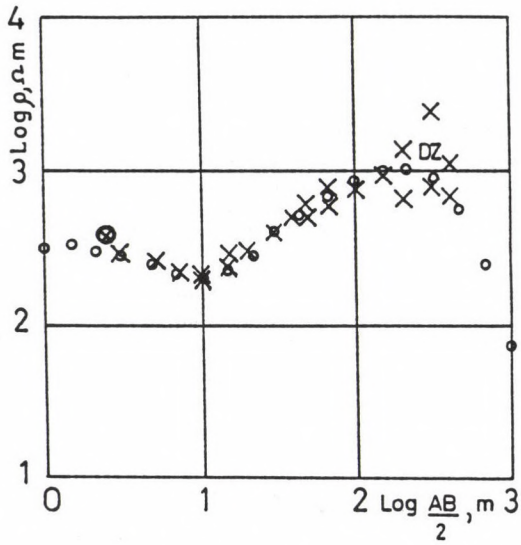

Fig. 5. The same as Fig. 4 for the MA2 sounding of Seymour Island

What is a brine layer? The formation of permafrost is viewed as a refining process in which the salts are segregated to the bottom of the frozen layers during freezing (Baskov and Zaytsev 1973, Ginsburg and Neizvestnov 1973, Larin et al. 1973, Yasko 1973). We also expect brine concentration due to the ion migration in

the frozen state (Anasimova 1978). Such a conductive zone has been observed by others, such as Osterkamp using well logs in Northern Alaska. Consequently, we expect brine concentration at the interface between the permafrost and unfrozen rocks.

Now, in the progressive search for the EJB and GHB unidimensional theoretical curves in the interval CD, we have used the two parallel experimental MT curves ALB and CMD as the limit of a zone of error-bars: we have tried to stay between these two curves. The layer sections of the two sites MA1 and MA2 are the same with the exception of the permafrost layers (Fournier et al. 1990). Figure 3 shows the resulting MT curves for the two sites, situated at a distance of 2 km from each other: for MA1: EQJBF and for MA2: GPHBF. The MA1 sounding was made in a different interval of periods, thus it gives a complete definition of the ICL between

Table II. Detailed layers content for the MT soundings of the Sections I and II

SECTION I					
Seymour Island (MA1) 9 layers			Seymour Island (MA2) 10 layers		
	Thickness in km	Resistivity in ohmm		Thickness in km	Resistivity in ohmm
			SC	0.0020	320.
SC	0.0028	25.	NC	0.01	165.
NC	0.130	2000.	NC	0.140	2000.
NC	0.018	0.6	NC	0.018	0.6
NC	4.25	160.	NC	4.25	160.
NC	1.5	1.7	NC	1.5	1.7
NC	72.	2000.	NC	72.	2000.
NC	23.	7.	NC	23.	7.
SC	400.	2000.	SC	400.	2000.
SC	Infinite	2.	SC	Infinite	2.
James Ross Island (BR1) 10 layers			James Ross Island (BR2) 10 layers		
SC	0.070	2000.	SC	0.070	2000.
NC	0.0105	0.6	NC	0.0105	0.6
NC	0.35	200.	NC	0.35	200.
NC	0.45	3.5	NC	0.45	3.5
NC	6.6	200.	NC	6.6	200.
NC	0.5	4.	NC	0.5	3.
NC	70.	2000.	NC	70.	2000.
SC	20.	10.	SC	20.	10.
SC	408.	2000.	SC	408.	2000.
SC	Infinite	2.	SC	Infinite	2.
James Ross Island (HI1) 6 layers			Vallecitos (VA) Andes Chain Arg. 5 layers		
SC	0.10	2000.	active layer	0.0015	800 SC
NC	0.09	0.38	permafrost	0.070	4100 NC
NC	?	?	subpermafr.	0.063	31 NC
NC	72.	2000.	basement lay.	3.8	4800 NC
SC	20.	10.		Infinite	75 SC
SC	410.	2000.			
SC	Infinite	2.			

Table II. (contd.)

SECTION II					
Robertson Island (RO1) 8 layers			near Larsen Nunatak (LA1) 6 layers		
	Thickness in km	Resistivity in ohmm		Thickness in km	Resistivity in ohmm
SC	0.165	2000.	SC	0.135	2000.
SC	0.026	0.5	NC	0.210	0.25
NC	1.7	200.	NC	42.	2000.
NC	0.5	1.6	SC	15.	7.
NC	27.	2000.	SC	450.	2000.
SC	7.	7.	SC	Infinite	2.
SC	450.	2000.			
SC	Infinite	2.			
near Pedersen Nunatak (PE1) 8 layers			near Cape Fair Weather (FA1) 8 layers		
SC	0.105	2000.	SC	0.270	2000.
SC	0.0148	0.25	NC	0.042	0.25
NC	0.84	200.	NC	1.3	200.
NC	1.7	2.	NC	0.9	2.
NC	24.	2000.	NC	42.	2000.
SC	10.	3.	SC	7.	7.
SC	460.	2000.	SC	450.	2000.
SC	Infinite	2.	SC	Infinite	2.

SC: Supposed Content
NC: Normal Content

78 and 100 km depth, with 7 ohmm resistivity. This is very important because it is a unique case of complete ICL determination from among all the MT soundings of this synthesis. The experimental section of curve BK is considered erroneous. Table II shows that the section of these two MT soundings begin with a permafrost layer of about 0.130 and 0.140 km thickness having a resistivity up to 2000 ohmm; then, a brine layer of 0.018 km at 0.6 ohmm, a resistive layer of 4.25 km at 160 ohmm (Tertiary + Upper Cretaceous), a 1.5 km layer at 1.7 ohmm (Lower Cretaceous + Upper Jurassic), a very resistive layer of 72 km at 2000 ohmm (partly crust and partly upper mantle), an ICL of 23 km thickness at 7 ohmm (MA1 only); then two added layers to avoid the eventual "sandwich or mutual effect" (Duhau et al. 1988): a 400 km layer at 2000 ohmm (upper mantle) and the top of the ultimate conductive layer (UCL) at 500 km depth — the UCL having 2 ohmm resistivity. The characteristics of these two final layers were chosen in the range generally accepted for the Earth (Berdichevsky et al. 1976). We add here that the cited Upper-Lower Cretaceous contact is discussed by Del Valle and Fourcade (1986), Del Valle et al. (1988), Fournier et al. (1989, 1992). This contact is supposed to be the contact between the resistive and the conductive layers constituting the sedimentary lower part of the studied basin. Consequently, it should be possible to follow it in the region of the Antarctic Peninsula by the MTS method if this supposed MT contact is realistic. Figure 7 shows the sections of the Marambio MTS: MA1 and MA2.

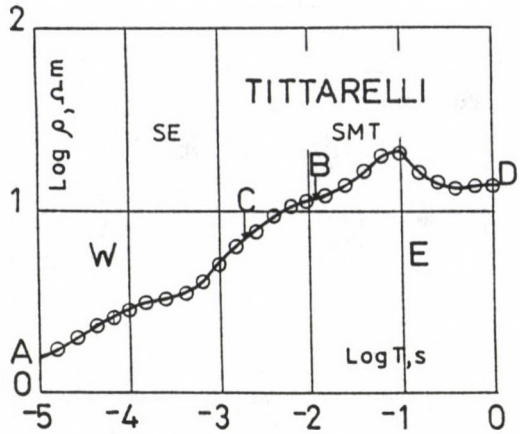

Fig. 6. For Planchez site: full line JK: Experimental curve from the VES origin; fill line LM: Experimental curve from the MTS origin; JLM: corresponding calculated MT curve for the site. For Tittarelli site: full line AB: Experimental curve from the VES origin; full line CD: Experimental curve from the MTS origin. CB is the interval presenting a perfect overlapping of the two curves of the two origins; ACBD is the corresponding calculated MT curve for the site

The Marambio MT soundings were studied by Pomposiello et al. (1988). It was shown that the tensorial unidimensional interpretative curve was obtained for an angle of NG 57°E, Rhomin, direction parallel to the general trend axis of the island and also parallel to the Peninsula Coast. We have made an attempt for a bidimensional interpretation of the results: it appears that it is a combination of the topographic profile and of the sea water depths, briefly the island effect that affects the TM curve and pushes it above the TE curve. The theoretical curve is very near to the measured curve. Below the site, the sedimentary layers should be mostly regular (with a great wave length for the layer's ondulations) (2D inversion programme by courtesy of Wannamaker).

The resistivity of the permafrost of Seymour Island has been determined with the combination of the two kinds of soundings: VES and MTS, about 2000 ohmm, or more. To simplify, we shall use systematically this value for the resistivity of the permafrost of the two other islands, James Ross and Robertson, for glacier ice cover of Robertson Island and also for the basement layers in the crust and for the upper mantle layers. It does not change anything if we use values higher than 2000 ohmm. We shall not research each time the exact minimum value of the high resistivities as it appears in the layer sections.

Section I.21 James Ross Island: Here, we must give a general remark concerning the iterative method that we shall use to complete soundings by addition of layers by analogy of the results obtained in previous soundings.

We know that in the MT method the first layer is not completely defined (only its integrated conductivity — or resistivity —), but there exists a maximum thickness that must not be exceeded by the addition of a combination for the first layers by analogy (Fournier et al. 1987). In Fig. 7 the layers in the columns between the two

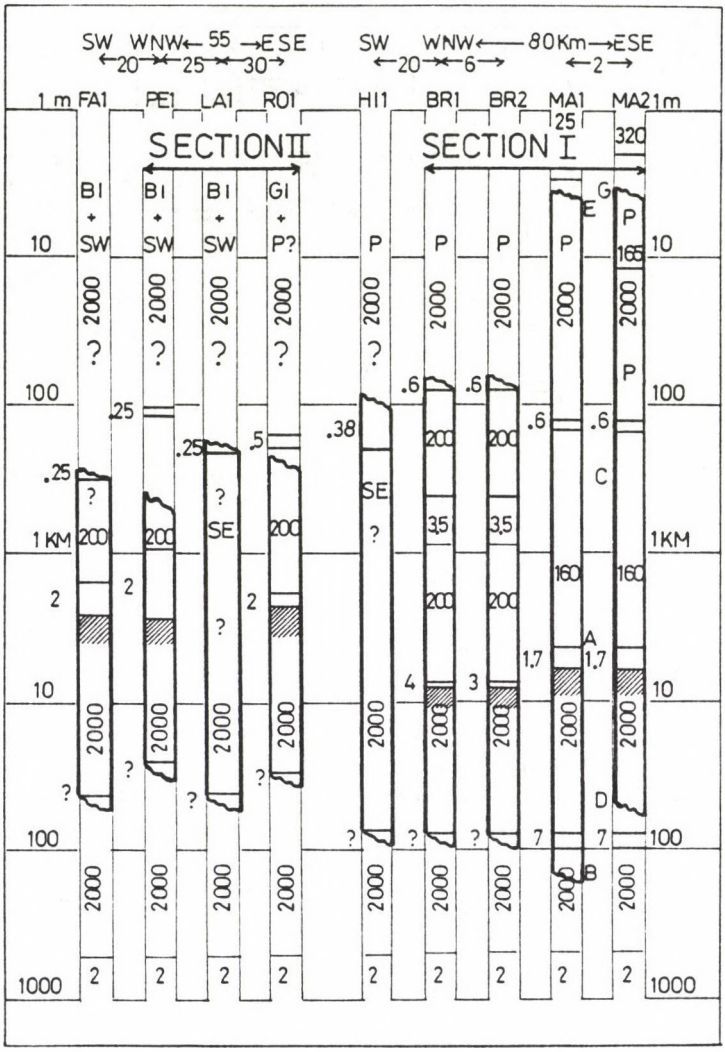

Fig. 7. Graphic representation in log scale, for the MT results of the soundings of the Section I + HI1 and the Section II + FA1. See the position of the soundings sites in Fig. 2. MA1 and MA2 on MA site — Marambio Base on Seymour Island. BR1 and BR2 on BR site — Brandy sites on James Ross Island. HI1 on HI site — Hidden Lake site on James Ross Island. RO1 on RO site — on the center of Robertson Island. LA1 on LA site — near the Larsen nunatak, on the Larsen Barrier. PE1 on PE site — near the Pedersen nunatak, on the Larsen Barrier. FA1 on FA site — near the Cape Fair Weather, on the Larsen Barrier. BI: Barrier Ice, GI: Glacier Ice, SC: supposed content, SW: Sea Water, SE: Screen Effect, NC: normal content, P: Permafrost

reinforced lines constitute the true, observed, section in the soundings. The other layers are incorporated by analogy with the results obtained in previous soundings. These supplements have a high probability to be realistic as discussed later.

Section I.22 James Ross Island — Brandy sites: for the geology see Del Valle

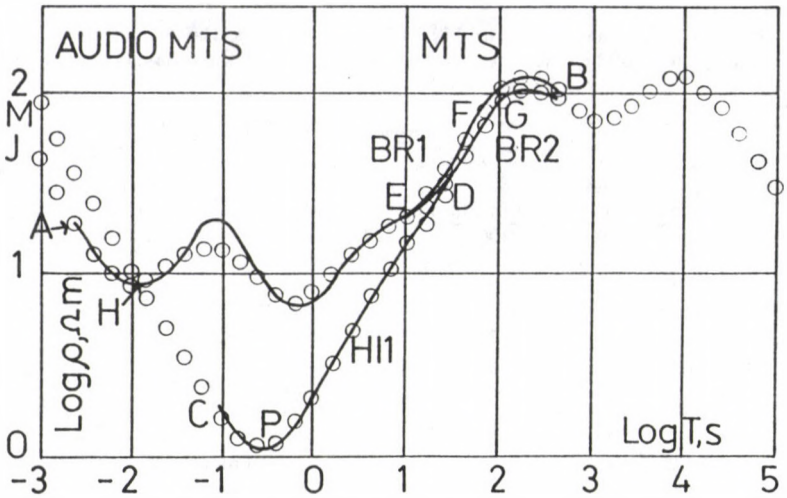

Fig. 8. MT theoretical unidimensional sounding curves for James Ross Island sites: for BR1: experimental results: full line AHEFB, for BR2: experimental results: full line AHEDGB, for HI1: experimental results: full line CDGB. The circles give the calculated corresponding curves

et al. (1982), Del Valle and Fourcade (1986), Rinaldy (1992) and Leguizamón and Mamani (1993), in Fig. 2 the position BR of the two MT sounding sites BR1 and BR2, situated 6 km from each other. There is no glacier ice on these two sites. A detailed description of the study is given by Mamani et al. (1993).

For the first layers, BR1 and BR2 give nearly the same section (only the difference: 0.5 km at 4 ohmm for BR1 and 0.5 km at 3 ohmm for BR2 (see Table II(1)). We interpret it by analogy with the Marambio soundings MA1 and MA2, Fig. 8: we complete the AH full line by JA dots till the presence of an acceptable thickness of permafrost, like the EQJC dots obtained for the MA1 sounding of Marambio, with the combination of the "VES transform in MTS with MTS". In this case, for the Brandy sties, we define a maximum thickness for a permafrost layer 0.070 km at 2000 ohmm + an adjustment brine layer 0.0105 km at 0.6 ohmm. (See Fig. 7 in the columns BR1 and BR2, and see also Table II.) Then follows for BR1 and for BR2 a sedimentary cover of 8 km: AHEF and AHEDG, a resistive layer, partly crust, partly upper mantle, and the top of the ICL at a depth of 78 km. The tectonic angle (strike) is NG 32°E for BR1 and Rhomax, and is Ng 21°E for BR2 and Rhomax also.

In the same way as for Marambio, a bidimensional interpretation was made for the BR1 and BR2 sites. A bay effect near the sounding site BR1 gives a contrast to Marambio: the TM curve is below the TE one. For BR2, more inside the country, the situation is the same, but with another cause of the relative position of the two curves, TM below the TE one. The structure below the sites, originally horizontal, is inclined: more in the BR1 case, and it is mescladed with basaltic intrusions and effusions (stronger at BR1 than at BR2).

A seismic sounding was made in the same region as BR1 (Keller and Diaz 1990).

Table III. Comparisons of layers for the MTS of Brandy I (and II) between A) the first normal determination; B) the results of the search to introduce minimum km of layers with the same trace of curve

A			B		
	layer thickness in km	layer resistivity in ohmm		layer thickness in km	layer resistivity in ohmm
SC	0.0700	2000.	SC	0.0700	2000.
NC	0.0105	0.6	NC	0.0105	0.6
NC	0.35	200.	NC	0.35	200.
NC	0.45	3.5	NC	0.45	3.5
NC	6.6	200.	.....1 km		
NC	0.5	4.		AV	
.....8 km			NC	4.	300.
	AV		NC	0.1	1000.
NC	70.	2000.	NC	1.	8.
SC	20.	10.	.....6 km		
SC	408.	2000.		AV	
SC	infinite	2.	NC	70.	2000.
			SC	20.	10.
			SC	408.	2000.
			SC	infinite	2.

SC: Supposed Content AV: Approximate Value
NC: Normal Content

It gives a sedimentary basin of 4.5 km thickness for "Brandy seismic section". Our results are not in accord with these depths for the basin, we have proposed 8 km (Table II). The sounding lines are at a distance of 0.2 km from each other. But the principle of non-uniqueness of the MT method permits us to look for a layers combination that does not change the trace of the sounding curve, but allows a shorter way for the interpretation. See Table III, for an understanding two layer combinations: A) the first with a sedimentary thickness of 8 km, and B) the second with only 6 km and containing a layer of a thickness 0.100 km at 1000 ohmm resistivity between two layers having lower resistivities.

We know also that the electric basement may be different in depth from the seismic basement. This remark explains the two km difference between the results of the two methods.

Section I.3 James Ross Island — Hidden Lake site: HI1 site is situated 20 km SW of the Brandy sites. As with Brandy there is no glacier ice on the site.

We interprete the Hidden Lake sounding HI1, Fig. 8, by analogy with soundings Brandy BR1 and BR2: we complete the full line CP by the MC dots so that we add an acceptable permafrost thickness of 0.1 km at 2000 ohmm + a very conductive layer (brine layer?) 0.090 km at 0.38 ohmm. A conductive-resistive contact is seen at point P on the full line curve CPDB. Then the top of the ICL follows at 72 km depth. But it is impossible to see something more because of the strong screening effect by the supposed brine layer. The tectonic angle or strike is Ng 39°E (Rhomin).

The HI1 sounding curve resembles LA1 one of Section II — See Fig. 9. We

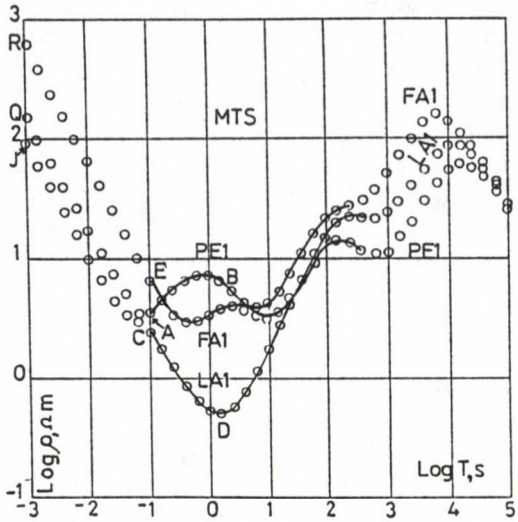

Fig. 9. MT theoretical unidimensional sounding curves for the sites situated on the Larsen Barrier in the Section II: for LA site: experimental results: full line LA1, for PE site: experimental results: full line PE1, for FA site: experimental results: full line FA1. The dots give the calculated corresponding curves

must be careful with the results of the sounding HI1, being apparently surprising for the upper part of its section.

For the three soundings of the James Ross Island, as for the Marambio soundings we add the same two final layers to avoid the "mutual effect" (Duhau et al. 1988). Table II gives the complete list of layers and Fig. 7 shows the sections.

Figure 7 shows that it is easy to connect at Section I the results of Seymour Island and James Ross Island: the existence of a permafrost; a brine layer, a little higher below James Ross Island than below Seymour Island; the supposed contact Upper-Lower Cretaceous; the sedimentary basement; the ICL (only the top of the ICL is defined at James Ross Island).

Section II. The region of this section seems to be a kind of basaltic tumor, defined by the Nunataks Foca (Fig. 2). These nunataks are basaltic islands, generally small, outcropping through the Larsen Barrier. The Nunataks Foca are a modern geomorphology feature made of subaerial basalts with AR/K radiometric ages ranging between 4 ± 1 my (Pliocene) and 0.2 my (recent) (Del Valle et al. 1983c). These nunataks are located 300 km east of a Mesozoic subsidence zone in the northwestern edge of South Shetlands Islands (Fig. 1). Subsidence stopped in the Early Tertiary. Consumption of oceanic crust was replaced during the Pliocene by rifting and basaltic volcanism on the Southeastern margin of South Shetland Islands (Fig. 1).

For this Section II, see the following studies: Del Valle et al. (1982, 1983a, 1983b, 1988), Fournier et al. (1989, 1992) and Muñoz et al. (1992).

Section II.1 Robertson Island: see in Fig. 2 the position of the sounding site. Figure 3 shows the MT interpretative curve RN obtained for site RO1. This curve

is the first one obtained on glacier ice in this synthesis: see Lefevre and Fournier (1957) and Fournier et al. (1957). We remark that the curve RN has the same trend as the one of Marambio CMD in Fig. 3.

For the Robertson Island sounding RO1, we complete the first layers by analogy with the Marambio soundings — Fig. 3 —, but with the restriction of connecting the two curves: full line RN and WR dots obeying the law of conductivity at the point R — as for Marambio MA1: full line CD with EQJC dots continuously at point C in Fig. 3. For the Robertson sounding we introduce a maximum thickness of 0.165 km of snow + glacier ice + permafrost (?) at a resistivity of 2000 ohmm + a brine layer 0.026 km at 0.5 ohmm. It is the first sounding having a glacier ice layer. Then we see 1.7 km Tertiary (?) + Upper Cretaceous having a resistivity of 200 ohmm (or more) and below, the Lower Cretaceous + Upper Jurassic (?) 0.5 km at 1.6 ohmm. The sedimentary basin has a thickness 2.4 km. After it a contact follows at a depth of 29 km, that is the top of the ICL, much higher than for Section I situated 200 km NE. (See Table II, Figs 3, 7 and Table IV.)

As with Marambio, the interpreted tectonic direction is Ng 33°E, Rhomin, Table IV: see Fig. 2. The sedimentary basin below the island is thinner than below Marambio, but with the same succession of layers — see Fig. 7.

Section II.2 Larsen site: on the Larsen Barrier, 5 km SW of the Larsen nunatak (Fig. 2). We interprete the Larsen sounding — curve LA1 in Fig. 9 — by analogy with the Brandy BR1 and BR2 soundings, we complete the CD full line by QC dots so we get an acceptable thickness of that ice barrier (0.135 km at 2000 ohmm) + a sea water thickness (0.210 km at 0.25 ohmm). Then we see a basin basement contact (?), without knowing the content of this basin because of the screening effect caused by the sea water layer (Fournier et al. 1989).

In the same way as Pomposiello et al. (1988) give the tensorial results for the study of Marambio, for the site Larsen, with a tectonic direction of Ng 50°E, Rhomin, (Table II, Fig. 7 and Table IV).

Section II.3 Pedersen site: on the Larsen Barrier, 5 km NE of Pedersen nunatak (see Fig. 2).

We interprete the Pedersen sounding PE1 full line curve by analogy with the Marambio MA1 sounding and the Robertson RO1 sounding. We complete the full line AB by JA dots in Fig. 9 with an acceptable thickness of the ice barrier (0.105 km at 2000 ohmm) + a sea water thickness (0.0148 km at 0.25 ohmm) without trespassing against the thickness maximum permitted in this case by the AB sounding full line curve. We see a sedimentary basin of 2.7 km thickness. The top of the ICL is at a depth of 27 km (Table II, Fig. 7 and Table IV).

The curve PE1, full line, Fig. 9, is the tensorial unidimensional interpretation having the direction Ng 40°E, Rhomin, the same trend as the two precedent sites.

We must note that for the Pedersen and Robertson Islands soundings, the 2D model response almost coincides for periods longer than 5–10 s with the response of a 2D model structure that synthesizes 1D results after Muñoz et al. 1992.

Section II.4 Fair Weather site: on the Larsen Barrier, 6 km SE of the Peninsula Coast (Fig. 2).

For the Fair Weather sounding, FA1 we complete the full line curve by analogy

Table IV. Partial results of the studies on the NE end of the Antartic Peninsula

Sounding	Brine Layer				Depth of	Intercalated Conductive Layer				Tectonic	Sounding
	1	2	3	4	the bottom	5	6	7	8	Angle	
Name	in km	in km	in km	in ohmm	in km	in km	in km	in km	in ohmm	(strike)	Site
MA1	0.137	0.155	0.018	0.6	5.9	78	101	23	7	Ng 57°E	See Fig. 2
MA2	0.137	0.155	0.018	0.6	5.9	78	101	23	7	Ng 55°E	See Fig. 2
BR1	0.070	0.080	0.010	0.6	8.0	78	?	?	?	Ng 32°E	See Fig. 2
BR2	0.070	0.080	0.010	0.6	8.0	78	?	?	?	Ng 21°E	See Fig. 2
HI1	0.100	0.190	0.090	0.38	5.0?	72	?	?	?	Ng 39°E	See Fig. 2
RO1	0.165	0.191	0.026	0.5	2.4	29.4	?	?	?	Ng 33°E	In the center of Robertson I
LA1	0.135	0.345	0.210	0.25	0.35	42.3	?	?	?	Ng 50°E	5 km SW of Larsen Nunatak
PE1	0.105	0.119	0.014	0.25	2.7	26.7	?	?	?	Ng 40°E	5 km NE of Pedersen Nunatak
FA1	0.270	0.312	0.042	0.25	2.5	44.5	?	?	?	Ng 36°E	6 km East of the coast

Remark: Ng means north geographic
On earth without snow: MA1, MA2, BR1, BR2, HI1
On earth with snow: RO1
On Barrier of Larsen, over snow: LA1, PE1, FA1
Brine Layer: 1: Depth of the top of the Brine Layer; 2: Depth of the bottom of the Brine Layer; 3: Thickness of the Brine Layer; 4: Resistivity of the Brine Layer
Intercalated Conductive Layer: 5: Depth of the top of the I.C.L.; 6: Depth of the bottom of the I.C.L.; 7: Thickness of the I.C.L.; 8: Resistivity of the I.C.L.

with the soundings of Brandy BR1 and BR2 and Larsen LA1 — RE dots in Fig. 9. We add 0.270 km of ice barrier at 2000 ohmm + 0.042 km of sea water at 0.25 ohmm. We see a layer combination analogous to that for the Pedersen site, but with a sedimentary thickness of only 2.5 km. The top of the ICL appears at a depth of 44.5 km (Table II, Fig. 7 and Table IV). The curve FA1, full line (Fig. 9), is the tectonic unidimensional interpretation. The direction is Ng 36°E, parallel to the axis of the local syncline and also parallel to the Peninsula Coast (Rhomin). We must be careful because the site is not far from a cape of the coast and perhaps a "cape effect" perturbation has modified the true position of the sounding curves.

Now, we see in Fig. 7, that it is possible to connect for Section II between the Robertson Island and the Larsen Barrier sites results, as we have done for Section I. But in this Section II, only the Lower Cretaceous + Upper Jurassic (?) is observable. The thickness of the ice barrier at Larsen is in general agreement with the values given by Skvarka (Del Valle et al. 1983c). We see a variable sea water depth, and also the mean top of the ICL: 36 km, with great dispersion, half in depth as at Section I: 78 km with a slight dispersion. We must try to understand the relatively high dispersion of the depth values for the top of the ICL in Section II results considering that:

1. The Marambio soundings were made relatively easy at the base.

2. The James Ross soundings were made with helicopter support, giving a good access to equipment at the soundings sites.

3. The Section II soundings were made first using airplanes, then an itinerant column of tractors and sledges with a minimum load of equipment, the people walking on the snow. In these conditions of work, it is difficult to do as a high quality work as at the former sites of Section I.

Section I and Section II remark: it seems in conclusion that these approaches are no fantasy because we stay between the limits of reality: integrated conductivity on the one hand and thickness of the Larsen Barrier given by Skvarka, on the other hand by Del Valle et al. (1983).

Remarks

Here, I give a list of the names of the 22 engineers and scientists that have worked with me in the Antarctic Peninsula, in the analysis, in the inversions and in the interpretation of the results: see Table V.

The historic fact that has permitted us to carry out this research, going on for already 15 years, or better said, the genesis of this "Scientific Andventure", was the acceptance of my suggestion by Capitan de Navio, don Roberto Manuel Martinez Abal, Director of the Argentine Antarctic Institute, to do MT research in the Antarctic Peninsula, on November 12th of 1976, in his office at Buenos Aires (from 1–79 to 10–93).

Table V. List of the 22 authors that have contributed with me in the electromagnetic, electric and seismic studies in the NE final part of the Antarctic Peninsula

Enrique Borzotta (1)	Marcelo Keller (3)
Enrique Buk (2)	Saturnino Leguizamón (5)
Bibiana Castiglione (1)	Arturo Nicolas Maidana (1)
Arturo Corte (2)	Manuel Jesus Mamani (1)
Rodolfo Del Valle (3)	Francisco Medina (3)
Jose Demicheli (4)	Carlos Emilio Moyano (1)
Maria Trinidad Diaz (3)	Miguel Muñoz (6)
Jose Miguel Febrer (3)	Jorge Nuñez (3)
Nestor Fourcade (3)	Alvaro Peretti (3)
Juan Carlos Gasco (4)	Maria Cristina Pomposiello (7)
Horacio Irigoin (4)	Jorge Venencia (1)

(1) CRICYT-IIACE-GEOFISICA, Casilla de Correo 131, 5500 Mendoza, Argentina
(2) CRICYT-GEOCRIOLOGIA, Casilla de Correo 330, 5500 Mendoza, Argentina
(3) Instituto Antartico Argentino, Cerrito 1248, 1010 Buenos Aires, Argentina
(4) Centro Espacial de San Miguel, Avenida Mitre 3100, 1663 San Miguel, Argentina
(5) Departamento de Sensores Remotos, IIACE, Casilla de Correo 131, 5500 Mendoza, Argentina
(6) Departamento de Geologia y Geofisica, Universidad de Chile, Casilla 2777, Santiago, Chile
(7) Centro de Investigaciones en Recursos Geologicos, CONICET, Ramirez de Velazco 847, 1405 Buenos Aires, Argentina

Conclusions

The MT results below Seymour Island and James Ross Island — Section I — are very interesting: we see a sedimentary basin with a thickness of about 6 km below the two islands cited. On the other hand, for Section II, the 80 km long profile (4 sites) shows a regular sedimentary thickness of about 2.5 km, exept for the Larsen site. An important fact common to the two regions studied is the content of the basin composed of a relatively resistive upper layer underlain by a very conductive layer (2 ohmm). It is provisionally proposed, that the contact between these two resistive layers could be the contact between the Lower and Upper Cretaceous. If this turns out to be true, it will be possible to follow this contact towards NE of the Antarctic Peninsula by the magnetotelluric method, eventually on icebergs and on ice-pack using the W method as is described by Del Valle et al. (1988), Fournier et al. (1989, 1992).

There is an intermediate conductive layer (ICL) below Seymour Island extending from 78 to 100 km depth and having a true resistivity of 7 ohmm. Below James Ross Island, the top of this ICL is at 78 km depth, without knowing its content. But for Section II, the top of this ICL is found at a depth of 36 km in average for the four MT sounding sites located over Robertson Island and on the Larsen Barrier.

We have also confirmed the existence of a brine layer below the permafrost that exists on the top of Seymour and James Ross Islands, and probably also below Robertson Island, depending on the thickness of the glacier ice cover, that is not present on the top of the two first islands already cited.

Aknowledgements

I am indebted to each of the 22 scientists and engineers who have participated with me in these expeditions in the Antarctic Peninsula, in the analysis, in the inversions and in the interpretation of the results giving me the opportunity to do this synthesis, see Table V.

The analysis and 1D and 2D inversions were made using six different computer centers belonging to three countries:

1) Argentine Antarctic Institute, Buenos Aires, Argentine,
2) Geophysical Center of Garchy, Garchy, France,
3) Air Force Spatial Center of San Miguel, San Miguel, Argentine,
4) Computer Center of the INPE, Sao Jose dos Campos, Brasil,
5) Computer Center of the Argentine Air Force, Buenos Aires, Argentine,
6) Computer Center of the CRICYT (CONICET), Mendoza, Argentine.

We are indebted to each of these six computer centers for permitting us to do the analysis of the records coming from the field and inversions for the interpretations.

We are particularly thankful to Capitan de Navio Roberto Manuel Martinez Abal and to Dr. C A Rinaldi, then Directors of the Argentine Antarctic Institute, for their help in the preparation of the successive campaigns.

Thanks are due to Dr. Richard Branham, Director of the Computing Division of the "Centro Regional de Investigaciones Cientificas y Tecnicas de Mendoza" for his help in the computations and for his English corrections of this manuscript, and also to Ana Maria Chaher for typing all the successive versions of the manuscript and doing the figures.

References

Anasimova N P 1978: Cryogenous metamorphization of chemical composition of subsurface water (exemplified in Central Yakutia). USSR Contrib. Permafrost, 2nd Int. Conf. Washington, Nat. Acad. Sci., 365–370.

Baskov E A, Zaytsev I A 1973: Principal features of hydrogeological structures within the URSS permafrost area. Proc. 2nd Int. Permafrost Conf., Yakutsk, URSS, p. 169.

Benderitter Y, Dupis A, Febrer J, Fournier H G 1978: In: Mémoire du B.R.G.M., Bureau de Recherches Géologiques et Minieres, 91, 71–78.

Berdichevsky M N, Fainberg E B, Rotanova N M, Smirnov J B, Vanyan L L 1976: *Ann. Geophys.*, 32, 143–155.

Borzotta E, Febrer J M, Fournier H G, Gasco J C, Maidana A N, Mamani M J, Moyano C E 1984: In: 13a. Reun. Cient. de la Asoc. Argent. de Geofis. y Geodes., San Carlos de Bariloche, Argentina

Cagniard L 1948: In: Ann. de l'Inst. de Phys. du Globe de Strasbourg, 4, 3, 3–28.

Dalziel W D 1983: Personal communication

Del Valle R A, Fourcade N M 1986: La cuenca sedimentaria pos-Triásico del Extremo Nororiental de la Peninsula Antártica. Inst. Antártico Argentino, Buenos Aires, Contrib. 323.

Del Valle R A, Fourcade N M, Medina F A 1982: In: Craddock C (Ed.), Antarctic Geosciences . Univ. Wisconsin Press, 275–280.

Del Valle R A, Fourcade N M, Medina F A 1983a: Inst. Antártico Argentino, Buenos Aires, Contrib. 276, 14.

Del Valle R A, Demicheli J, Febrer J M, Fournier H G, Gasco J C, Irigoin H, Keller M A, Pomposiello M C 1983b: *Note C.R.Ac.Sci.*, Ser. II, 296, 1189–1194.

Del Valle R A, Fourcade N M, Medina F A 1983c: Inst. Antártico Argentino, Buenos Aires, Contrib. 287.

Del Valle R A, Diaz M T, Febrer J M, Fourcade N H, Fournier H G, Gasco J C, Keller M A, Medina F A, Nuñez H C , Pomposiello N C 1988: *Acta Geod. Geoph. Mont. Hung.*, 23, 265–286.

Duhau S, Osella A M, Moyano C E 1988: Effects of the conductosphere on MT soundings. In: 9th Workshop Electromagnetic Induction in the Earth and Moon, Sotchi, Ukraina

Fournier H G 1993: Procedures and critiques for their use to obtain an acceptable synthesis with the resultst coming from 6 campaigns on the Antarctic Peninsula using 4 geophysical methods: magnetotellric, audiomagnetotelluric, electrical and seismic soundings from 1979 till 1992. 7th Sci. Assem. of the Int. Ass. of Geomag. and Aeron., Buenos Aires, Argentina

Fournier H G, Lefevre C, Albertinoli P, Bauer A, Blum A, Cagniard L 1957: Publ. de l'Assoc. des Ing. de la Fac. Polyt. de Mons, Belgique, 4, 1–2.

Fournier H G, Ward S H, Morrison H F 1963: Magnetotelluric evidence for the low velocity layer. Space Sci. Lab. Univ. Calif., Berkeley Techn. Rep. 222 (89), 4 (76), p. 22.

Fournier H G, Keller M A, Demicheli J, Irigoin H 1980: Contrib. Cient. del Inst. Antár. Argen., 4, 37–47.

Fournier H G, Corte A E, Mamani M J, Maidana A N, Borzotta E 1986: Actas de la Cu. Reu. de la Subcom. Latinoam. sobre la impor. de los proce. perigl., p. 57.

Fournier H G, Corte A E, Gasco J C, Moyano C E 1987: *Cold Reg. Sci. and Technol.*, 14, 155–161.

Fournier H G, Demicheli J, Gasco J C, Febrer J M, Del Valle R A, Keller M A, Pomposiello M C, Borzotta E 1989: *Cold Reg. Sci. and Technol*, 17, 49–59.

Fournier H G, Buk E M, Corte A E 1990: *Cold. Reg. Sci. and Tehcnol.*, 17, 301–307.

Fournier H G, Demicheli J, Gasco J C, Febrer J M, Del Valle R A, Keller M A, Pomposiello M C, Borzotta E 1992: Bul. Mens. de l'Ass. des Ing. de la Fac. Polyt. de Mons, Belgique, 10, 3–8.

Ginsburg G D, Neizvestnov Y V 1973: In: Proc. 2nd Int. Permafrost Conf., Yakutsk, 171–172.

Grikurov G E 1978: Geology of the Antarctic Peninsula. Amerind. Co. New Delhy, New York

Keller M A, Diaz T 1990: *Rev. Brasil. de Geofis.*, 8, 1–6.

Larin E M, Marov G P, Holmyansky M A, Neizvestnov Y V 1973: In: Permafrost 2nd Conf., Washington Nat. Acad. Sci., 176–177.

Lefevre C, Fournier H G 1957: *C. R. Acad. Sci.*, 244, 2410–2412.

Leguizamón S, Mamani M J 1993: In: Seminario ERS-1/SPOT de aplicaciones antarticas, Puntas Arenas, Chile

Maillet R, Doll H G 1933: *Beitrage Ang. Geophys.* 3, p. 109.

Mamani M J, Maidana A N, Fournier H G, Castiglione B, Venencia J, Peretti A 1993: Results from an AMT and MT campaign on the James Ross Island at the North East end of the Antarctic Peninsula. In: 7th Sci. Assemb. of the Inter. Assoc. of Geomag. and Aeron., Buenos Aires, Argentina

McGinnis L D, Nakao K, Clark C C 1973: In: 2nd Int. Conf. Permafrost, Yakutsk, 136–145.

Muñoz M A, Fournier H G, Mamani M J, Borzotta E 1992: *Acta Geod. Geoph. Mont. Hung.*, 27, 65–86.

Pomposiello M C, Fournier H G, Febrer J M, Gasco J C 1988: *Geoacta*, 15, 167–178.

Rinaldi C A (Ed.) 1992: Geologia de la Isla James Ross. Inst. Antar. Argent., CIRGEO, Buenos Aires

Yasko V G 1973: In: Permafrost 2nd Int. Conf. Washington Nat. Acad. Sci., 176–177.

Acta Geod. Geoph. Hung., Vol. 29(1–2), pp. 39–56 (1994)

HYPOCENTER DETERMINATION OF LOCAL EARTHQUAKES USING GENETIC ALGORITHM

I BONDÁR[1]

[Manuscript received January 13, 1994]

Genetic algorithms belong to the class of global optimization techniques. Unlike local, deterministic techniques, such as the least squares methods, they require no derivative information but use only misfit function evaluations, thus avoiding the linearization of the problem. Genetic algorithms start from a randomly chosen population of binary coded model parameter sets and use probabilistic transition rules to guide the search. Models with lower misfit values are reproduced and mated by the genetic operators selection, crossover and mutation with higher probabilities, while poorer models tend to die off, analogously to the principle of the survival of the fittest in biological evolution. Since the algorithm evaluates the objective function from many parts of the search space it is not likely to get trapped into a local minimum, but produces near optimal solutions. The averaged results of repeated runs of the algorithm provides a robust estimate for the global optimum.

We applied the genetic algorithm to localize the hypocenters of local earthquakes, so that the objective function, the weighted RMS travel time residual is minimized. In our implementation the search space includes the hypocenter coordinates and the crustal velocity model parameters, thus allowing the joint inversion of velocity model parameters with the hypocenter location. The power of genetic algorithm is illustrated on two examples, a dynamite explosion event with known hypocenter and some events selected from the earthquake sequence of 1985, Berhida, Hungary. The results show that genetic algorithm can perform better than an iterative matrix inversion that requires a good starting model.

Keywords: Berhida earthquake; explosion seismology; genetic algorithm; hypocenter; robust estimation; seismology

Introduction

The most commonly used techniques in hypocenter determination are the variants of the least squares method. These are local methods that strongly rely on using local information on the gradient of the objective function in order to improve the starting model in an iterative fashion (Lee and Lahr 1975, Klein 1978, Lienert et al. 1986). Therefore their success depend on the initial guess regarding the hypocenter location. If the starting model is far from the global minimum, local methods are prone to get trapped in a local minimum. Moreover, they need a reliable velocity model of the underlying Earth structure, as well.

[1]Seismological Observatory, Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences, H–1112 Budapest, Meredek u. 18, Hungary

1217-8977/94/$ 4.00

Stochastic global search methods, such as Monte Carlo methods, simulated annealing and genetic algorithms require no derivative information, thus avoiding the linearization of the multiparameter non-linear optimization problem. They use random processes to search the model space and to find better models.

Both simulated annealing and genetic algorithms have their analogies in natural optimization systems, i.e. in thermodynamics and biological evolution. Comparisons of the two methods can be found in Davis (1990), Ackley (1990) and Sambridge and Drijkoningen (1992). The superiority of genetic algorithms over the Monte Carlo method, which is a memoryless random walk over the model space, is shown by Sambridge and Drijkoningen (1992). Stoffa and Sen (1991) describes a combination of simulated annealing and genetic algorithms applied for inversion of plane-wave seismograms.

The primary development of genetic algorithms and their theoretical background is originated from Holland (1975). Genetic algorithm research is a rapidly evolving field of artificial intelligence and has already found many scientific and engineering applications. More recent summaries of the field have been given by Goldberg (1989) and Davis (1990).

Genetic algorithms differ from traditional optimization and search techniques in several ways. They use probabilistic transition rules to guide the search, not deterministic ones; they search from a population of models; they work with binary coded model parameters formed into bit-strings that are often called chromosomes, analogously to natural genetics, not the parameters themselves; and they use objective (fitness or cost) function information, not derivatives or other auxiliary knowledge. These algorithms are based on the principle of "survival of the fittest" by using some simple genetic operators: selection, crossover and mutation.

Genetic algorithms

Genetic algorithms start with a randomly chosen population of models. The model parameters are coded as binary strings and the resulting bit-strings are concatenated to build up a chromosome, which is considered one instant of the model space. For each model parameter we define a pair of bounds, a_i and b_i, such that $a_i \leq x_i \leq b_i$ and a resolution interval Δ_i, such that

$$N_i = 2^{\ell_i} - 1 = \frac{b_i - a_i}{\Delta_i} \tag{1}$$

where N_i denotes the number of discrete values of the ith model parameter, and ℓ_i stands for the length of the bit-string (i.e. the number of bits in the string) on which the model parameter within the given range and resolution can be represented. When all bits in the string are zero, the lower bound a_i, if all bits are one, the upper bound b_i is represented (see Fig. 1). Each model parameter can have different search ranges and resolution, which in most problems can be chosen *a priori*. If we

have M model parameters, the discrete model space will contain

$$N = \prod_{i=1}^{M} N_i \tag{2}$$

models.

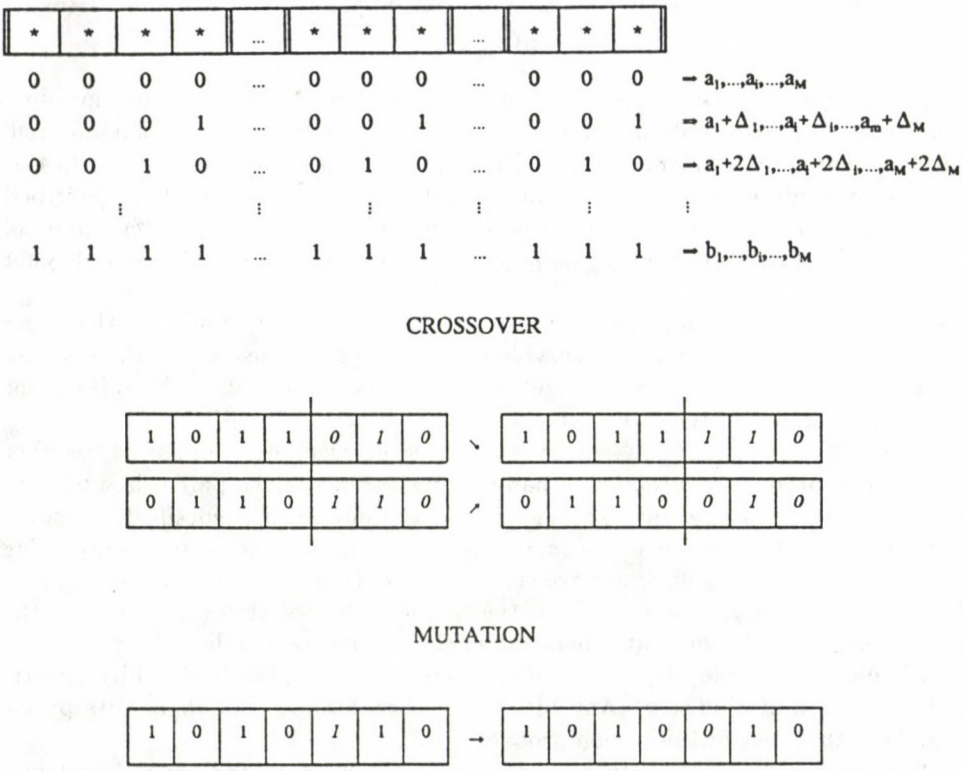

Fig. 1. Genetic algorithms work with binary coded model parameters, not the parameters themselves. The model parameters are coded into a bit string between the lower bound a_i and upper bound b_i linearly with resolution Δ_i, and concatenated yielding a "chromosome" (upper). The genetic operator crossover mates two chromosomes together by exchanging bits (italic) to the right of the randomly selected crossover position. The result is two offsprings that inherit information from both parents (middle). Mutation is the random change of a bit (italic) in a chromosome (lower)

The objective function (called fitness function in maximization or cost function in minimization problems) is evaluated for each member of the population, and models are selected for reproduction based on their fitness values.

The simplest selection method, stochastic sampling, defines the probability of selection for the kth model parameter vector $\mathbf{m}_k = m_k(x_1, x_2, \ldots, x_M)$ as the ratio

of the model's fitness value to the sum of all fitness functions:

$$P_s(m_k) = \frac{\phi(m_k)}{\sum_{k=1}^{n} \phi(m_k)} = \frac{\phi(m_k)}{n\phi_{\text{avg}}} \tag{3}$$

where n is the size of population. In minimization problems the probability of selection using the cost function is defined as

$$P_s(m_k) = \frac{\phi_{\max} - \phi(m_k)}{n(\phi_{\max} - \phi(m_k))} \,. \tag{4}$$

The selection process is continued until n models have been chosen for reproduction. In our implementation however, we use a more advanced selection method, called stochastic remainder selection without replacement (Goldberg 1989). In this selection procedure the expected count of each model is determined as described above, but each model will be replicated in direct proportion to the integer part of its expected count, while the fractional part is used as an additional probability for selection.

The selection procedure ensures that the more successful models with higher fitness or lower misfit values will survive and reproduce themselves at the expense of poorer models, while below average models will tend to die off in the subsequent generations, analogously to the principle of the survival of the fittest.

After the selection process was finished, the selected models are paired together randomly to produce $n/2$ couples of parent chromosomes. Each parent is subjected to the crossover (partial bit exchange) with its mate with a specified crossover probability P_c. If crossover is to occur, a crossover site is selected randomly along the bit-string, and two offsprings are created by exchanging the bits to the right of the crossover position (see Fig. 1). If the couple is not selected for crossover, the parents propagate to the next generation of the population unaffected.

The members of the offspring population are then subjected to the third genetic operator, the mutation step. Any bits of the chromosomes can change its parity (Fig. 1) with a specified mutation probability P_m.

The resulting new generation of chromosomes are evaluated, reproduced and mated over and over again, until a specified generation count is reached, or the population becomes homogeneous, i.e. the average misfit value approaches the minimal misfit in the population.

The three basic genetic operators play different roles in the algorithm. The reproduction step ensures the survival of the fittest from generation to generation via the selection mechanism. However, if the population size is too small, the problem of premature convergence may arise. This could happen when the initial population contains an extraordinary individual being close to a local extremum in the model space, which has a relatively high fitness value compared to the other members of the population. The above average individual would be reproduced by the selection process more frequently then its middling colleagues, and soon would take over a significant proportion of the later generations. The undesirable effect

of premature convergence can be avoided by scaling the fitness function (Goldberg 1989), or by simply increasing the population size.

The crossover operator is responsible for the mixing and sharing of information between the members of population (for more sophisticated crossover operators see Booker 1990).

Mutation can be considered as a background process, aiming to maintain the genetic diversity in the population, which would otherwise be exhausted by the previous two steps, and occasionally to introduce new genetic material. However, mutation probability should be kept low (Goldberg (1989) recommends $P_m \ell / L$, where L is the length of chromosomes), otherwise the algorithm becomes similar to a random walk over the search space, destroying the beneficial effects of the reproduction and crossover steps.

The robust searching features of genetic algorithms are attributed to the way the "schemata" are processed (Holland 1975). The schema is the term used to describe classes of strings with common set of elements. For example, consider the schema $H = 1**1*0$, where the asterisk is the "don't care symbol", i.e. $*$ can represent either the value 1 or 0. Strings belonging to the region of the search space designated by a schema (often called hyperplane) are the instances of that schema, e.g. the string 110100 is an instance of the above defined schema. Every string of length k is an instance of exactly 2^k distinct schemata.

The schemata can be characterized by two properties: schema order and defining length. The order of schema H, denoted by $o(H)$ is the number of fixed positions present in the template, while the defining length of a schema H, denoted by $\delta(H)$, is the distance between the first and last specific string position. In the example above, the schema $H = 1**1*0$ has defining length $\delta(H) = 5$ and order $o(H) = 3$. Using these notations, the expected number of copies for a particular schema in the next generation under reproduction, simple crossover and mutation can be calculated. The fundamental theorem of genetic algorithms (Goldberg 1989) states that the expected count m, of a schema H at iteration $t+1$ is bounded by the expression

$$m(H, t+1) = m(H, t) \frac{\phi_s(H)}{\phi_{\text{avg}}} \left[1 - P_c \frac{\delta(H)}{\ell - 1} - P_m o(H) \right] \tag{5}$$

where ϕ_s is the average fitness of all the strings in the population which are the instances of schema H. The theorem shows that schemata with above average performance, low order and short defining length will be sampled at exponentially increasing rates. Holland (1975) has estimated that the number of beneficially processed schemata at each generation is of the order n^3, despite the processing of only n chromosomes at each iteration. This feature, commonly referred as implicit parallelism, causes the genetic algorithm to explore the search space and exploit the highly fit models in a highly parallel fashion.

However, while the theory indicates sampling rates and search behaviour in the limit, any implementation uses a finite population of models. Estimates based on finite samples inevitably have a sampling error associated with them. Repeated iterations compound the sampling error and lead to search trajectories much different from those theoretically predicted (Booker 1990). Moreover, the convergence of the

algorithm to the global optimum of the objective function is not guaranteed. But, since the algorithm evaluates the objective function from many parts of the search space in parallel, it is not likely to get trapped into a local extremum when proper choice is made of the size of the population and probabilities of crossover and mutation. Genetic algorithm performs well even in problems intentionally designed to deceive the algorithm (Goldberg 1990), which indicates that genetic algorithms are not easily misled.

Hypocenter determination using genetic algorithm

For the determination of the hypocenter of an earthquake we use the first P and S (if available) arrivals at a number of stations. The stations are given by their coordinates (latitude, longitude, elevation) and delay times. The underlying crustal model is assumed to consist of horizontal, homogeneous layers over a homogeneous halfspace.

The multiparameter search space is built up from the hypocenter coordinates (latitude, longitude and depth) and the crustal model parameters (layer thicknesses, P velocities and P/S velocity ratios). The inclusion of crustal model parameters in the search space allows us to improve the underlying velocity model simultaneously with the determination of the hypocenter. However, any of the parameters can be fixed to a specified value, thus disabling its involvement in the search space. For example, in case of explosions or quarry blast, where the hypocenter coordinates are known, the method can be used to fine tune the velocity model. Each variable parameter is defined by a pair of bounds and a resolution interval. The bounds are automatically adjusted, so that equation (1) is satisfied.

Our objective is to minimize the root mean square of the weighted travel time residuals over the stations:

$$\phi = \frac{\sum_{i=1}^{N} (w_i Tres_i)^2}{\sum_{i=1}^{N} w_i^2} \tag{6}$$

where N is the number of stations, w_i denotes the applied weights, and $Tres$ is the travel time residual at a station. The travel time residual is defined as

$$Tres = Tobs - Tcal - Tdel \tag{7}$$

where $Tobs$ is the observed arrival time, $Tcal$ is the calculated arrival time and $Tdel$ denotes the station delay time.

The weights in the objective function are defined as a product of three independent weights: the first one describes the reading quality of the phase, which can be 0, 0.25, 0.5, 0.75 or 1; the second is a residual weight which ensures to weight down stations with large residuals; and the third one weights down the more distant stations. The residual and distance weights are determined dynamically during the run, similarly to that of the HYPOINVERSE program (Klein 1978).

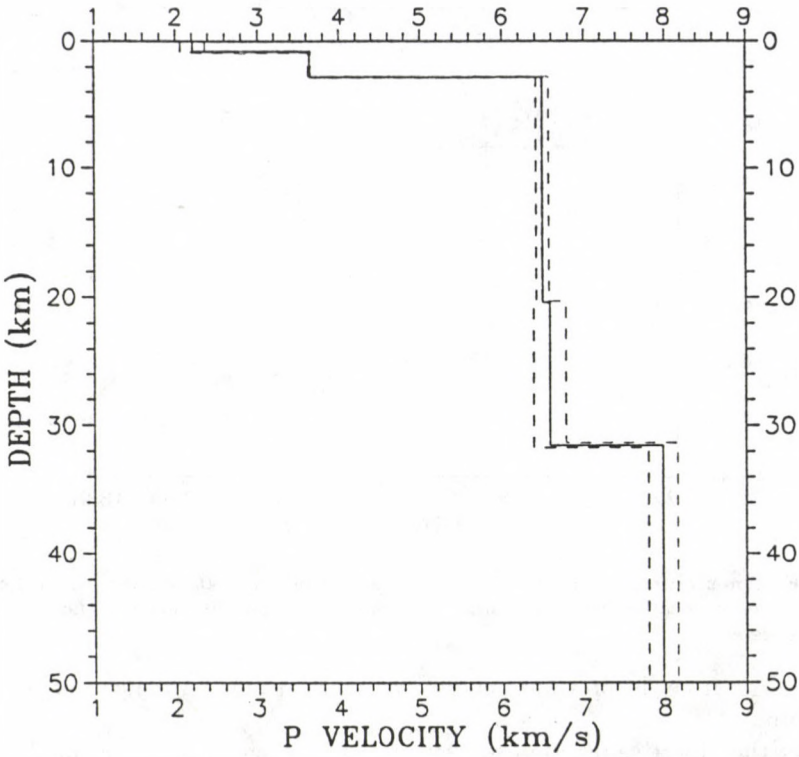

Fig. 2. Mean crustal velocity model (solid line) of the explosion November, 1990 and standard error ranges (dashed lines) computed from the results of 50 runs of the genetic algorithm

In order to evaluate the objective function, each chromosome in the population is decoded one by one to get an instance of parameter values in the model space. The travel times of the direct wave and all possible refracted waves under the parameter values (i.e. hypocenter coordinates, layer thicknesses, P and S velocities represented by the chromosome) are computed and the first arrivals are determined for each station. The S waves are considered pseudo-P waves in the computation.

Since no closed formula exists to determine the travel time of the direct wave from a focus deeper than the first layer, we use an iterative method to find the raypath between the focus and the stations by changing the angle of emergency from the focus. The iteration is stopped when the ray hits the station within a specified precision (usually 1 meter).

As we allow the inclusion of velocity model parameters in the search space, such velocity models may be generated that neither direct wave, nor refracted waves can reach a station. If this situation occurred, the chromosome representing that model is discarded and replaced by the best individual from the previous generation. If no individuals were discarded in the population then the worst solution, i.e. the individual having the largest RMS residual in the population is replaced by the

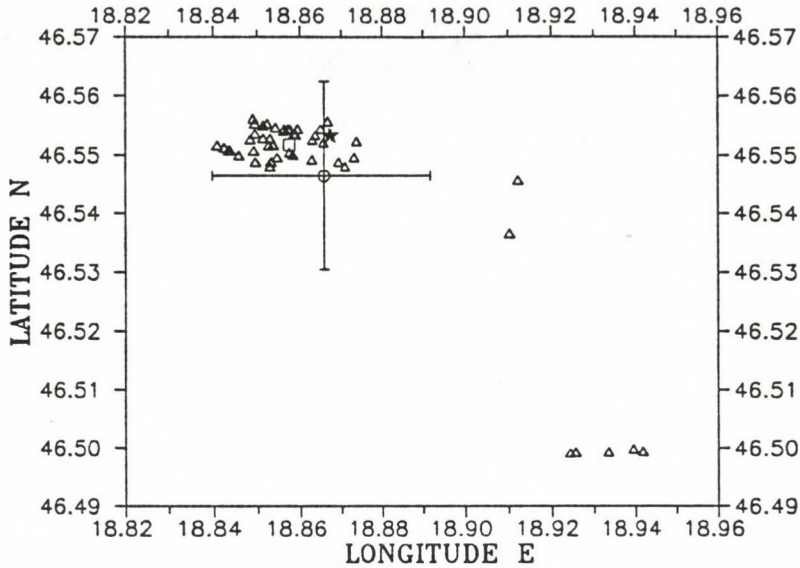

Fig. 3. Epicenter estimates for the explosion event November, 1990. Median (square) and mean (circle) epicenter location with standard errors computed from 50 runs and the true epicenter (star) are shown

best solution from the previous generation, which speeds up the convergence of the algorithm.

After the objective function was evaluated for each chromosome in the population, the individuals are ready to reproduced and mated by the genetic operators to make up the next generation.

When the algorithm finishes, the parameters of the best solution during the run, i.e. the model parameters (hypocenter coordinates and crustal model parameters) that yielded the smallest RMS travel time residual are reported. The origin time of the earthquake, the largest azimuthal gap between the stations, as well as the epicentral distances, epicenter-station azimuths, emergence angles from the focus, applied weights, travel times and travel time residuals regarding the best solution for each station are also reported.

Results

In the followings we demonstrate the performance of the genetic algorithm on two examples: an explosion and a shallow earthquake in Hungary.

On 20 November, 1990, 15:00:55.63 (GMT) 200 kg of dynamite was exploded 10 meters below the surface in the vicinity of Paks nuclear power plant, Hungary, in order to determine the structural response of the plant against seismic waves (Mónus and Szeidovitz 1991). The shotpoint was located at 46.5533N, 18.8672E and two permanent and five temporal (installed just for this purpose) stations recorded the event.

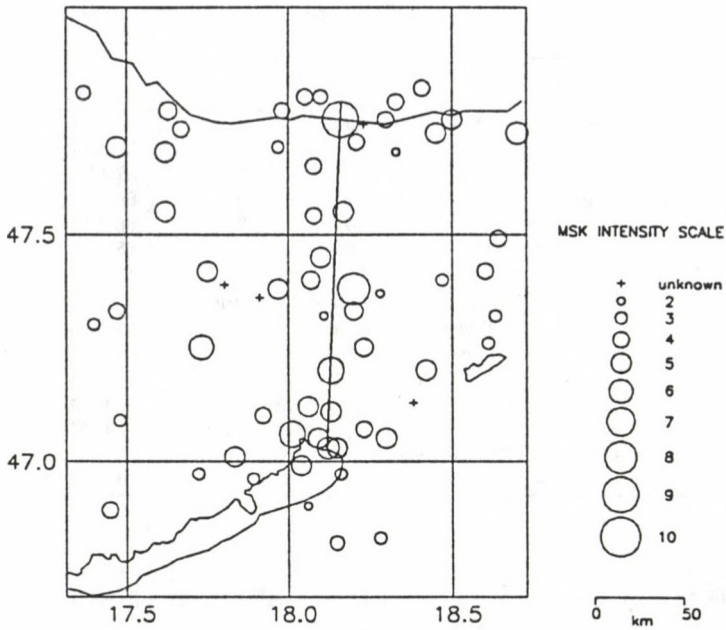

Fig. 4. Historical seismicity in the region of Berhida, Hungary. Prior to the 1985 earthquake sequence some 700 earthquakes were felt along the Komárom-Balatonkenese line (solid line) since 1599. Only the epicenters of the largest events are shown

Since the hypocenter coordinates are known, we used first the genetic algorithm to fine tune the crustal model parameters. Table I lists the input parameters for the algorithm. In each run we computed 50 generations with $P_c = 0.66$ crossover probability, $P_m = 0.0192$ mutation probability and 50 members in each population. The length of chromosomes were 52 bits, thus defining some $4 \cdot 10^{15}$ discrete models in the search space. Table II shows the best solution over the 50 runs of the program. Note, that it determined the origin time of the explosion as 55.624 s, which is in excellent agreement with the true origin time. The mean values and standard error ranges of the crustal model parameters computed from the results of the 50 runs are shown in Fig. 2.

In the next step we executed the program 50 times again with the same input values, but now the crustal model parameters were fixed to the mean values obtained from the previous experiment and the hypocenter coordinates are allowed to vary, so that we could see how precisely the genetic algorithm approaches the true hypocenter. The hypocenter coordinate bounds were chosen as 45.6809–47.3193N, 16.6809–18.3193E and 0.001–1.025 km with 0.0001°, 0.0001° and 0.001 km resolution steps for the latitude, longitude and focal depth, respectively. The latitude and longitude ranges define a quite big area (more than 20000 km^2), which makes the localization of the hypocenter more difficult for the genetic algorithm. With these bounds and resolutions the search contained some $3 \cdot 10^{11}$ discrete models.

Table III shows the best solution over 50 runs of the program. The epicenter

Table I. Input parameters for genetic algorithm 20 November 1990 explosion event

Station parameters							
No.	station	φ (°N)	λ (°E)	h (km)	delay (s)	P arrival (s)	P weight
1	PAK	46.5718	18.8439	0	0	56.700	0.75
2	MMA	46.5188	18.8492	0	0	57.160	0.75
3	SZI	46.6084	18.8840	0	0	57.810	1.00
4	UZD	46.5911	18.5804	0	0	60.400	0.50
5	GYA	46.6913	18.7734	0	0	60.430	0.75
6	MEV	46.1135	18.1123	0	0	68.700	0.75
7	PSZ	47.9194	19.8944	0.94	0	89.200	0.50

Crustal velocity model parameters								
No.	layer thickness (km)				P velocity (km/s)			
	min	max	resolution	bits	min	max	resolution	bits
1	0.50	0.81	0.01	5	1.64	2.27	0.01	6
2	1.70	2.01	0.01	5	3.63	4.27	0.01	6
3	17.30	17.94	0.01	6	5.94	6.58	0.01	6
4	10.80	11.44	0.01	6	6.24	6.88	0.01	6
5	–	–	–	–	7.68	8.32	0.01	6

Table II. Best solution over 50 runs of genetic algorithm with fixed hypocenter coordinates, 20 November 1990 explosion event

Station parameters							
No.	station	distance (km)	azimuth (°)	emergence angle (°)	P travel time (s)	P residual (s)	P weight
1	PAK	3.11	339.28	90.18	1.370	−0.295	0.750
2	MMA	3.81	180.00	38.71	1.603	−0.067	0.750
3	SZI	6.22	10.00	38.71	2.267	−0.082	0.999
4	GYA	16.90	335.29	20.44	4.183	0.623	0.742
5	UZD	22.34	280.96	20.44	5.018	−0.242	0.490
6	MEV	75.82	230.11	20.44	13.248	−0.172	0.536
7	PSZ	170.55	26.67	20.44	28.210	5.366	0.019

Crustal velocity model and focus parameters				
No.	thickness (km)	P velocity (km/s)	focus parameters	
1	0.81	2.27	latitude (°)	46.5533N
2	2.01	3.63	longitude (°)	18.8672E
3	17.63	6.50	depth (km)	0.01
4	11.36	6.74	origin time (s)	55.624
5	∞	8.07	rms residual (s^2)	0.095
			azimuthal gap (°)	153.33

location estimates of the 50 runs, as well as the true epicenter, the median and the mean epicenter with its error ranges are shown in Fig. 3. Although there are some poorer solutions, the genetic algorithm robustly localized the epicenter: either the median or the mean estimate for the epicenter location falls within 0.01° distance away from the true epicenter. The focal depth estimations are a bit poorer, the best solution gave 176 m for the focal depth, while the averaged depth estimate over the 50 runs is 320 m with 262 m standard error.

Table III. Best solution over 50 runs of genetic algorithm with fixed crustal velocity model parameters, 20 November 1990 explosion event

			Station parameters				
No.	station	distance (km)	azimuth (°)	emergence angle (°)	P travel time (s)	P residual (s) weight	P
1	PAK	2.20	0.00	94.57	0.998	−0.292	0.750
2	MMA	3.81	180.00	37.40	1.559	−0.393	0.750
3	SZI	6.22	18.38	37.40	2.222	−0.405	1.000
4	GYA	16.31	339.34	19.92	4.031	0.405	1.000
5	UZD	21.10	280.97	19.92	4.768	−0.362	0.500
6	MEV	75.11	229.40	19.92	13.089	−0.383	0.749
7	PSZ	170.88	27.03	19.92	28.242	4.964	0.198

Crustal velocity model and focus parameters				
No.	thickness (km)	P velocity (km/s)	focus parameters	
1	0.801	2.211	latitude (°)	46.5553N
2	2.003	3.640	longitude (°)	18.8522E
3	17.616	6.491	depth (km)	0.176
4	11.115	6.577	origin time (s)	55.994
5	∞	7.980	rms residual (s^2)	0.388
			azimuthal gap (°)	152.97

Table IV. Events used in computation from the earthquake sequence Berhida, Hungary

event	date (yr–mo–dy h:m)
1	85–08–15 04:13
2	85–08–15 04:16
3	85–08–15 04:18
4	85–08–15 04:28 (main shock)
5	85–08–15 04:44
6	85–08–15 05:29
7	85–08–15 08:58
8	85–08–15 09:05
9	85–08–15 09:54
10	85–08–15 10:53

Our next example is one of the largest earthquakes in Hungary that has taken place in this century. The earthquake of magnitude $M_b = 4.7$ (NEIC, ISC) occurred in the western part of the Pannonian Basin near Lake Balaton on 15 August, 1985, 4h 28m (GMT). The shock caused moderate damage in the Berhida-Peremarton epicenter region and slight damage was reported from Budapest, too. The quake was felt throughout western Hungary and southwestern Slovakia. It was also felt at Zagreb, Croatia and in Burgenland and at Vienna, Austria. The maximum intensity has been estimated VII on the MSK-64 scale. Tóth et al. (1989) have determined the focal mechanism as strike-slip faulting with nearly horizontal compressional stress axis oriented E-W.

Earthquakes in the Pannonian Basin are mostly crustal events, sometimes fol-

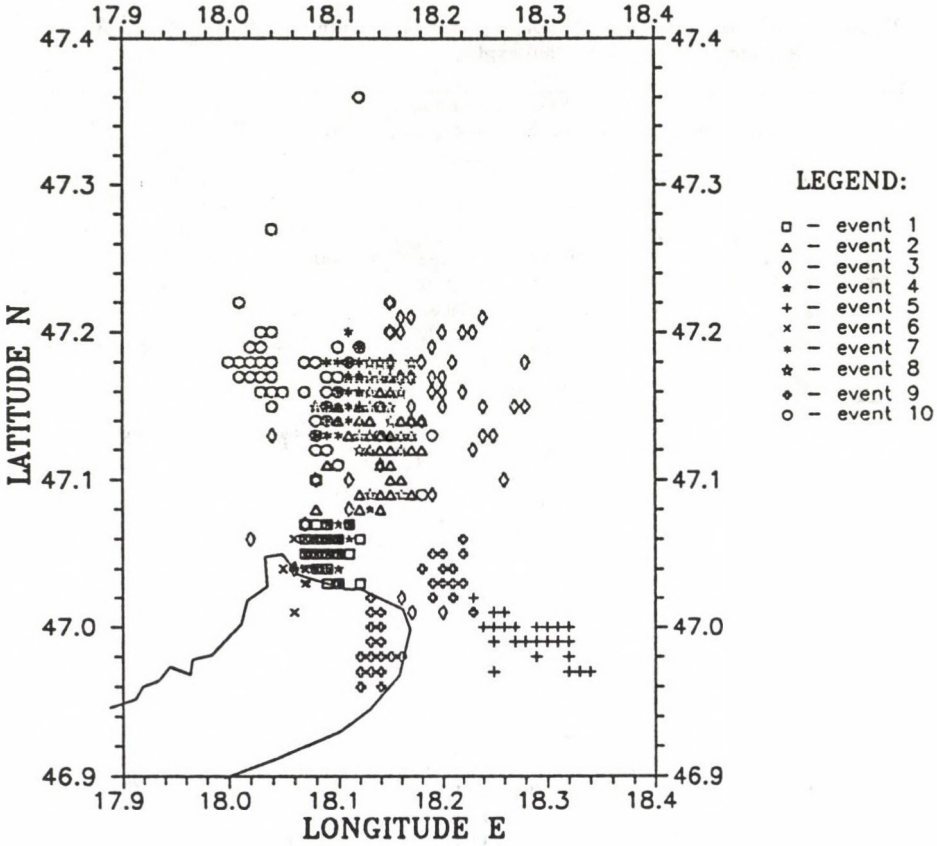

Fig. 5. Epicenter estimates for the 10 events selected from the 1985 Berhida earthquake sequence. The epicenters were computed 50 times for each event

lowed by a series of aftershocks. The relatively strong earthquake at Berhida, 15 August, 1985 was preceded by 3 foreshocks and followed by 29 aftershocks with a magnitude greater than 3.0. A large number of aftershocks occurred in the area after the major shocks — at least 226 aftershocks were identified from 15 August, 1985 to 30 July, 1986.

Berhida lies just on the Komárom-Balatonkenese line which is well-known for its relatively high seismicity. Prior to the 1985 Berhida earthquake sequence, some 700 earthquakes were felt along this line since 1599 (Zsíros et al. 1988). The epicenters of the largest events are shown in Fig. 4.

We determined the hypocenter of the 3 foreshocks, the main shock and 6 aftershocks. The events taken into consideration are listed in Table IV. The program was executed 50 times for each event, so that statistics could be computed on the results. The population size and the number of generations were 50 again, the

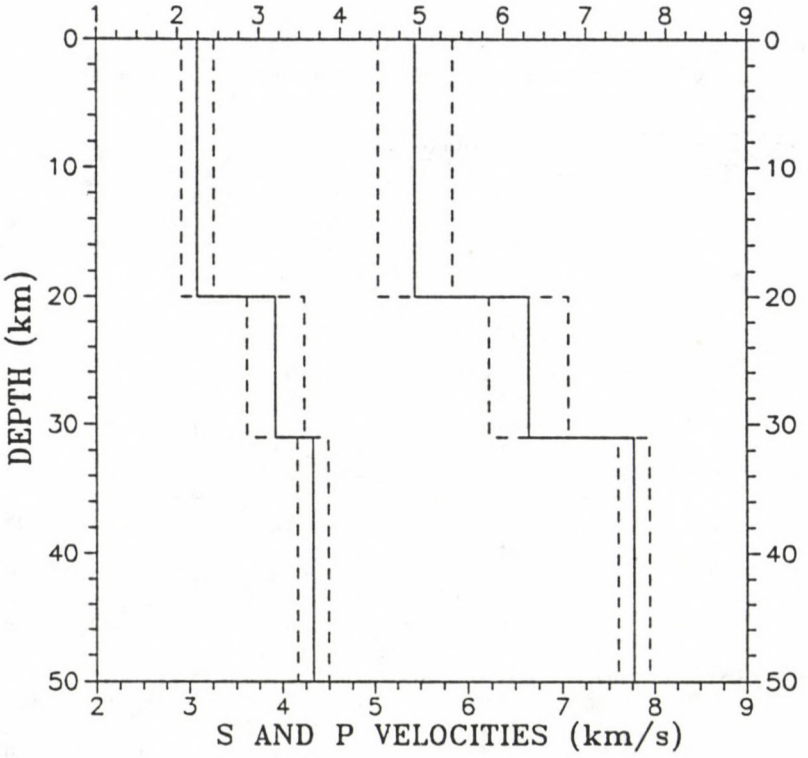

Fig. 6. Mean crustal velocity model (solid line) of the 1985 Berhida events with standard error ranges (dashed lines) computed from the results of 500 repeated runs

Table V. Crustal velocity model parameter and hypocenter coordinate ranges for Berhida events

Crustal velocity model parameters									
No.	layer thickness	P velocity (km/s)				V_P/V_S velocity ratio			
		min	max	res.	bits	min	max	res.	bits
1	20.0	4.97	6.25	0.01	7	1.59	1.75	0.01	6
2	11.0	5.94	7.22	0.01	7	1.67	1.82	0.01	6
3	∞	7.39	8.67	0.01	7	1.71	1.86	0.01	6

Source parameters											
latitude (°N)				longitude (°E)				focal depth (km)			
min	max	res.	bits	min	max	res.	bits	min	max	res.	bits
46.94	47.58	0.01	6	17.94	18.57	0.01	6	3.0	9.4	0.1	6

crossover and mutation probability were chosen as $P_c = 0.66$ and $P_m = 0.0196$, respectively. The same source coordinate and crustal model parameter bounds and resolution were taken for all of the events, as shown in Table V, so the search space contained some $2 \cdot 10^{15}$ discrete models for each event.

Table VI. Comparison of genetic algorithm results with hypocenter estimates of different agencies for Berhida events

event	agency	date	seconds	φ (°N)	λ (°E)	h (km)
1	CSEM	85–08–15	14.53±0.22	47.09±0.02	18.02±0.03	10.
	NEIC	4:13	12.0	47.0	18.1	15.
	ISC		12.8±0.87	47.08±0.031	18.03±0.05	9.±7.9
	GA		11.74±0.56	47.05±0.013	18.09±0.012	5.98±1.8
2	CSEM	85–08–15	6.89±0.43	47.14±0.04	18.14±0.05	10.
	NEIC	4:16	6.0	47.2	18.1	10.
	ISC		5.8±0.60	47.25±0.049	18.23±0.064	10.
	GA		5.15±0.619	47.12±0.022	18.14±0.023	7.03±1.5
3	CSEM	85–08–15	26.01±1.16	47.35±0.11	18.19±0.11	10.
	NEIC	4:18	21.5	47.1	18.2	10.
	ISC		25.2±0.71	47.36±0.012	17.91±0.08	10.
	GA		21.61±0.564	47.15±0.053	18.18±0.06	6.86±1.4
4	CSEM	85–08–15	49.43±0.11	47.07±0.01	18.07±0.02	10.
	NEIC	4:28	46.9	47.0	18.1	10.
	MOS	main	49.6	47.04	18.00	33.
	LDG	shock	51.3	47.0	18.0	
	HRVD		51.9±2.2	47.11±0.27	17.9±0.18	10.
	ISC		47.4±0.13	47.06±0.016	18.01±0.019	10.
	GA		46.78±0.539	47.05±0.011	18.09±0.012	5.48±1.7
5	CSEM	85–08–15	32.87±0.37	47.03±0.02	18.10±0.04	10.
	NEIC	4:44	31.4	47.0	18.1	10.
	ISC		31.0±1.1	47.06±0.034	17.99±0.064	3.±10
	GA		30.16±0.345	46.99±0.011	18.28±0.029	8.51±0.9
6	CSEM	85–08–15	19.19±0.19	47.10±0.02	18.06±0.02	10.
	NEIC	5:29	17.3	47.0	18.0	9.
	MOS		17.4	46.95	18.28	33.
	LDG		20.7	47.0	18.0	
	ISC		18.0±1.6	47.04±0.029	18.01±0.039	11.±13
	GA		16.75±0.596	47.05±0.010	18.08±0.012	6.32±1.8
7	CSEM	85–08–15	57.12±0.56	47.10±0.05	18.13±0.04	10.
	NEIC	8:05	55.9	47.1	18.1	10.
	ISC		57.0±1.9	47.2±0.12	18.1±0.12	10.±16
	GA		56.53±0.763	47.16±0.017	18.11±0.010	5.83±1.8
8	CSEM	85–08–15	32.91±0.64	47.35±0.07	18.2±0.06	10.
	NEIC	9:05	28.7	47.1	18.2	10.
	ISC		31.0±1.6	47.3±0.12	18.17±0.091	5.±12
	GA		27.85±0.308	47.15±0.022	18.14±0.018	4.28±1.3
9	CSEM	85–08–15	7.24±0.31	47.01±0.03	18.14±0.03	10.
	NEIC	9:54	6.3	47.0	18.1	10.
	ISC		6.0±2.4	47.0±0.13	18.1±0.13	4.±18
	GA		5.53±0.665	47.01±0.028	18.17±0.036	5.36±1.8
10	CSEM	85–08–15	19.12±0.44	47.2±0.04	18.01±0.06	10.
	NEIC	10:53	16.0	47.0	18.0	10.
	ISC		17.4±0.37	47.14±0.033	18.05±0.053	10.
	GA		18.03±1.061	47.17±0.041	18.07±0.045	6.44±1.6

GA refers to genetic algorithm estimates, other data are read from ISC Bulletin

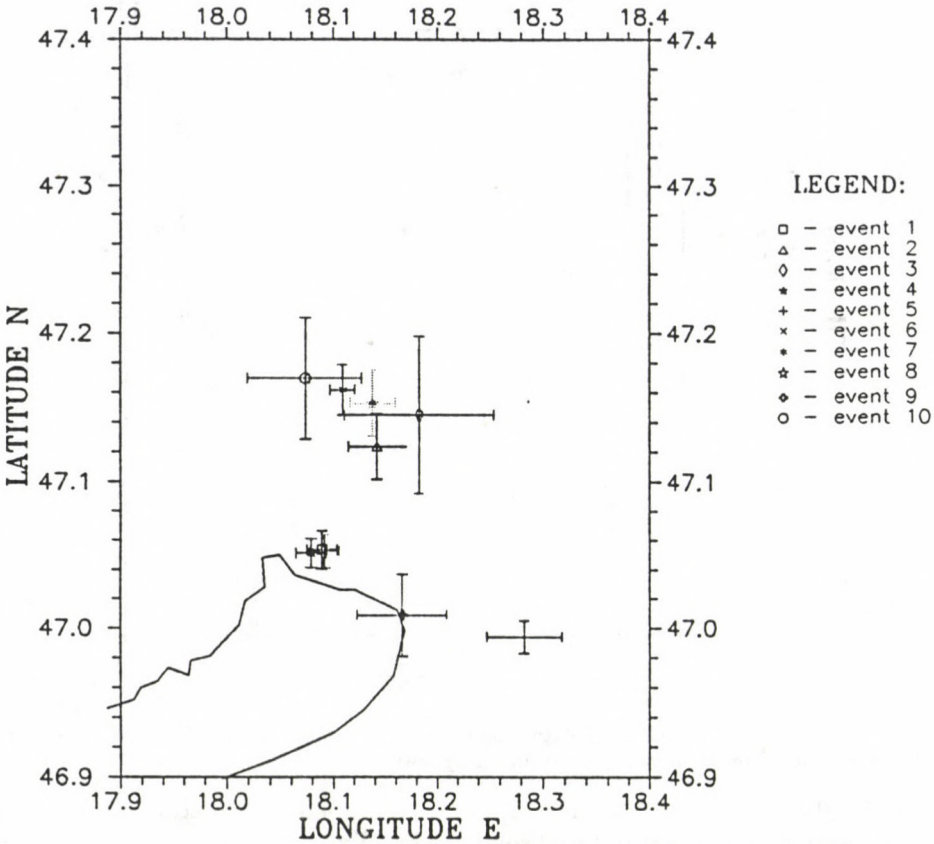

Fig. 7. Mean epicenter locations for the 1985 Berhida events with standard error ranges, each computed from 50 repeated runs

Figure 5 shows the computed epicenters of the events, while the mean values of the crustal velocity (P and S) parameters with their standard error bounds, which were computed from the results of the 500 runs can be seen in Fig. 6.

Figure 7 displays the averaged epicenters over the 50 runs and the standard error ranges for each event. The mean hypocenter depths of the events together with their error ranges are shown in Fig. 8. The obtained focal depth estimates indicate a rather shallow focus around 6 km deep in the crust, in accordance with the depth distribution of focuses in the Pannonian Basin (Zsíros et al. 1989).

Discussion and conclusions

Genetic algorithms are a class of search algorithms that have been used in many optimization problems. The algorithm requires minimal information, i.e. it uses only the misfit function evaluations in controlling the search. It is independent of the details of the forward problem and needs no derivative information, but, owing

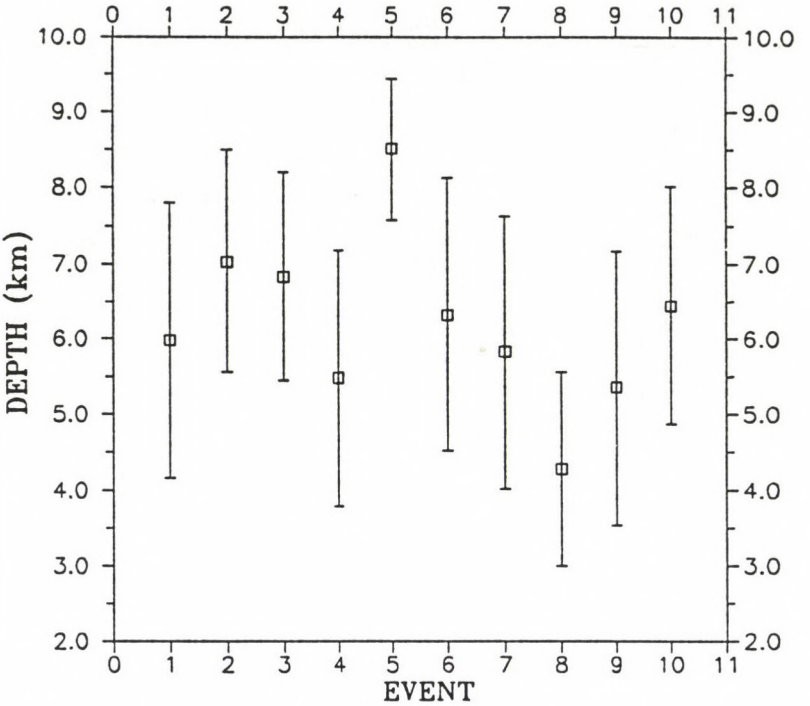

Fig. 8. Focal depth estimates for the 1985 Berhida events. Mean values (square) with standard error ranges computed from 50 repeated runs are shown

to its implicit parallelism, is able to efficiently exploit information in the model search space and find near optimal solutions rapidly.

We recall here, that reproduction copies models according to their objective function values, i.e. the models with lower misfit values have higher probabilities of getting copied. Therefore, in the end of a genetic algorithm run the resulting population will contain many copies of one or more very good models that are close to the global minimum. Thus, by repeating the genetic algorithm with different initial random populations, we can acquire a set of good models that can be used to calculate the mean model and its standard error bounds, which provides a robust estimate for the global optimum.

We have shown that in the case of the 20 November, 1990 explosion event the mean hypocenter estimate computed from 50 repeated runs of the genetic algorithm provides an excellent solution for the true hypocenter.

For the 10 selected event from the 1985 Berhida earthquake sequence Table VI lists the mean hypocenter estimates obtained from the repeated runs of the genetic algorithm and the estimates given by international agencies as read from the ISC Bulletin. Figure 9 shows the epicenter estimates of the different agencies and the mean estimate obtained from the repeated genetic algorithm runs for the main shock, 15 August 1985, 4:28 that has been recorded by more than 300 stations

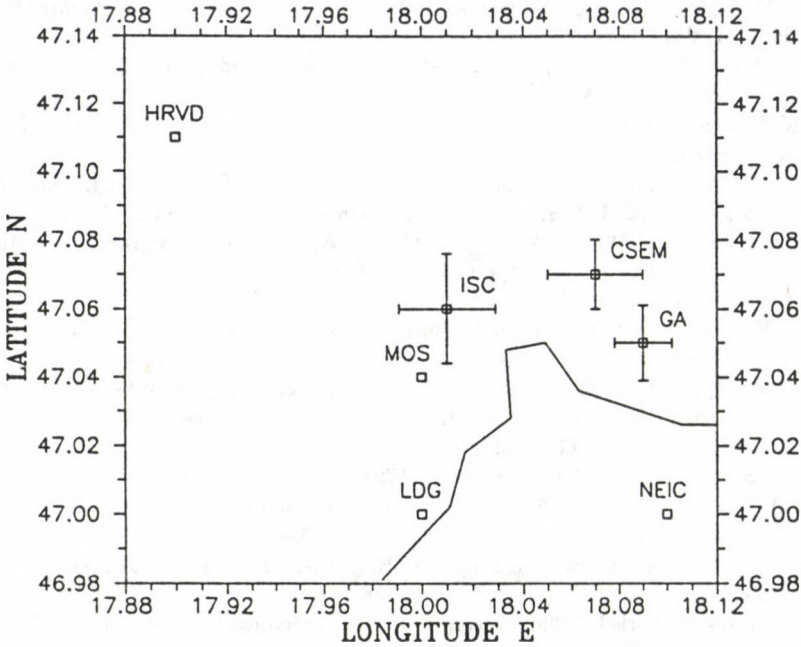

Fig. 9. Comparison of genetic algorithm epicenter estimate for the 1985 Berhida main shock with epicenter estimates of different agencies. Error bars are shown where they are available

all over the world. The estimate given by HRVD seems to be an outlier, but the other estimates, including the genetic algorithm (labelled GA in the Figure) fall close to each other. Furthermore, comparing the various estimates in Table VI, it can be concluded that the estimates provided by the genetic algorithm are in good agreement with that of the international agencies, but in the case of focal depth estimations, genetic algorithm gives much more consistent estimates with definitely less variance than the agencies.

Acknowledgment

The author would like to thank Péter Mónus who provided many helpful and stimulating discussions.

References

Ackley D H 1990: In: Genetic Algorithms and Simulated Annealing, L Davis (Ed.), Pitman, London, 170–204.

Booker L 1990: In: Genetic Algorithms and Simulated Annealing, L Davis (Ed.), Pitman, London, 61–73.

Davis L (Ed.) 1990: Genetic Algorithms and Simulated Annealing, Research Notes in Artificial Intelligence, Pitman, London

Goldberg D E 1989: Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Reading, MA

Goldberg D E 1990: In: Genetic Algorithms and Simulated Annealing, L Davis (Ed.), Pitman, London, 74–88.

Holland J H 1975: Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor

Klein R W 1978: Hypocenter Location Program HYPOINVERSE Part I: Users Guide to Versions 1, 2, 3 and 4, U.S. Geol. Surv. Open-File Rept. 78-694.

Lee W H K, Lahr J C 1975: HYPO71 (revised): A Computer Program for Determining Hypocenter, Magnitude and First Motion Pattern of Local Earthquakes, U.S. Geol. Surv. Open-File Rept. 75-311.

Lienert B R, Berg E, Frazer L N 1986: *Bull. Seism. Soc. Am.*, 76, 771–783.

Mónus P, Szeidovitz Gy 1991: Seismic Noise Measurements at the Paks Nuclear Power Plant (in Hungarian), Technical Report 2-1991, Seismological Observatory, Geodetic and Geophysical Res. Inst., Hungarian Academy of Sciences

Sambridge M, Drijkoningen G 1992: *Geophys. J. Int.*, 109, 323–342.

Stoffa P L, Sen M K 1991: *Geophysics*, 56, 1794–1810.

Tóth L, Mónus P, Zsíros T 1989: *Gerlands Beitr. Geophysik*, 98, 312–321.

Zsíros T, Mónus P, Tóth L 1988: Hungarian Earthquake Catalog (546-1986), Seismological Observatory, Geodetic and Geophysical Res. Inst., Hungarian Academy of Sciences, Budapest

Zsíros T, Mónus P, Tóth L 1989: *Gerlands Beitr. Geophysik* 98, 146–154.

Acta Geod. Geoph. Hung., Vol. 29(1–2), pp. 57–80 (1994)

RECENT HORIZONTAL DEFORMATION IN THE PANNONIAN BASIN MEASURED WITH EXTENSOMETERS

P VARGA[1] and T VARGA[2]

[Manuscript received February 3, 1994]

The extensometric network and the observed strains of the Pannonian Basin are described. Deformation data are interpreted in a comparison with the stress measurements obtained worldwide, in Europe and in Hungary.

Keywords: deformation; extensometer; Pannonian Basin; sress-strain relation

Introduction

High precision $10^{-9} - 10^{-11}$ extensometric (strain) data have been used recently to study local tectonic conditions. Results of observations at stations equipped with such instruments are strongly influenced by local effects (e.g. topography, cavity). The interpretation of the data obtained from the network of such stations is difficult, because each single site equipped with strainmeters is differently disturbed by local effects. Nevertheless succesful experiments were made to use extensometric data for tectonic purposes in connection with seismicity in active areas (Latynina and Karmaleeva 1978, Agnew 1986). An attempt was realized to study the regional recent tectonic activity with strainmeters in an intraplate and "qutie" area in the Pannonian Basin (Varga et al. 1993). All strainmeter sites of this experiment were equipped with identical quartz tube extensometers operating under similar conditions and the equipment was maintained at every station in the same manner.

The main goal of the present study is to prove the objectivity of the extensometric data in terms of real geological processes. It was found that local extensometric observations have a comparable accuracy in regional interpretation to other methods of earth sciences.

Stress and strain fields of our planet, of Europe and of the Pannonian basin

The determination of the stress and strain fields is an important problem of geosciences. Stresses, deformations and movements are often regarded as a consequence of the processes in the Earth's mantle.

[1]Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences, H–9401 Sopron, POB 5, Hungary

[2]Eötvös Loránd Geophysical Institute of Hungary, H–1145 Budapest, Columbus 17–23, Hungary

1217-8977/94/$ 4.00

In the frame of a broad international cooperation (Zoback et al. 1989) the global features of the tectonic stress field were determined on the basis of earthquake focal mechanisms, well breakouts, *in situ* stress measurements (hydraulic fracturing and overcoring), studies of young geological deformational features. The stress regime of our planet can be characterized as follows (Zoback 1992)

— in most places a uniform stress field exists throughout the upper brittle crust

— the interplate regions are determined by compression in which the maximum principal stress is horizontal

— active extensional tectonism can be observed in topographically high areas both on the continents and in oceanic areas.

It is remarkable that stress orientations together with the relative stress amplitudes are uniform over broad geological regions. There are different sources of the regional stress field with the common feature that they are connected to the plate tectonic activity (e.g. forces acting at the plate boundaries, viscous drag on the subducting slabs).

In Europe directions determined by shallow focus earthquakes are very coherent (Udias et al. 1989). A similar situation is valid in case of other methods of stress field determination. The regional patterns of the stress distribution in Europe show three regional areas (Müller et al. 1992):

— the maximum compressive horizontal stress ($S_{H_{max}}$) has a NW to NNW orientation in Western Europe,

— a WNW-ESE $S_{H_{max}}$ orientation in Scandinavia with a large variability of $S_{H_{max}}$ orientations

— an E-W $S_{H_{max}}$ orientation and N-S extension in the Aegean Sea and Western Anatolia.

These different stress fields — based on 1500 stress orientation determinations — are the result of plate-driving forces acting on the boundaries of the Eurasian plate. An overview of the first results of stress determination in the Pannonian Basin was published in 1990 by Dövényi and Horváth.

The $S_{H_{max}}$ directions in the Pannonian Basin can be related to the well known NW-SE direction of western Central Europe (Grünthal and Strohmayer 1992). A similar $S_{H_{max}}$ orientation was described for Přibram (Czech Republik) (Peška 1993) and for different parts of Austria (Kohlbeck et al. 1980).

The scatter of $S_{H_{max}}$ directions in the Pannonian Basin is considerable. The orientation distribution changes significantly from the NW-SE direction typical in the western part of the Pannonian Basin to the NNE-SSW direction typical in its central and eastern parts (Müller et al. 1992). In addition a NE-SW $S_{H_{max}}$ orientation was also observed in the focal mechanisms of the seismic events in the central and eastern part of the Pannonian Basin. According to Müller et al. (1992) the $S_{H_{max}}$ orientations of this region might result from the superposition of the roughly

NW stress orientation in Western Europe and the NNE oriented compression acting along the NE coast of the Adriatic. Becker (1993) comes to the conclusion that "On average the absolute maximum horizontal stress is oriented to WNW-ESE in the overall Pannonian Basin with the only difference that it is compressional at the Earth's surface in the western and probably in the eastern part of the Pannonian Basin but tensional in its centre." The estimated maximum stress magnitudes scatter between 0 and ± 10 MPa.

On the distribution of the horizontal displacements in a planetary scale, there are first of all evidences for the rigid plate motions. In this direction the methods of space geodesy (satellite laser ranging SLR, lunar laser ranging LLR, very long baseline interferometry VLBI) gave a lot of very important information. The Global Positioning System (GPS) is able to monitor regional and local tectonic motions by repeated measurements. Of course, the increased accuracy of space geodetic methods leads to a more pronounced role of gravimetry in detecting the mechanics of the deformations or displacements (Groten 1991). The new geodetic methods had not sufficient accuracy in some cases and the need of repeated measurements distributed over an interval of several years made them time consuming. To monitor the recent tectonic activity, strain and tilt observations of high accuracy (10^{-10}) are run mainly in active seismic areas.

Extensometers

The first strainmeters (extensometers) were developed in the 19th century, but due to high technical requirements the first equipment of this type was made by Benioff as late as in 1935, who called it as "linear strain seismograph". The equipments of this type are able to measure the relative displacement of two points at a distance of some 10 meters from each other with an accuracy of 10^{-3} micrometers what means 10^{-10} relative units. The importance of the strainmeters lies in the fact that phenomena with periods from some 10 s up to years can be reliably recorded with them corresponding to the wide spectrum of movements taking place in the solid Earth (Table I). Thus they connect the frequency bands of seismic waves and of phenomena studied usually by means of geodetic methods. It is worth mentioning — even if according to authors' knowledge it has not been studied in detail — that the amplification of the extensometers at different frequencies is practically constant. Due to this feature of the strainmeters they are often used in attempts to predict seismic events (Rikitake 1976, Asada 1982, Kurskeev 1990).

From the visual point of view the most characteristic phenomena on strainmeter records are earth tides. The analysis of the records shows that the main diurnal and semidiurnal waves can be detected with a signal to noise ratio of 10 or higher, the phase delay of the tidal waves can be obtained with a considerable accuracy, too. As example the harmonic analysis results of an extensometer record of the Budapest Geodynamical Observatory for 1984-85 obtained by the Venedikov (1966) method is shown. The well pronounced differences in the amplitudes and phase delays of the best determined diurnal (0_1 and K_1) and semi-diurnal (N_2, M_2 and S_2) waves can be explained by the loading influence of oceanic tides especially strong in Europe

Table I. Deformation phenomena in the solid Earth

Type of phenomena	Period (in s)	Relative strain
Waves from distant earthquakes	$10^{-1} - 10^{-2}$	$10^{-6} - 10^{-9}$
Microseisms	$(1 - 20) \cdot 10^{0}$	$10^{-9} - 10^{-11}$
Free oscillations	$10^{2} - 10^{3}$	$10^{-8} - 10^{-11}$
Tides	$2 \cdot 10^{4} - 1 \cdot 10^{6}$	$10^{-7} - 10^{-8}$
Seasonal processes (meteorological, hydrological)	$3 \cdot 10^{7}$	$10^{-5} - 10^{-7}$
Annual variation of axial rotation speed (estimated elastic strain)	$3 \cdot 10^{7}$	$10^{-9} - 10^{-10}$
Annual speed of tectonic phenomena	$3 \cdot 10^{7}$	$10^{-5} - 10^{-7}$

for M_2 and S_2 consituents (Pertzev and Ivanova 1991). The residual noise spectrum of the data shows a 0.59 nano-strain noise in the frequency band of semi-diurnal waves and a 0.71 nano-strain noise in case of diurnal waves. In fact these noise level values — as it is usual in case of extensometers — are almost by one order higher than in case of the records obtained by recording gravimeters. They are mainly of natural (e.g. meteorological) but also of technical origin. The lower accuracy of instrument calibration is another reason why strainmeters are not competitive with recording gravimeters in the study of the solid Earth's tides at present. According to the authors' experience a strainmeter can be calibrated with an accuracy of 5 percent while gravimeters with a reliability of 0.5 percent. It is reasonable to try to reduce the record noise and to increase calibration accuracy of strain records because in tidal domain they carry — in principle — a greater amount of information on the Earth's inner structure than gravity and tilt observations preferred recently. (For instance the liquid core resonance characterised usually by the difference of amplitude ratios of 0_1 and K_1 waves is 1.8 percent in case of gravity, 5.8 percent in case of tilt, and higher than 15 percent in case of strain observations (Varga 1976)).

Another conspicuous phenomenon on the strain records are earthquakes. Of course extensometers can only record, longer seismic waves. To give an idea on the resolving power of the extensometers used, a long periodic seismic wave package of the earthquake of January 15, 1993 (Japan, Hokkaido $M_b = 6.9$, $M_s = 7.0$) is shown (Fig. 1) in comparison with the record obtained at the Piszkéstető Observatory, Hungary ($\varphi = 47°57'$, $\lambda = 19°57'$) with the very broad band seismograph of the type STS-2.

The long term strain variations can be connected to meteorological and hydrological phenomena as well as to recent tectonic activity. To separate these from the instrumental drift is a complicated task. To get valuable long periodic strain signals, very stable conditions and careful maintenance must be guaranteed over long time intervals. The separation of signals of the order of 10^{-7} (Table I) can be particularly useful since they may be in connection with recent regional tectonics.

The strainmeters (extensometers) are of three types: wire, rod (or bar) and laser instruments. The common basic principle of them is that the displacement gradient

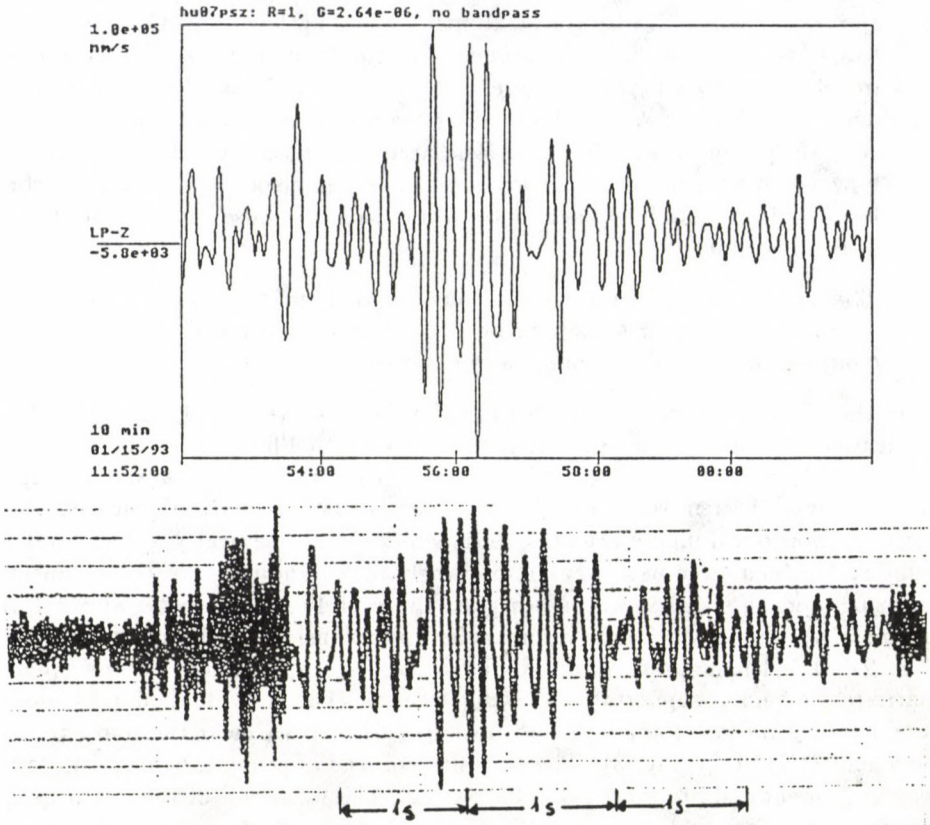

Fig. 1. Long periodic surface wave package of the earthquake of January 15, 1993 (Japan, Hokkaido) recorded with STS-2 very broad band seismograph and with 21.3 m long quartz tube strainmeter of Budapest Geodynamical Observatory

should be constant for the baselength of the equipment (what cannot be perfectly realized). The continuous measurement of realistic external (among them, tectonic) signals requires the following very strict constructional conditions:

— Perfect attachment of the strainmeter to the rock and of the different instrument parts to each other. This requirement is of first order importance because for perfect observations, the homogeneous behaviour of a basically inhomogeneous system is needed.

— For the observations with strainmeters extremely stable *thermal conditions* are needed. The diurnal temparature variations must be kept far below 0.1 C°, the annual ones must not exceed some centigrades. For the reduction of instrumental disturbances, materials with low thermal conductivity are needed.

— Stable and high quality *reference* of the length and a satisfactory *method of the comparison* of it with the strainmeter baselength is needed. As reference a basis connected to the meter standard and measurement by differential laser interference are suitable. There are different methods of comparison, but neither of them allows a calibration better than some percents, and this is not enough for the earth tide data. The accurate and systematic calibration is, however, important not only for the tidal components of the strain records, but also for the study of the long term variations, because it shows the stability of the instrument.

— The *transducers* of strainmeters must be of high accuracy and sensitivity. To produce a sensor satisfying both conditions mentioned is theoretically a complex problem of metrology and engineering.

In the present study quartz tube rod extensometers were used, the basic idea of which was developed at the Shmidt Institute of Earth Physics, Moscow (Latynina and Karmaleeva 1978). To increase the sensitivity a specially high quality capacitive transducer system was constructed in the Geodetic and Geophysical Research Institute, Sopron, Hungary (Mentes 1981) which worked several years without any technical problem even in sites with extremely high humidity. For the extensometer suspended on stainless wire silica glass tubes of 45 mm diameter and 2–3 mm thickness were used which have a linear thermal extension coefficient of $0.45 \cdot 10^{-6}$. (Fig. 2). The lenghts of the individual tube pieces are 2.0–2.5 m. The suspended quartz rod is held by supports with levelling screws. The individual quartz tubes are assembled by means of adhesive and invar profile pieces (linear thermal extension coefficient $0.7 \cdot 10^{-6}$) (Fig. 3). The adhesive consists of cement, quartz sand and a two-component resin to accelerate the setting. The fixed end of the extensometer is connected to the rock by a metallic rod which is adjoined to the quartz tube by a magnetostrictive calibration unit (Fig. 4). This device consists of a coil with a permendur-core of a resistivity about 40 ohm. The linear deformations of the calibration unit caused by the effect of currents of different intensities were determined by a laser interferometer. In order to reach high accuracy and high resolution, optical and capacitive sensors were used simultaneously in the first working period. Both transducers were installed at the free end of the extensometer system and show excellent agreement (Latynina et al. 1984, Mentes 1991). The displacement at the free end of the measuring system is indicated by a capacitive sensor, and can be recorded with a resolving power which enables observations with a relative sensitivity of 10^{-10} or even better if the external — natural and artificial — noise conditions are favourable. For the processing and interpretation, the determination of the sensitivity with adequate accuracy meant an important difficulty. A study of the calibrating unit showed that the scale value of the record can be determined with an inner accuracy of 1.5 percent at the stability of 1.0 percent of the calibration current. Taking into account the error of the interferometric calibration it can the spoken only about a calibration accuracy of about 5 percent as already mentioned. The built-in calibration unit enabled the linearity control of the total system, too. During routine recording, calibration is carried out automatically once in a day.

A detailed study of the mechanical part is given by Latynina et al. (1978). The following are to be mentioned:

— Forces hampering the free motion of the quartz tube rod are especially dangerous when the fitting of the individual tubes is not perfect. Forces acting against the motion of the rod at the suspension wires are proportional to the strain. Hence this problem has not an importance of first order.

— Elastic deformations of the quartz tube rod can be estimated if the sensors are installed at different places along the instrument. On this basis it was found that errors of such a type are not significant.

The influence of meteorology on the strain measurements must not be neglected. The problem of possible external thermal influences will be discussed later on in detail. The seasonal thermoelastic deformations reflected in extensometric records may reach a considerable magnitude ($10^{-5} - 10^{-7}$) at a depth of some ten meters

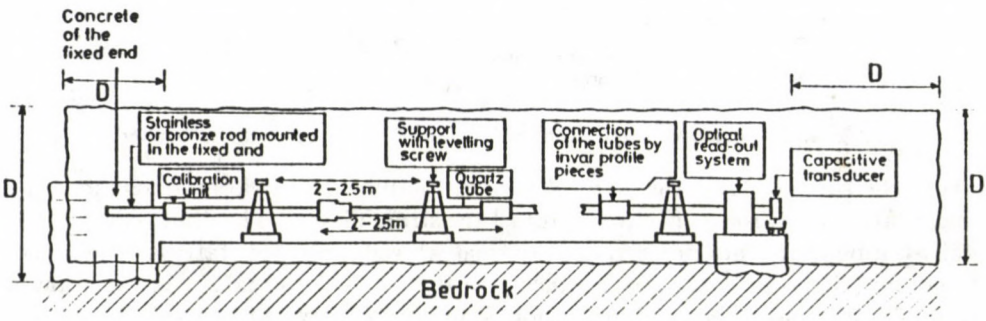

Fig. 2. The principal scheme of quartz rod strainmeters used in the Pannoninan Basin's network

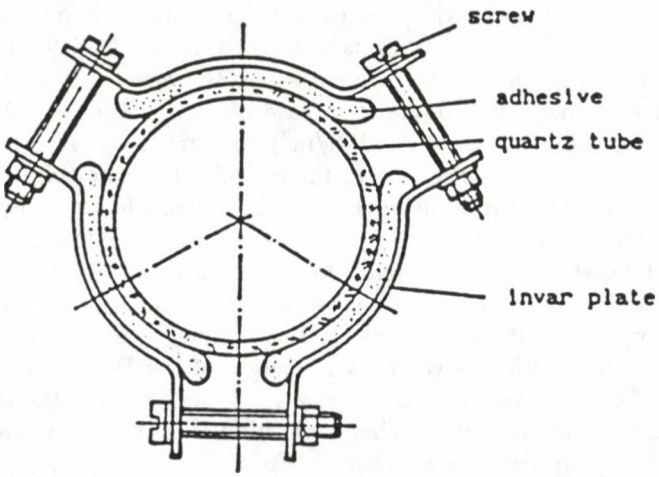

Fig. 3. Sectional drawing of the attachment of the tubes (Mentes 1991)

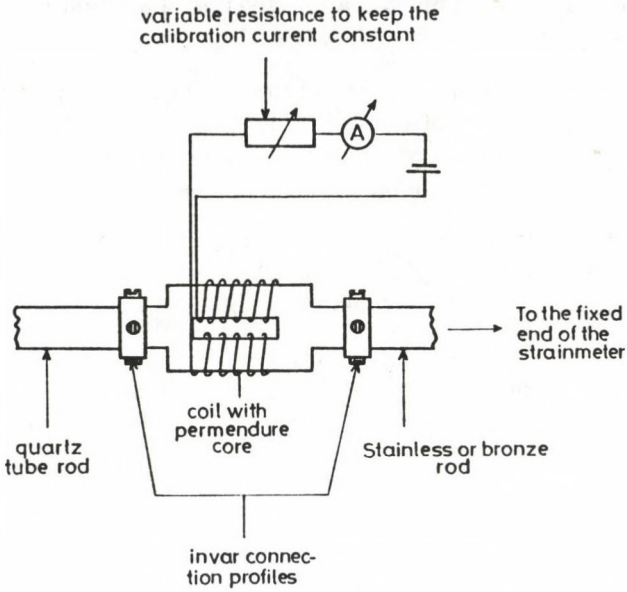

Fig. 4. The magnetostrictive calibration unit of the quartz rod strainmeters

under the surface. From the point of view of thermoelastic deformations monolitic and relatively soft rocks effectively protect the strain meters. The direct effect of the daily temperature variation is relatively small at well protected stations, they are of the order $10^{-2} - 10^{-1}$ centigrade which give $10^{-9} - 10^{-8}$ diurnal strain variation in case of silica glass. The yearly variation is about 1°C at well protected sites which means an annual strain $10^{-7} - 10^{-6}$ and this is easy to separate from the long term variations.

Variations of the atmospheric pressure can be hardly determined due to their irregular behaviour. The main error is connected to the length variation of the quartz rod induced by air pressure which is $\Delta p/3k$ (where Δp is the air pessure variation and k is the bulk module). In case of quartz $k \sim 4 \cdot 10^{10}$ N/m^2 and at strong barometric changes (2–3000 N/m^2) the strains caused by air pressure variations reach 10^{-8} which is similar to the magnitude of the tidal strain.

Atmospheric moisture influences the strain in dependence of hydrogeological conditions. High porosity leads to bigger strains up to 10^{-7}. As far as the surface water induced by strain is concerned, close correlation between level and strain was found at some places (Erpel tunel, Germany) (Wilmes 1983). In case of the Budapest Geodynamic Observatory, the observed extensometric material does not show any connection with the water-level changes of the Danube which is flowing in a distance of about 2 km. A connection with the variation of the level of karstic water was found in several cases. This connection is, however, a complicated one (Fig. 5) and needs further investigation.

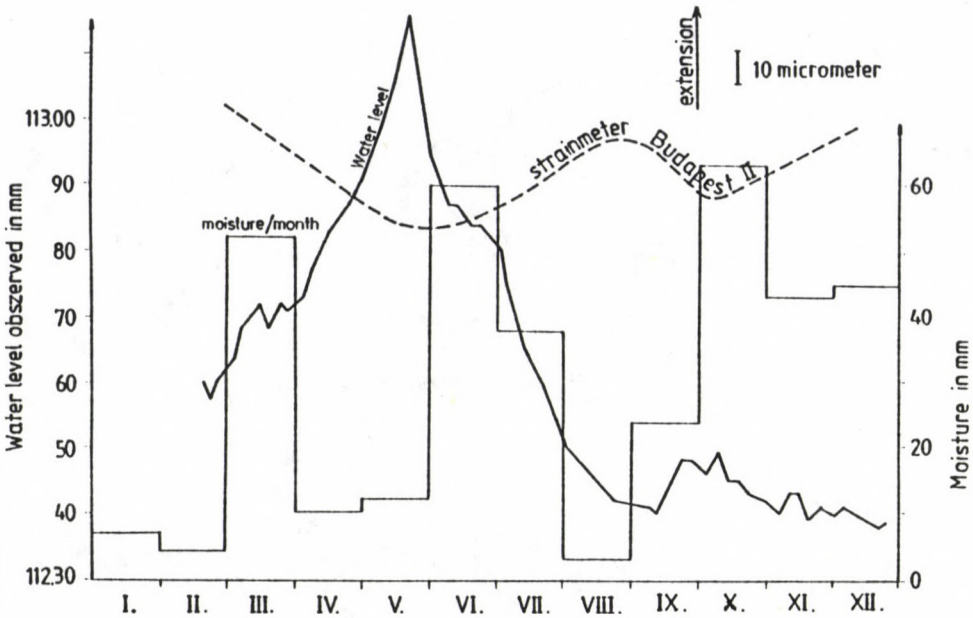

Fig. 5. Karst water level variations under the Budapest Geodynamical Observatory and the corresponding residual strain variations

Extensometric network of the Pannonian Basin

Data are used from all stations in operation at this time. Records of adequate length are at disposal from the stations Budapest and Sopron (Fig. 6). The records of the observatory operating in Pécs was not used because it was installed in 1991 and its records did not reach sufficient length. The Hungarian stations are operating in close cooperation with the stations Vyhné (Slovakia) (Brimich and Latynina 1988) and Beregovo (Ukraine) (Latynina et al. 1992).

The *Budapest Geodynamical Observatory* (Latynina et al. 1984) is situated in the NW part of the Hungarian capital (Fig. 7) in a natural cave formed in limestone by thermal water. The level of the karstic water is about hundred meter deeper than the level of the station. The strainmeters (Fig. 7) equipped by capacitive transducers are at a distance 30–35 m both from the entrance and from the surface. In order to decrease the cavity effect the posts of the strainmeters are mounted in a way that the quartz-tubes lie almost in the axis of the galleries, and the concrete pillars of the free and fixed ends are from the ends of the site at a distance of about the diameter of the quasi-cylindrical tunnels. The most important parameters of the two equipments installed are listed in Table III. The annual temperature variation at the sites of the strainmeters is $\leq 1°C$, while the diurnal change does not exceed 0.1°C. To reduce the influence of the surface temperature variation a system of hermetic doors were installed in the tunnels between the surface and the strainmeters. This station is at the same time the Hungarian 1st order absolute

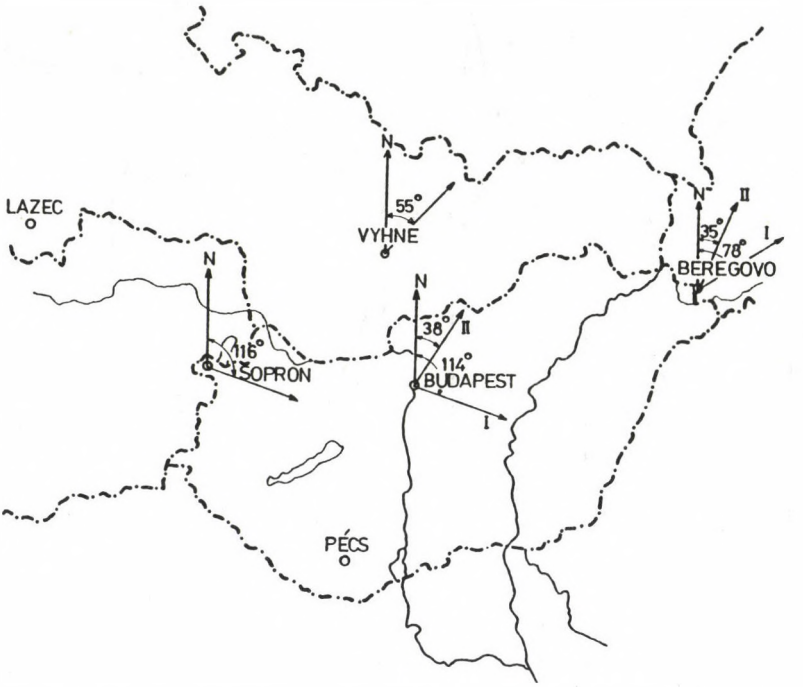

Fig. 6. The extensometer network operating in the Pannonian Basin

gravimetric point, permanent gravimetric records are run with several instruments. Special laboratory devices for gravimeter calibration are also available here. For a more correct interpretation of the strain observations at the station, barometric and temperature variations are regulary recorded and geophysicists from the Mecsek Uran Mining Company (Pécs) record the radon gas content, too.

The Sopron Geodynamical Observatory (Mentes 1991) is in the western part of the city (Fig. 8) in an artificial tunnel driven in gneiss. The distance of the extensometer from the entrance is about 30 m; the overlaying of the gallery is 60 m thick and is separated from the gallery by thermal insulation. Due to technical reasons the instrument is not in the tunnel axis. The temperature variation is less than 0.5 °C per year and 0.05°C per day. Seismographs and a recording gravimeter are run in the station, too. The most important characteristics of the quartz tube strainmeter can be found in Table III. A microbarograph developed in the Geodetic and Geophysical Research Institute is regulary used at the Sopron Observatory for the interpretation of the extensometer records.

Station Beregovo (Ukraine) (Latynina et al. 1992) is 10 km far from the city of Beregovo in a tunnel driven into tuff. The vicinity of the station is characterized by a seismic activity higher than in the Pannonian Basin. Distance of the equipment from the surface is more than 15 m, the station is carefully isolated from the temperature variations.

Station Vyhné (Slovakia) (Brimich et al. 1988) is a gallery driven in granite. The instrument is installed 130 m from the entrance. The relative depth under the surface is 50 m. The seasonal change of the temperature is 0.2 °C, the diurnal one is almost one order of magnitude less. The extensometer is equipped with an optical (resolution $5 \cdot 10^{-9}$) and a photoelectrical recording system (sensitivity $5 \cdot 10^{-10}$) (for details see Table III). The air pressure variations recorded at the meteorological observatory in Mlynany (30 km westward from Vyhné) were used for the interpretation of the aperiodic deformation.

When the above mentioned four stations are used as a network in order to achieve further geophysical information, the following additional conditions must

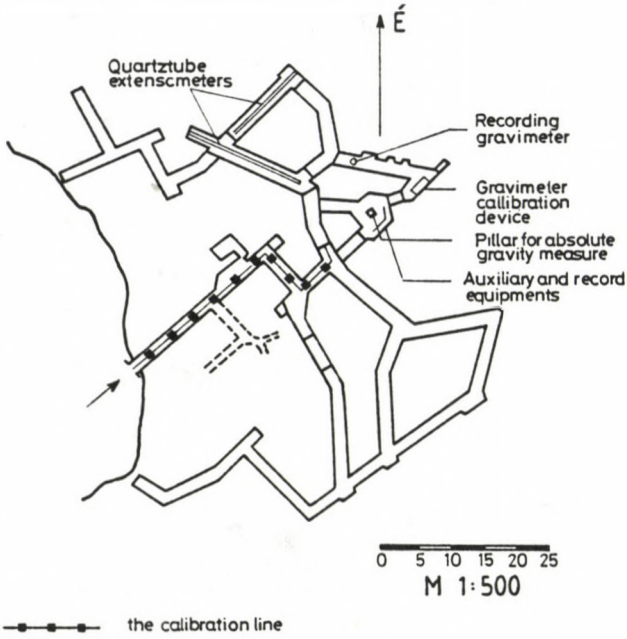

Fig. 7. Scheme of the Geodynamical Observatory Budapest

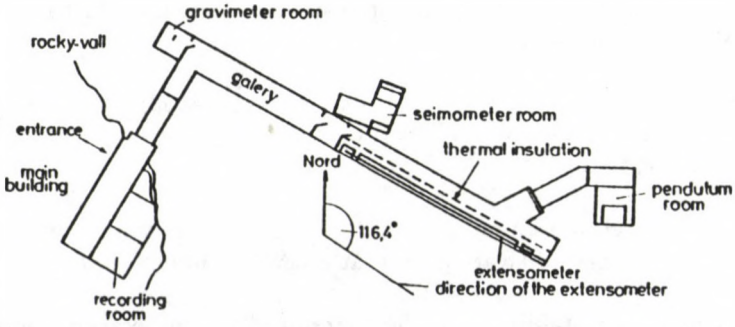

Fig. 8. Scheme of the Sopron Observatory

Table II. Earth tidal observations carried out with the 21.3 m long quartz tube extensometer at Budapest Observatory in 1984–85 according to Venedikov's (1966) method.

Wave	Amplitude ($\cdot 10^{-9}$ relative)	Phase	h/l (from measurement)	h/l (from theoretical model)
Diurnal				
O_1	5.36 ± 0.36	$-1.3° \pm 3.8°$	0.153	0.146
P_1	5.53 ± 0.33	$2.6° \pm 3.4°$	0.130	0.189
Semi-diurnal				
N_2	0.55 ± 0.19	$9.0° \pm 20.3°$	0.130	0.146
M_2	0.67 ± 0.17	$-10.3° \pm 3.7°$	0.125	0.146
S_2	1.26 ± 0.16	$-16.5° \pm 7.1°$	0.111	0.146

In the two columns on the r.h.s. h and l are the Love and Shida numbers correspondingly.

Table III. Quartz tube extensometric stations in the Pannonian Basin

Station	Latitude	Longitude	Azimuth	Length of the instrument (in m)	Reference
BeregovoI (Ukraine)	48.2°	22.7°	73°	27.5	Latynina et al. 1992
BeregovoII (Ukraine)	48.2°	22.7°	37°	11.4	Latynina et al. 1992
Vyhné (Slovakia)	48.5°	18.5°	55°	20.5	Brimich and Latynina 1988
BudapestI (Hungary)	47.6°	19.0°	114°	21.3	Varga T. et al. 1993
BudapestII (Hungary)	47.6°	19.0°	38°	13.8	Varga T. et al. 1993
Sopron (Hungary)	47.7°	16.5o	116°	22.0	Mentes 1991

be fulfilled:

— Regular (daily) local calibration of the strainmeters, for a fast recognition of possible defects in the records.

— Systematic intercalibration of the strainmeters operating at different sites. For this purpose a special mobile calibration device was developed and successfully tested by Mentes (1993).

— The meteorological data, as temperature (both in the tunnel and at the surface), pressure, humidity are needed at every strainmeter site.

— Similar long-term stability must be guaranteed in temperature (diurnal variations less than one tenth of centigrade, seasonal ones less than 1°C),

— The electronic (capacitive, inductive etc.) transducer devices must be compared with an optical read-off system of the same resolution power ($\sim 10^{-10}$) at least once, as far as possible at the beginning of the observations.

— To avoid later uncertainties all steps of the data processing must be done in the same way as far as possible with the same computer programs.

The data observed by the strainmeter network of the Pannonian Basin

The main parameters of the strainmeters used in the Pannonian Basin are listed in Table III. The network of extensometers were installed — except Pécs — during the eighties (Fig. 6). Two of the observatories (Beregovo and Budapest) are equipped with two instruments while at Vyhné and at Sopron only one instrument is running. All these stations were mounted — as mentioned earlier — in similar manner but their geological environment is quite different. Beregovo and Budapest lie in the sedimentary part of the Pannonian Basin while Vyhné and Sopron are close to the margin of the area and they are connected to the Alpine-Carpathian range.

A considerable data set was collected first of all at the stations Beregovo, Budapest and Vyhné. From Beregovo the records of both strainmeters installed (Latynina et al. 1992) were used in the interval 1986–1991 (Fig. 9); for Vyhné two time intervals were selected: 1984–1987 and 1989–1991 (Fig. 10); for both strainmeters of Budapest the data from 1990 till 1992 (Fig. 11); and from of Sopron the data of the years 1991–1993 (Fig. 12) were used. In case of all the six strainmeters hourly values expressed in micrometer were used. The scale of the record was determined on the basis of the magnetostrictive calibration device mounted into each instrument. (The intercalibration of the individual strainmeters using the calibration device developed by Mentes (1993) is the task of the nearest future.)

It should be mentioned that during the studied time intervals there were no important changes in the recording conditions. The interpretation of data obtained at the Budapest Geodynamical Observatory are somewhat complicated due to the hydrological conditions. The Danube flows two kilometers far from the station. No correlation was found between the water level of the river and the strainmeter data so far. It seems that even the moisture does not influence the strainmeters at this station. On the contrary, the level of the karstic water (measured in the Mátyáshegy cave as a level variation of a small underground lake called Agyagos) shows in this obervatory in 1992 close correlation with the residual curves of one of the extensometers installed in the azimuth N 114° (Fig. 5). This phenomenon can be of great importance in the practical use in environmental studies of the strainmeters in the future, and must be taken into account when the data are interpreted in terms of tectonics.

It is an important question with respect to the future use of the extensometers whether the aseismic and non-tidal parts of the records are of instrumental origin (so called instrumental drift) or they are caused by (local or regional) phenomena.

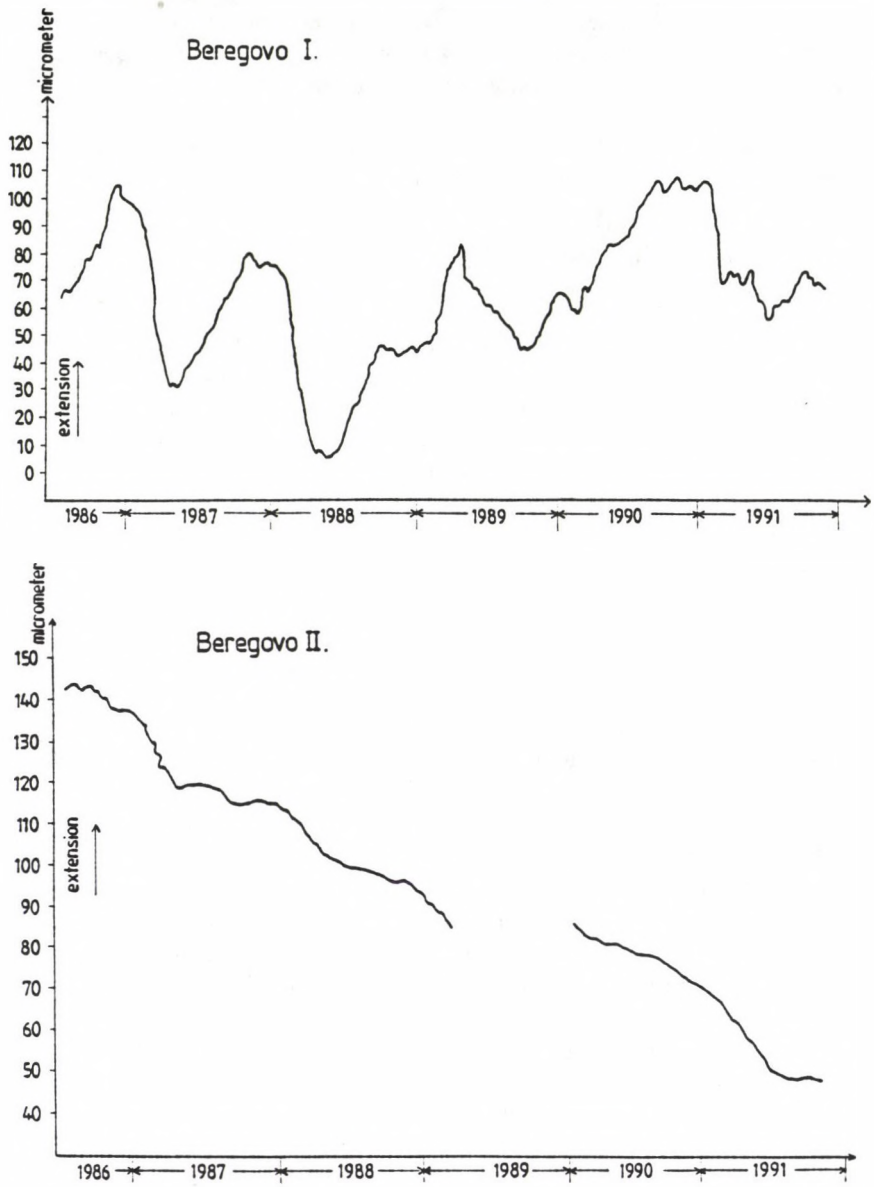

Fig. 9. Long-periodic strain variations observed at Beregovo Observatory in the time interval 1986–1991

In this respect experiments carried out with strainmeters installed parallel in the same tunnel are important (Latynina 1975a, 1975b, Latynina and Rizaeva 1976) because they show that the instrumental drift is the same or almost the same in case of them. To investigate whether similarities are of local or of regional origin in case of the stations of the Pannonian network, the amplitude spectra were calculated

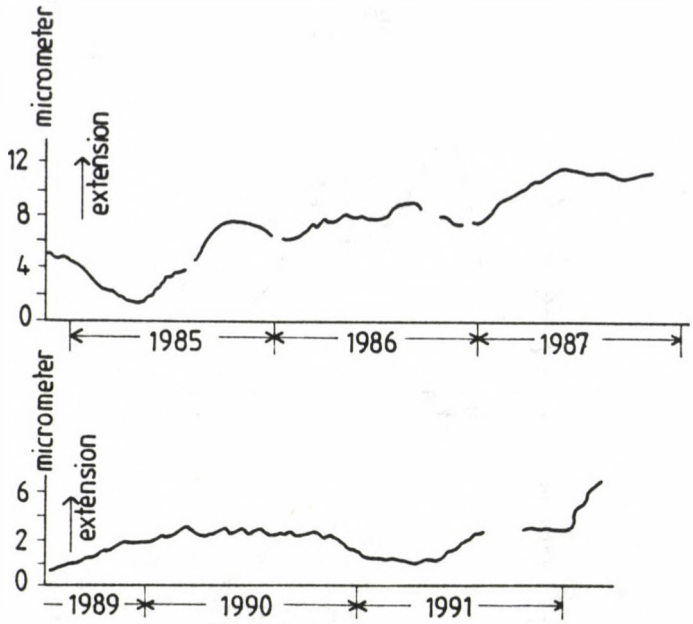

Fig. 10. Long-periodic strain variations observed at Vyhné Observatory from 1984 till 1987 and from 1989 till 1991

for the frequency band $1 \cdot 10^{-3} - 2 \cdot 10^{-2}$ cph (T = 1000 – 50 in hours). Three from the five calculated spectra are based on synchronous data sets (Sopron and the two instruments of Budapest), while two were recorded in different time-intervals (Vyhné and Beregovo).

All these spectra show that the "energy content" is significantly higher at low frequencies than in case of the higher ones. After the removal of the exponential parts of the spectra, the number of the significant frequency peaks in the spectra of the different records were compared. A spectral peak was accepted for significant if it was 3 or more times higher than the main noise level of the given spectrum. The results of this investigation (Table IV) show that the number of the coincident spectral peaks in case of the records obtained during the same time interval is considerably high: there are 13 coincidences out of 15 cases between the two strainmeters of Budapest and 9 coincidences out of 18 in case of the first three equipments listed in Table IV. The number of coincidences is probably significantly reduced in the latter case by the fact that the interval of the observations carried out in Sopron and in Budapest do not exactly coincide. It is remarkable that the amplitudes of peaks in case of these three equipments installed in different azimuths are very similar. This last circumstance shows that at the stations of the present network the cavity and the topographic effects are not extreme high in the studied frequency band. As far as the physical interpretation of the spectral peaks is concerned, the nine coinciding peaks are probably of regional meteorological origin.

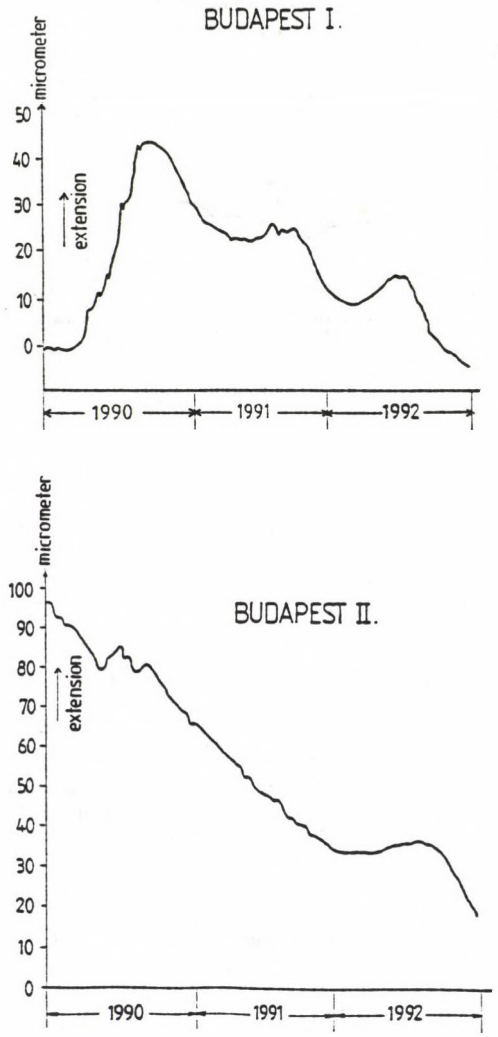

Fig. 11. Long-term variations observed at Budapest Observatory (1990–1992)

The four peaks recorded by the two strainmeters of Budapest alone — when neglecting the imperfect time-interval coincidences — are of course not of instrumental origin. They can be connected to local meteorological events or explained by the above mentioned hydrological conditions. (These frequencies are in cph: $4.5 \cdot 10^{-3}$, $13.5 \cdot 10^{-3}$, $15.5 \cdot 10^{-3}$, $19.0 \cdot 10^{-3}$). Two or three frequencies being present at the same time in all the five records ($1.5 \cdot 10^{-3}$, $3 \cdot 10^{-3}$, $10.5 \cdot 10^{-3}$ cph) are probably random coincidences or they are characteristic features of the meteorology of the Pannonian Basin in the studied time-interval 1989–1991.

The recorded data (Figs 9–12) show that in case of all stations used in this study

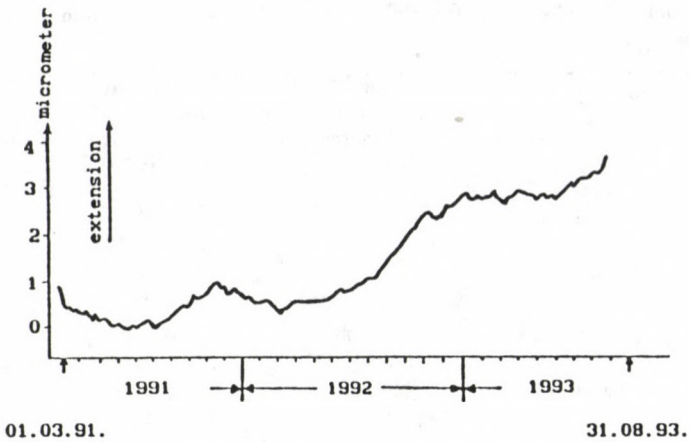

Fig. 12. Long-term variations obserwed at Sopron Observatory (1991–1993)

Table IV. Frequency peaks (in cph) recorded at stations Beregovo, Budapest, Sopron and Vyhné

Station	Epoch	Record length (in days)	Number of significant peaks	Number of significant peaks recorded with			
				two	three	four	five
				instruments			
Budapest I	01.01.91–04.09.91	247	15	13	9	3	2
Budapest II	01.01.91–04.09.91	247	15				
Sopron	11.01.91–04.10.91	177	18				
Vyhné	23.07.90–14.12.90	145	13				
Beregovo	02.08.89–28.12.89	149	12				

the long periodic variations y are mainly composed from a quasi-annual component and from a monotonous, almost linear component, what can be simply modelled as

$$y = a + bt + c \cdot \sin(\omega \cdot t + e)$$

where a, b, c and e are the arbitrary constants to be determined, t is the time in hours, ω is circular frequency, i.e. $\omega = \frac{2\pi}{T}$ (T =8766 hours = 1 year). For later consideration constants b (it describes the longperiodic "secular" strain), c (the amplitude of the yearly wave), and e (the phase shift between the observed and astronomical annual wave) are important (Table V).

The examination of the annual drift rates (b) and their geophysical interpretation is the task of the next section of this study. The seasonal strain variations can be studied with the help of the amplitude (c) and phase shift (e). The yearly wave

Table V. Annual drift and seasonal (yearly) variations observed at the strainmeter stations of the Pannonian Basin

Station	Epoch	Annual drift in micron $\cdot y^{-1}$ (and in relative unit $\cdot y^{-1}$)	Yearly wave in micron $\cdot y^{-1}$ (and in relative unit $\cdot y^{-1}$)	Phase shift of yearly wave relative to meteorological in y
Beregovo I.	1986–1991	−4.27 ($-1.55 \cdot 10^{-7}$)	13.90 ($5.05 \cdot 10^{-7}$)	0.26
Beregovo II.	1986–1991	−27.56 ($-2.41 \cdot 10^{-6}$)	4.35 ($3.82 \cdot 10^{-7}$)	0.26
Vyhné	1984–87, 1989–91	0.90 ($4.39 \cdot 10^{-8}$)	0.45 ($2.10 \cdot 10^{-8}$)	0.32
Budapest I.	1990–1992	−1.69 ($-7.93 \cdot 10^{-8}$)	10.4 ($4.88 \cdot 10^{-7}$)	0.36
Budapest II.	1990–1992	−31.42 ($-2.27 \cdot 10^{-6}$)	1.2 ($8.69 \cdot 10^{-8}$)	0.26
Sopron	1991–1993	1.51 ($6.81 \cdot 10^{-8}$)	0.4 ($1.82 \cdot 10^{-8}$)	0.12

amplitudes are small and vary between $5 \cdot 10^{-7}$ and $2 \cdot 10^{-8}$. Remarkable is that the phase shift values (e) are very similar at each station and they are ~ 0.3 years in average. The only exception is Sopron where the phase shift is less (~ 0.1 year). To interpret the influence of the seasonal strain variation two different possibilities present themselves. The first way is a study of the partial differential equation of thermo-conductivity in a special case of a flat boundary where the boundary condition at the surface ($z = 0$) is described by a harmonic wave $y(0,t) = C \cdot \sin(\omega \cdot t)$. In this way the solution can be obtained in a closed form

$$C(z,t) = C_0 \cdot e^{-(\omega/2\alpha^2)^{\frac{1}{2}} \cdot z} \cdot \cos\left(\sqrt{\frac{\omega}{2\alpha^2}} \cdot z - \omega t\right)$$

(where α is the thermal conductivity of the rocks, its typical value is $4 \cdot 10^{-3}$ cm^2 s^{-1}). This expression describes the two laws of Fourier according to which the surface amplitude C_0 decreases exponentially with the increase of the depth z, and in the rocks the surface temperature wave has a phase shift $\sqrt{\frac{\omega}{2\alpha^2}} \cdot z$. If an annual wave is taken, the corresponding frequency is $\omega = 1.9924 \cdot 10^{-7}$s^{-1}. Some solutions of the above equation (Table VI) clearly show that even if the yearly surface temperature variation is 40 °C, at a depth of 10 m an annual wave with an amplitude of $C = 0.27$°C only exists which is clearly not enough to explain the annual strain waves of $10^{-7} - 10^{-8}$ in case of silica glass tubes which have a thermal expansion coefficient of $4.5 \cdot 10^{-7}$. Another problem is connected with the phase values which are of about 0.3 years. On the basis of the data listed in Table VI this phase delay is typical at the depth of 5 m.

Another explanation of the yearly strain variation due to the surface temperature waves was given by Hvoždara and Brimich (1988). They show that thermoelastic

Table VI. Reduction of the amplitude and delay of the phase of the surface temperature variations at different depths z in case of rock thermal conductivity $\alpha = 4 \cdot 10^{-3} \mathrm{cm}^2 \mathrm{s}^{-1}$

z	Reduction of the amplitude which us unit at the (data are in relative units)	Phase delay values at different levels under the Earth's surface in years
10 m	$6.8 \cdot 10^{-3}$	−0.79
20 m	$4.6 \cdot 10^{-5}$	−1.59
30 m	$3.1 \cdot 10^{-7}$	−2.38
50 m	$1.5 \cdot 10^{-11}$	−3.97

deformations increase from the Earth's surface till a depth of 30 m, below they are constant. This explanation needs further investigation because the discussion of the problem given by Hvoždara and Brimich (1988) does not take into consideration the problem of the phase lag.

Therefore it can be supposed at this time that the quantities c and e in Table VI characterize the imperfectness of the thermoisolation of the strainmeters in the tunnel.

Geodynamical meaning of observed strainmeter data

Recently a very big progress has been achieved in the observation of geodynamical processes of a planetary scale. The mantle circulation produces shear stresses at the base of each plate. Intraplate stress sources are connected with lateral heterogeneities of the plate lithospheres (variations of crustal and lithospheric thickness, heat flow *etc.*); they were ignored earlier (Wuming et al. 1992). Motion, rotation of the plates are represented simply as rigid body motions (Gordon and Stein 1992). Analyses of magnetic anomalies over midocean ridges show that plate separation rates have a broad range between 12 mm $\cdot$ y^{-1} (Atlantic Ridge) and 160 mm$\cdot$ y^{-1} (East Pacific) over the past few million years. Convergence rates are of the same order. Therefore it can be calculated with rates of some ten mm $\cdot$ y^{-1}. The techniques of space geodesy (Very Long Baseline radio Interferometry – VLBI, Satellite Laser Ranging – SLR, the Global Positioning System – GPS) allow to detect and investigate plate velocities of this order averaged over a few years (e.g. in case of GPS measurements, it must be continued for at least a decade to reach a rate accuracy of 1 mm $\cdot$ y^{-1} at the 95 percent confidence level). It is important to mention that the velocities are similar to those averaged over million years. In this way, however, the steady plate motions can only be described, and there is no possibility to detect the intraplate displacements, tectonical processes, preseismic activities *etc.* which are by many orders of magnitude less than the rates of the plate movements.

For the study of intraplate processes highly sensitive tilt- and strainmeters seem to be of great importance. It was, however, shown earlier by Harrison (1976) that

in general each geometrical and material irregularity can produce distortions of the size of the signals on the records. Cave deformations of the location of the recording instruments and the effect of the surface topography should be first of all mentioned in this respect. Small cracks and other material inhomogeneities may produce large local effects. The distortion effect of caves, topography, local inhomogeneities, cracks is strongly influenced by the shape of the strain tensor to be observed (Emter and Zürn 1985).

The distortion influence is less in case of strainmeters as in case of tiltmeters and it can be further reduced if the instrument is installed near to the axis of the tunnel. It must not be forgotten that the longitudinal strain in the tunnel is magnified near its ends and most of the anomalous strain is concentrated within a distance of one tunnel diameter taken from the end of the tunnel (Harrison 1976).

Taking this into consideration one can make an attempt to interpret the intraplate deformational processes. Emter (1989) showd that drift rates of globally distributed highly sensitive tiltmeters are in most cases at least by one order of magnitude higher than tectonically reasonable tilt rates. This statement — probably in less degree — seems to be valid for most of the strain measurements, too. According to the authors' opinion the uncertainties in strain data interpretation can be reduced by a station system operating in the same manner within a regional tectonic unit. Varga (1984) determined the mean annual drift in planetary scale on the basis of 28 extensometric stations and got as result: $2.1 \cdot 10^{-6}$ y $^{-1}$. The question is whether this strain rate is big or not. According to Kasahara (1958) the maximum elastic deformation can be given by the ratio

$$\epsilon_{max} = P \cdot \mu^{-1}$$

where P is the strength of the rocks $(10^6 - 10^7)\mathrm{N/m}^2$ and μ is the shear or rigidity coefficient $(3 - 5) \cdot 10^{10}\mathrm{N/m}^2$. Therefore $\epsilon_{max} = 10^{-4} - 10^{-5}$. The value of μ is well known from seismology. The objectivity of the P values are somewhat more disputable because above value is the result of laboratory tests. It is known, however, that the stress drops in the largest earthquake foci (e.g. San Francisco, 1906, Izdu, 1930) do not exceed 10^7 N/m^2 (100 bars). The analysis of the isostatic anomalies gives a similar value, too, the isostatic anomalies are usually $\leq$ 20 mgl and in young orogenic areas $\leq$ 50 mgl, i.e. $\leq$ 125 bar $= 1.25 \cdot 10^{-7}$ N/m^2 (1 mgl corresponds in this case to 2.5 bar). The same value (10^7 N/m^2) can be estimated from heights of the geoid anomalies (about 50–100 m) and of the mean surface density. In a seismically active area the process of strain accumulation $t \sim 10^2$ years (y) (an estimated time interval between two earthquakes). In case of smooth and gradual stress, the accumulation of maximum strain rate pro year:

$$\epsilon_{max}/t = \frac{10^{-4}}{10^2 y} - \frac{10^{-5}}{10^2 y} = (10^{-6} - 10^{-7}) \cdot y^{-1}$$

can be obtained.

Taking into consideration the problems of strainmeter data interpretation described in the first part of this section, as well as the above obtained annual strain

rate values, some conclusions can be obtained on the Pannonian extensometric network data. As it is well known, the Pannonian Basin is characterised by a thin crust (25 km), by a pronouncedly high heat flow [(80 – 130) mW · m^{-2}], by a temperature gradient less than 50°C km^{-1}, by a low density of the upper mantle and by a moderate seismicity (one earthquake in a year with the intensity of I_0 = 4 MSK, two earthquakes with the intensity I_0 = 8 MSK during a century). It is worth mentioning that the earthquake activity of the basin margins is about 10 times higher than that of the central part.

Concerning the areal distribution and the geological conditions of the stations (Fig. 6), Beregovo and Budapest are located in the sedimentary part of the Pannonian Basin, while Vyhné and Sopron lie close to the margin of the basin and are connected with the Alpine and Carpathian ranges. The comparison of strain and stress data, being not necessarily connected to each other in a simple way, can give useful new information. Unfortunately there are for strains no detailed data of the regional distribution at disposal as for the stress (Zoback et al. 1989). In the central sedimentary part of the Pannonian Basin extension could not be detected in the strain. The same conclusion was obtained by the worldwide cooperation of stress measurements for intraplate regions. Small extension values were observed near the margins of the depressions (Vyhné $4.39 \cdot 10^{-8}$ and Sopron $6.81 \cdot 10^{-8}$) which can be connected with the extensional tectonism being worldwide typical for mountainous areas (Zoback et al. 1989). Because the orientation of the strainmeters is determined by the orientation of the cave where the instruments are installed, it is irrealistic to speak about the disturbation of maximum strain directions. Nevertheless, the maximum (compressional) strain detected in the Pannonian Basin is almost a northward oriented compression (Table IV): in Beregovo (27.6 micron $\cdot$ y^{-1} or $2.4 \cdot 10^{-6}$ y^{-1}) and in Budapest (31.4 micron $\cdot$ y^{-1} or $2.3 \cdot 10^{-6}$ y^{-1}). This orientation is not far from the main NW–SE maximum stress direction of Europe (Müller et al. 1992) which was also detected in stress observations in the Pannonian Basin (Dövényi and Horváth 1990, Grünthal and Strohmayer 1992). Strains (mainly compressions) observed in other directions are by one or two orders of magnitude less (10^{-7} $y^{-1} - 10^{-8}$ y^{-1})) (for orientation: the daily variation of the strain tides of the solid Earth is 10^{-8}).

The maximum elastic deformation $\epsilon_{max} = 10^{-4} - 10^{-5}$ is reached in N-S direction within 100 years in case of Beregovo II and Budapest II; in the sedimentary part of the Pannonian Basin it is softened by the high heat flow which significantly reduces the stress accumulation. In other places and directions of the network this critical value can be reached within $10^3 - 10^5$ years. It is possible that this slow strain accumulation is connected with the low seismic activity of the brittle marginal parts of the Pannonian Basin where the stations Vyhné and Sopron are installed. The elastic stress magnitudes can be estimated on the basis of strain measurements. If the elastic axial deformations plotted in the Cartesian coordinate system are connected with the corresponding axial stresses $\{\sigma_i\}$ according to Hooke's law as

$$\{\epsilon_i\} = [D] \cdot \{\sigma_i\}$$

where $i = x, y$ or z, and

$$[D] = (1 - \nu - 2\,\nu^2)^{-1} \begin{bmatrix} E(1-\nu) & E\nu & E\nu \\ E\nu & E(1-\nu) & E\nu \\ E\nu & E\nu & E(1-\nu) \end{bmatrix}$$

where E is Young's module (expressed in 10^{-11} Pa) and ν is Poisson's dimensionless ratio. If we consider the deformations e.g. along the axis x, then:

$$\epsilon_x = E(1 - \ \nu - 2\,\nu^2)^{-1}\,[(1-\nu)\sigma_x + \nu\,\sigma_y + \nu\,\sigma_z]\ .$$

It is usually supposed that the flat surface of the Earth is free of normal stresses ($\sigma_z = 0$) and the medium is isotropic in the horizontal direction ($\sigma_x = \sigma_y$). Thus:

$$\epsilon_x = E(1 - v - 2v^2)^{-1}\,[\sigma_x] = \frac{E\sigma_x}{1 - v - 2v^2}\ .$$

Since the measurements of strain variations ($\Delta\ \epsilon_{max}$) with strainmeters are based on the above equation, the corresponding stress variation ($\Delta\sigma_x$) can be obtained from:

$$\Delta\sigma_x = E(1 - \nu - 2\nu^2)^{-1}\Delta\epsilon_x\ .$$

Let be $E = 10^{11}$ Pa, $\nu = 0.25$. From strainmeter measurements for $\Delta\epsilon_x$ the annual rate $(10^{-6} - 10^{-8}) \cdot \mathrm{y}^{-1}$ was obtained. For these values the corresponding annual stress rates are $\Delta\sigma_x = 6.25(10^4 - 10)$ Pa. Becker (1993) estimated from *in situ* stress measurements stresses of the order of 10^6 Pa (with different sign and significance) for different sites in the Pannonian Basin which can accumulate according to the observed annual strain rate data during $10^2 - 10^6$ years if the strain is gradual and smooth. Thus is a rather strong condition, and it probably shows that the stress magnitude values obtained from *in situ* measurements are overestimated.

On the basis of investigations reported here it is concluded that

— strainmeter values are not in contradiction with the stress measurements as far as the directions are concerned, their typical value is of the order $10^{-7} - 10^{-8}$.

— stress values derived from strainmeter data are less than 10^5 Pa, the magnitude of the observed stress seems to be overestimated,

— a further development of the strainmeters would be necessary with the aim to simplify the installation of such equipments in smaller cavities; in this way there would be more chance to obtain the maximum strain values,

— the study of hydrological conditions is important for the improvement of the reliability of the strainmeter data,

— for more accurate studies of strains, the extensometers must be intercalibrated in the future.

Acknowledgement

This work was supported by the Hungarian Science Fund OTKA (Project No 2131).

References

Agnew D C 1986: *Reviews of Geophysics*, 24, No.3, 579–624.

Asada T ed. 1982: Earthquake Prediction Techniques. Their Application in Japan. University of Tokyo Press

Becker A 1993: *Terra Nova*, 5, No.4, 375–388.

Brimich L, Latynina L A 1988: *J. Phys. Earth.*, No.12, 3–9 (in Russian).

Dövényi P, Horváth F 1990: *Acta Geod. Geoph. Mont. Hung.*, 25, 257–266.

Emter D 1989: In: Rezente Krustbewegungen Seminar, Heft 39, Studiengang Vermessungswesen. Univ. der Bundeswehr, 125–139.

Emter D, Zürn W 1985: In: Earth Tides, Van Nostrand R, co ed. Harrison J C, 309–327.

Gordon R G, Stein S 1992: *Science*, 256, 333–342.

Groten E 1991: *J. Geodynamics*, 14, 1–7.

Grünthal G, Strohmayer D 1992: *J. Geoph. Res.*, 97, No.B8, 11805–11820.

Harrison J C 1976: *J. Geoph. Res.*, 81, 319–328.

Hvoždara M, Brimich L 1988: *Studia geoph. et geod.*, 32, 129–135.

Kasahara K 1958: *J. Phys. Earth*, 6, 15–22.

Kohlbeck F, Rock K-H, Scheidegger A E 1980: *Rock. Mech.*, Suppl. 9, 21–29.

Kurskeev A K 1990: Problems in Earthquake prediction, Nauka KSSR, Alma–Ata (in Russian)

Latynina L A 1975a: *Tectonophysics*, 29, 421–427.

Latynina L A 1975b: *J. Phys. Earth.*, 3, 16–26.

Latynina L A, Karmaleeva R M 1978: Deformographic measurements. Nauka, Moscow (in Russian).

Latynina L A, Rizaeva S D 1976: *Tectonophysics*, 31, 121–127.

Latynina L A, Szabó Gy, Varga P 1984: *Acta Geod. Geoph. Mont. Hung.*, 19, 197–205.

Latynina L A, Jurkevits O I, Baisarovits I M 1992: *Journal of Geophysics*, 14, No.2, 63–66.

Mentes Gy 1981: *Acta Geod. Geoph. Mont. Hung.*, 16, 269–280.

Mentes Gy 1991: *Mareés Terrestres, Bull. d'Information*, 110, 7936–7939.

Mentes Gy 1993: *Mareés Terrestres, Bull. d'Information*, 115, 8467–8471.

Mogi K 1985: Earthquake prediction. Academic Press

Müller B et al. 1992: *J. Geoph. Res.*, 97, 11783–11803.

Pertzev B P, Ivanova M V 1991: *J. Phys. Earth.*, No.10, 84–88 (in Russian).

Peška P 1993: *Studia geoph. et geod.*, 37, 141–149.

Rikitake T 1976: Earthquake prediction. Developments in Solid Earth Geophysics 9, Elsevier Scientific Publishing Co.

Udias A, Buforn E, Ruiz de Guanu J 1989: Catalogue of focal mechanisms of European earthquakes. Dept. of Geophysics, Universidad Complatense, Madrid

Varga P 1976: *Mareés Terrestres, Bull. d'Information*, 74, 4319–4332.

Varga P 1984: *Journal of Geophysics*, 55, No.1, 68–72.

Varga T, Latynina L A, Brimich L, Mentes Gy, Katona Gy, Varga P 1993: *Mareés Terrestres, Bull. d'Information*, 116, 8537–8545.

Venedikov A P 1966: *Comm. Obs. Royal Belgique*, No. 250, Ser. Geoph., 71, 463–485.

Wenczel H-G 1991: Earth tide analysis program systems ETERNA "Readme.1.st" of the distributed floppy.

Wilmes H 1983: *Deutsche Geodätische Kommission*, Reihe C, Heft Nr.281.

Wuming B, Vinick R, Froidevaux C 1992: *J. Geoph. Res.*, 97, 11729–11737.

Zoback M L 1992: *J. Geoph. Res.*, 97, 11703–11728.

Zoback M L, Zobak M D, Adams J, Assaumpcão M, Bell S, Bergman E A, Blümling P, Brereton N R, Dehant D, Ding J, Fuchs K, Gay N, Gregerson S, Gupta H K, Gvishiani A, Jacob K, Klein R, Knoll P, Magee M, Mercier J L, Müller B C, Paquin C, Stephansson O, Suarez G, Suter M, Udias A, Xu Z H, Zhizhin M 1989: *Nature*, 341, 291–297.

Acta Geod. Geoph. Hung., Vol. 29(1–2), pp. 81–105 (1994)

A THIN SHEET NUMERICAL STUDY OF THE ELECTROMAGNETIC FIELD OVER GEOMETRICALLY COMPLEX HIGH CONDUCTIVITY STRUCTURES: THE FIELD COMPONENTS AND THEIR RELATION WITH SOME 3-D MT INTERPRETATION PARAMETERS

L SZARKA[1], M MENVIELLE[2], P TARITS[3], A ÁDÁM[1]

[Manuscript received February 15, 1994]

The present paper is the first of a series of papers dealing with studies of the electromagnetic field over geometrically complex models. We used thin sheet modelling to describe the behaviour of the EM field in 3-D situations, and investigate the behaviour of parameters used in EM studies. The models are high-conductivity crustal three-dimensional thin structures having different geometrical complexity in a homogeneous lithosphere, underlain by a deep high-conductivity asthenosphere.

In this paper, we study the morphological characteristics of the five electromagnetic field components over structures of complex geometry. We also consider normalized quantities of common use in the analysis of electromagnetic investigations results: impedance components, tipper elements, tensor invariants c_1, c_2, and c_3 (where $c_1 = Z_{xx} + Z_{yy}$, $c_2 = Z_{xy} - Z_{yx}$, and $c_3 = Z_{xy} \cdot Z_{yx} - Z_{xx} \cdot Z_{yy}$), skew, and ellipticity. The five electromagnetic field components are mapped over three models having different geometrical complexity, and the main features of the behaviour of each component are outlined. The results of the numerical computations also show that the morphology of the normalized quantities we study, can be approximated by that of the combination of the electric components. The areal distribution of the electric component along the direction perpendicular to the inducing magnetic field gives the main characteristics of the antidiagonal elements of the impedance tensor, while the other electric component describes the diagonal elements of the impedance tensor.

Finally, it is pointed out that with increasing model complexity (1) the skew and the c_1 invariant (which indicate the corners of the structures) are getting worse and worse; (2) the Berdichevsky invariant (c_3) seems to better approximate the contour of the conducting structure than over simple rectangular models; (3) the c_2 invariant remains a reliable 'side' indicator; (4) ellipticity maps give useful information neither over relatively simple nor more complex models, because the denominator may be close to zero. Skew and ellipticity parameters thus appear to be useless to depict the geometry of 3-D structures.

Keywords: magnetotellurics; numerical modelling; skew; thin sheet; three-dimensional models

[1]Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences, H–9401 Sopron, POB 5, Hungary

[2]Laboratoire de Physique de la Terre et des Planetes (URA 1369), Bat. 504 – Université Paris Sud, F–91405 Orsay Cedex, France

[3]Université de Bretagne Occidentale Laboratoire de Geosciences Marines (URA 1278), 6 Avenue Le Gorgeu, F–29200 Brest, France

Introduction

Since the beginning of magnetotellurics, different MT parameters have been introduced with the aim of helping in the interpretation of electromagnetic soundings. For example, the characteristics of apparent resistivity (ϱ_a) and phase (φ) sounding curves were discussed early by Cagniard (1953) for structures, the conductivity of which only varies with depth (one-dimensional situations, 1-D).

When the structure extends towards infinity in one horizontal direction (two-dimensional situations, 2-D), not only sounding curves but also profiles and pseudosections were studied in E- and H-polarizations by many authors. Berdichevsky and Dmitriev (1976) made one of the first steps towards a qualitative study of the three-dimensional (3-D) distortions.

In the general 3-D situation, the electromagnetic field cannot be entirely described with two characteristic profiles or pseudo-sections. The 3-D problem is more complex, and areal maps should be then taken into consideration.

Before the availability of effective computer programs it was impossible to make any systematic study over 3-D structures. So far only a few studies of the electromagnetic field over 3-D structures have been published, mainly from the early stage of 3-D modelling most of them being reviewed by Vozoff (1986, 1991). Furthermore, the aim of most of these papers was more to demonstrate and justify algorithms than to give a complete description of 3-D anomalies. Some papers also addressing a particular feature of the behaviour of the electromagnetic field over 3-D structures have been published (e.g. Weidelt 1975, Reddy et al. 1977, Ting and Hohmann 1981, Wannamaker et al. 1984, Park 1985).

The thin sheet approximation allows the design of very efficient numerical modelling methods, offering almost a unique tool to study morphologic characteristics of the electromagnetic field over 3-D structures with small vertical over horizontal dimension ratio. Although thin sheet approximation deals with a limited class of structures, it fits in practice with a large number of geophysical situations, and reliable answers to a great number of questions encountered in exploration and fundamental geophysics can be derived from numerical investigations with thin sheet modelling programs.

The first bimodal thin sheet algorithm was proposed by Vasseur and Weidelt (1977), and it allowed to compute the electromagnetic field for a superficial thin sheet with an heterogeneous structure of finite extent overlying a stratified substratum. Algorithms dealing with the computation of the electromagnetic field for superficial thin heterogeneous structures extending towards infinity and overlying a stratified substratum (Dawson et al. 1982), or a 2-D substratum have been proposed afterwards.

Tarits (1989) made Vasseur and Weidelt's program operational for thin sheet with heterogeneous structures of finite extent embedded into a stratified medium (Weidelt (1987, personal communication) made the same). This program also allows to determine field components either at the surface, or at arbitrary depths, and therefore permits modelling of seafloor, mines, and drill holes situations. It is in fact a very powerful tool for investigating the behaviour of the electromagnetic field

in a large number of geophysical situations. We therefore used the Vasseur and Weidelt (1977) and Tarits (1989) algorithms to investigate the behaviour of the electromagnetic field components, and that of various parameters introduced for interpreting electromagnetic observations in 3-D situations.

The models we used have been designed for the results to be relevant for upper crustal electromagnetic studies. Due to the electromagnetic scaling rules, our results can be generalized for other situations, e.g. for the lower crust and upper mantle, or for very near-surface explorations.

It is worth noticing here that the related computations can be made on personal computers, thus providing low cost efficient tools for data interpretation.

In this study we thus discuss the behaviour of electromagnetic field components and the use of different interpretation parameters in certain 3-D situations. Performing such interpretation requires a clear knowledge of the links between the observations — field components or MT interpretation parameters — and the main geometrical and electrical features of the underlying conductivity heterogeneities. In our knowledge, although the electromagnetic field components are always computed in the process of numerical modelling, a full analysis of their behaviour has never been published. We show that the component mapping provides a quick-look and accurate sketch of the main geometrical features of the underlying conductive structures, that is proved to be very useful in the very first stage of the interpretation.

In fact, various parameters (e.g. impedance, tipper, *etc.*), normalized by the inducing field are generally used in the interpretation process instead of the field components themselves. Our modelling results allow us to discuss the efficiency of some normalized parameters of common usage, and to select the more powerful ones in terms of geometrical description of the structure.

We first present the numerical models we use, and recall the basic formulae and definitions. We then present maps of the electromagnetic components (E_x, E_y, H_x, H_y, and H_z), and of some normalized parameters over three 3-D structures. We discuss their main morphological features, and the relations that exist between the areal distribution of different normalized parameters and that of the electric field components. It will be demonstrated that skew and ellipticity are not acceptable interpretation parameters.

The numerical models

The selected models reflect one of possible crustal conductivity structure types in the Pannonian Basin (Ádám 1992). High-conductivity formations (dykes, layer-pieces, *etc.*) may be observed in a resistive lithosphere which is underlain here by a high conductivity asthenosphere at a depth of approximately 70 km.

We considered the case of conductive structures at a depth of 4 km in a stratified medium (Fig. 1). We computed the electromagnetic field for three models, corresponding to increasing geometrical complexity (see also Fig. 1):

1. a rectangular structure (called S), having two perpendicular symmetry axes and consequently a relatively simple 3-D anomaly

2. an L-shaped structure (called L), having one symmetry axis and a more complex 3-D anomaly
3. a non symmetric U shaped structure (called Y), having no symmetry axis and consequently the most complex 3-D anomaly.

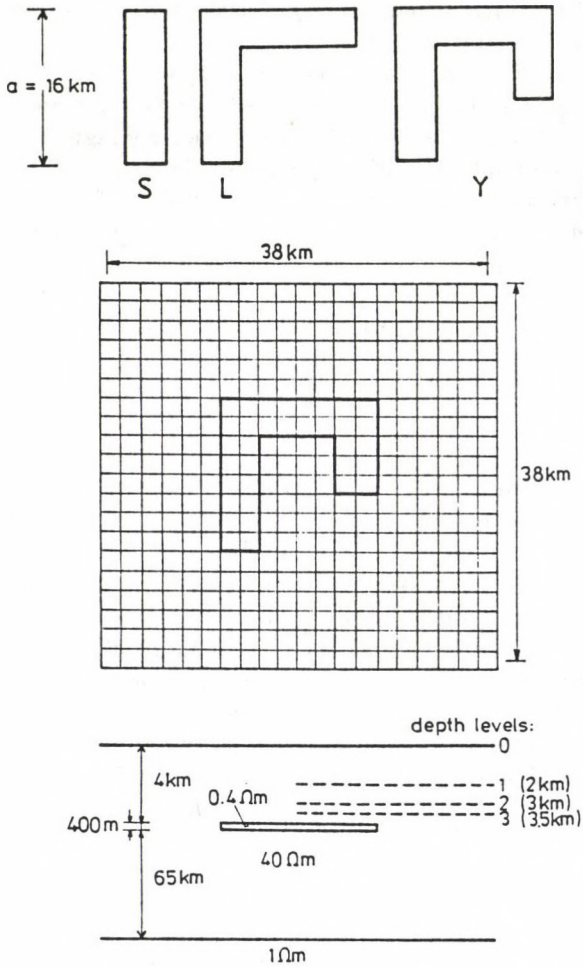

Fig. 1. Summary of thin sheet model parameters. a) Plan views of 3-D thin sheet models used in these studies. E, S, U and T denote different rectangular models; L and Y denote more complex shapes. b) Plan view of the thin sheet computation grid and a characteristic model cross section showing all resistivity and geometric parameters

The electromagnetic field was modelled over an area of 40 km × 40 km using a 20 × 20 mesh of 2 km × 2 km cells (see Fig. 1), and it was computed at four different depths: at the surface (level 0), at a depth of 2 km (level 1), at a depth of 3 km (level 2), and at a depth of 3.5 km (level 3). Five different, logarithmically approximatively

equidistant periods covering the range from the short period thin sheet limitation to the period corresponding to a dominant appearence of the asthenosphere were used: 14 s, 56.25 s, 225 s, 900 s and 3600 s.

Notations and basic formulae

For the sake of brevity the electromagnetic field components will be discussed hereafter in two groups: 1. the 'main components' denoted by capital letters, 'E' and 'H' (H is parallel to the direction of the inducing magnetic field; E is perpendicular to it); 2. the 'additional components' denoted by lower case letters, 'e', 'h', and 'z' (h is perpendicular to the inducing magnetic field; e is parallel to it). E and e stand for the horizontal electric field components; H and h stand for the horizontal magnetic field components. z stands for the vertical magnetic field component, which is considered as additional component. The horizontal plane is referred to a xy frame where the x axis points towards North, and the y axis points towards East. When the inducing magnetic field is along the x axis, the main components are E_y and H_x, while the additional components are e_x, h_y and z_0. When the inducing magnetic field is along the y axis, the main components are E_x and H_y, while the additional components are e_y, h_x and z_{90} (see also Table II). Using for the horizontal components the same notations as in Table I, the impedance tensor elements are given by:

$$Z_{xx} = \frac{e_x H_y - E_x h_y}{H_x H_y - h_x h_y} \tag{1}$$

$$Z_{xy} = \frac{E_x H_x - e_x h_x}{H_x H_y - h_x h_y} \tag{2}$$

$$Z_{yx} = \frac{E_y H_y - e_y h_y}{H_x H_y - h_x h_y} \tag{3}$$

$$Z_{yy} = \frac{e_y H_x - E_y h_x}{H_x H_y - h_x h_y} \tag{4}$$

Similarly, the tipper elements are given by:

$$Z_{zx} = \frac{z_0 H_y - z_{90} h_y}{H_x H_y - h_x h_y} \tag{5}$$

$$Z_{zy} = \frac{z_{90} H_x - z_0 h_x}{H_x H_y - h_x h_y} \tag{6}$$

Based on numerical modelling results, it will be demonstrated that these expressions can be simplified in the situations we consider.

The morphology of field components over three-dimensional structures

We only present here the maps for one period (T = 225 s), because they provide a relevant illustration of the conclusions drawn from the whole set of maps we

Table I. Main and additional electric and magnetic components as a function of the inducing magnetic field direction. In the coordinate system x points towards North, (y) points towards Earth

Clockwise from the North direction of the inducing H-field	Field components				
	E	e	H	h	z
0°	E_y	e_x	H_x	h_y	z_0
90°	E_x	e_y	H_y	h_x	z_{90}

Table II. Numerical values of the two parts in the nominator of the impedance tensor element for model L. The dominating terms are printed in bold

Z_{xx}	$\mathbf{e_x H_y}$	$E_x h_y$
	$10^{-5} \cdot 1$	$10^{-4} \cdot 10^{-3}$
Z_{xy}	$e_x h_x$	$\mathbf{E_x H_x}$
	$10^{-5} \cdot 10^{-3}$	$10^{-4} \cdot 1$
Z_{yx}	$\mathbf{E_y H_y}$	$e_y h_y$
	$10^{-4} \cdot 1$	$10^{-5} \cdot 10^{-3}$
Z_{yy}	$E_y h_x$	$\mathbf{e_y H_x}$
	$10^{-4} \cdot 10^{-3}$	$10^{-5} \cdot 1$

produced. The inducing magnetic field has a unit intensity and zero phase shift. (In this meaning all other field components can be regarded as normalized quantities.) Maps computed at different depths confirm all conclusions presented here but in this paper only surface field values will be shown.

General features of the behaviour of the components

Figure 2 shows 20 field component maps (real and imaginary parts of five field components in the two perpendicular inducing field directions) over the rectangular model. Figures 3 and 4 show similar maps for the L- and U-shaped models respectively. (It should be mentioned that in the plots the isoline values are not everywhere denoted. It helps better understanding of these figures that the "main" components have more symmetric and the "additional" components have more antisymmetric appearance.)

First of all, these maps show that the observed areal distribution of real and imaginary parts of each component are always similar to each other. This follows

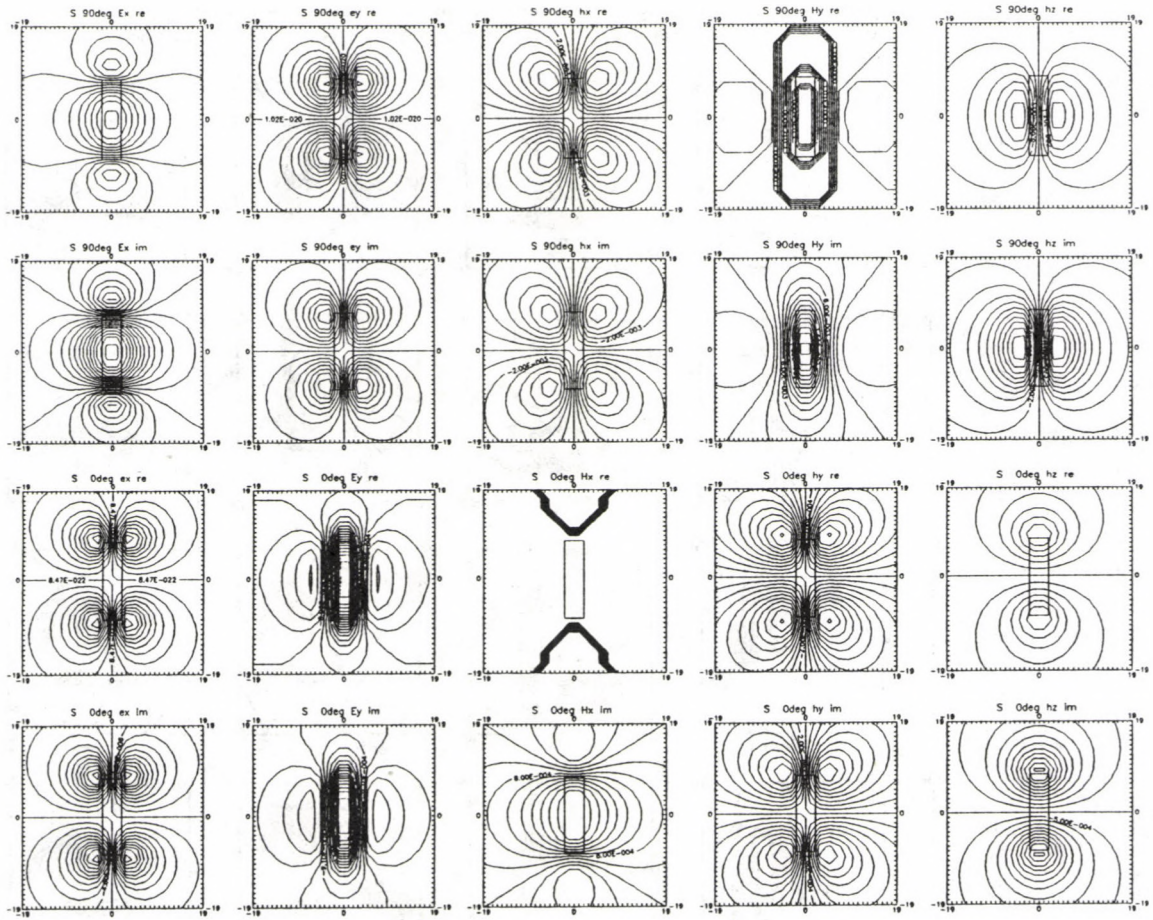

Fig. 2. Real and imaginary parts of electric, horizontal and vertical magnetic field components in two ($\Theta = 90°$ and $\Theta = 0°$) polarizations over model 1 (model name: S) at a period of 225 s

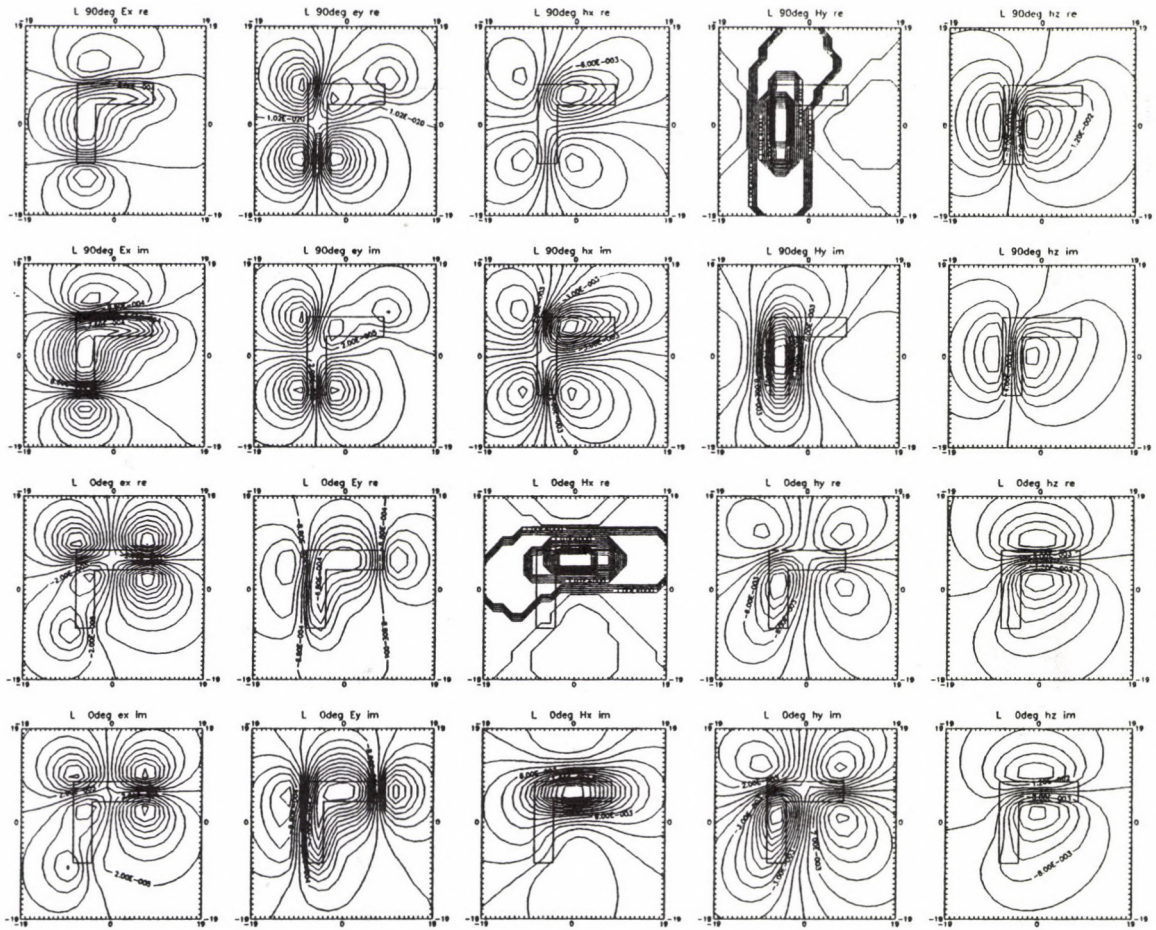

Fig. 3. Real and imaginary parts of horizontal electric, horizontal and vertical magnetic field components in two ($\Theta = 90°$ and $\Theta = 0°$) polarizations over model 2 (model name: L) at a period of 225 s

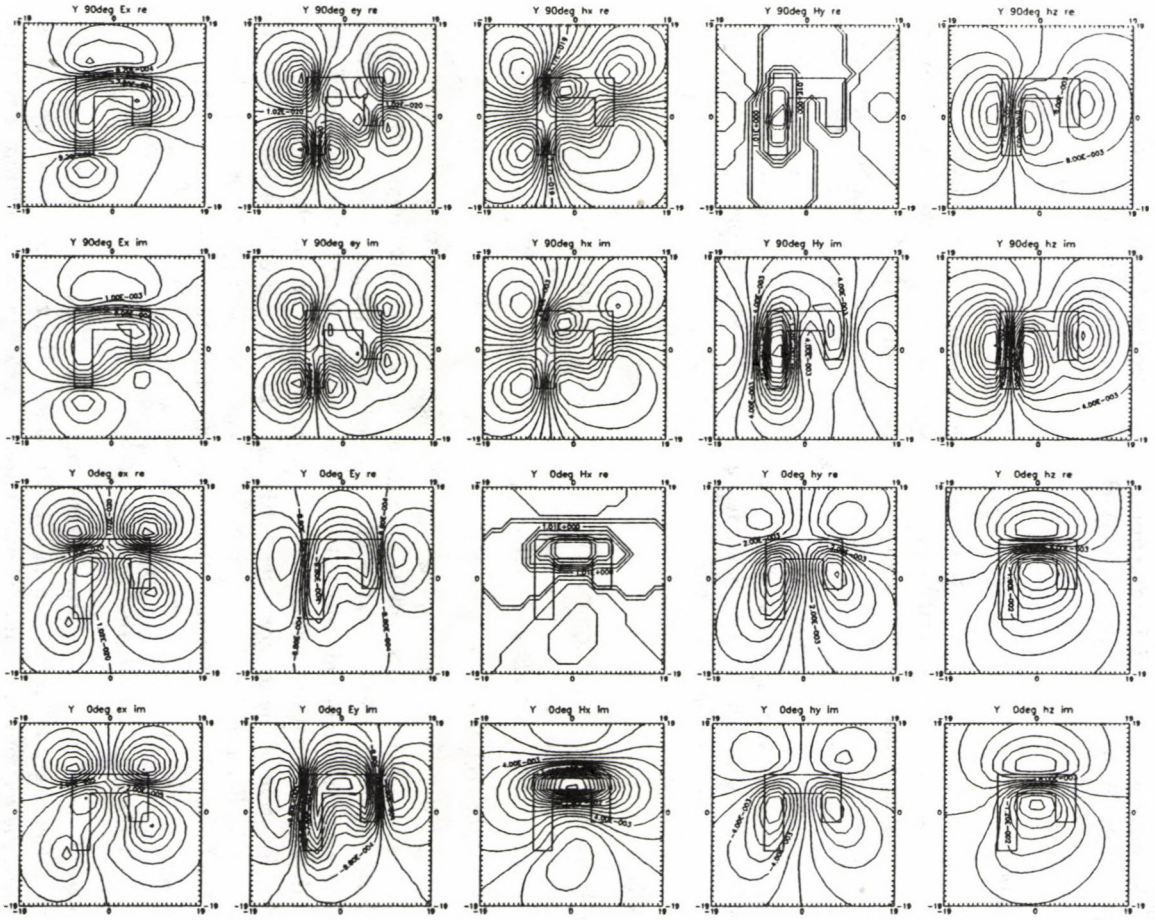

Fig. 4. Real and imaginary parts of horizontal electric, horizontal and vertical magnetic field components in two ($\Theta = 90°$ and $\Theta = 0°$) polarizations over model 3 (model name: Y) at a period of 225 s

from the fact that the inductive effects are negligible in the conductive structures for the models we consider. The deflection of the electric currents by the conductivity heterogeneities is then mostly driven by galvanic effects, and each field component is the product of a position-dependent term by a period-dependent term (e.g. Le Mouel and Menvielle 1982, Menvielle 1987). The behaviour of the electromagnetic field at the surface of the model corresponds to the already described current channelling, or static distortion (Larsen 1975) effects. A review of the distortion effects can be found in Menvielle (1987).

E, H, and z components

The E, H and perhaps the z maps over the rectangular model (Fig. 2) are fairly well known. E is characterized by a central minimum and two 'side' maxima, H by a central maximum and by two 'side' minima, and z by two antisymmetric anomalies at the edges of the structure whose direction is perpendicular to that of the inducing magnetic field (the 'E-polarization' edges). Besides, the E and H maps show the existence of a zone, between the central anomaly and the side anomalies, where the E and H values do not differ very much from their undistorted values. Figures 3 and 4 show that these basic features can be still recognized, although distorted, over more complex structures.

e and h components

The behaviour of e and h components reminds to the main components, but the additional components are more severely distorted over more complex structures than the main components E and H.

e and h over the rectangular model

The e and h components appear over the rectangular model just outside its corners. They are characterized by two positive and two negative peaks, depending on the inducing field direction. When E is parallel to the longer sides of the model, the extrema of e are definitely shifted from the corners towards the middle of the longer sides. This situation is due to the more important current channelling effect along the direction of the longer sides of the model than along the other direction. Such a sharp difference for the extrema of h cannot be seen; they seem to be at corners for both inducing field directions. As expected, the e and h zero lines (or zones) coincide with the symmetry axes.

e and h over the L- and U-shaped models

Figures 3 and 4 show two e and h zero zones (there are still two of them, but they are not as sharp as those seen in Fig. 2). One of the zero zones can be always observed in the sections of the model which are elongated in the direction of E. The center line of this zone connects the E and H side anomalies through the central anomaly.

e and h extrema do not change in the same way with increasing model complexity. In the e-maps one can still see the peaks corresponding to those present

on the rectangular model, but one or more further extrema will appear. They are still developed at corners and model bends. It is remarkable that the 'new' extrema appear almost exactly at corners. A visible shift towards the middle of the longer sides can only be seen if the model section is definitely elongated in the direction of the actual E. 'Inner' extrema may also appear over model bends.

In the h-maps the 'old' and 'new' extrema merge into each other.

Relation between the electromagnetic field components and the normalized parameters

For the period range and the models we consider, the integrated density of the additional currents is negligible compared to that of the main currents. h is therefore expected to be negligible compared to H, as it is found in the numerical results. This implies $eh \ll EH$, $Eh \ll eH$ and $h_x h_y \ll H_x H_y$ (see Table II for the numerical values). The expressions of the impedance tensor elements can be then reduced to the following simplified expressions, which are in practice valid in many geophysical situations, and then often used without any numerical justification:

$$Z_{xx} \sim e_x/H_x \tag{7}$$

$$Z_{xy} \sim E_x/H_y \tag{8}$$

$$Z_{yx} \sim E_y/H_x \tag{9}$$

$$Z_{yy} \sim e_y/H_y \tag{10}$$

The expressions of the tipper elements similarly reduce to:

$$Z_{zx} \sim z_0/H_x \tag{11}$$

$$Z_{zy} \sim z_{90}/H_y \tag{12}$$

The areal variations of H are much less than those of E and e (see Figs 2, 3 and 4). The areal distribution (the "morphology") of the impedance elements is therefore similar to that of the electric components. The areal distribution of normalized parameters over 3-D structures can therefore be accounted for in terms of electric field components. This is illustrated by a comparison between the corresponding maps in Figs 5, 6 and 7 on the one hand, and in Figs 2, 3 and 4 on the other hand. In the general case, the electric field components and the impedance tensor elements bear almost the same information, directly allowing a 'quick look' interpretation of the observations, even in presence of 3-D complex structures. In the case of a very conductive structure, however, the variations of the magnetic field over the structure may become significant (see e.g. Vasseur and Weidelt 1977, Ádám et al. 1986) depending on the frequency, and the electric field components may provide a better information than the tensor elements.

The tensor invariants and the skew

Tensor invariants have been introduced by many authors with the aim of helping in the electromagnetic data interpretation. These quantities are independent on the

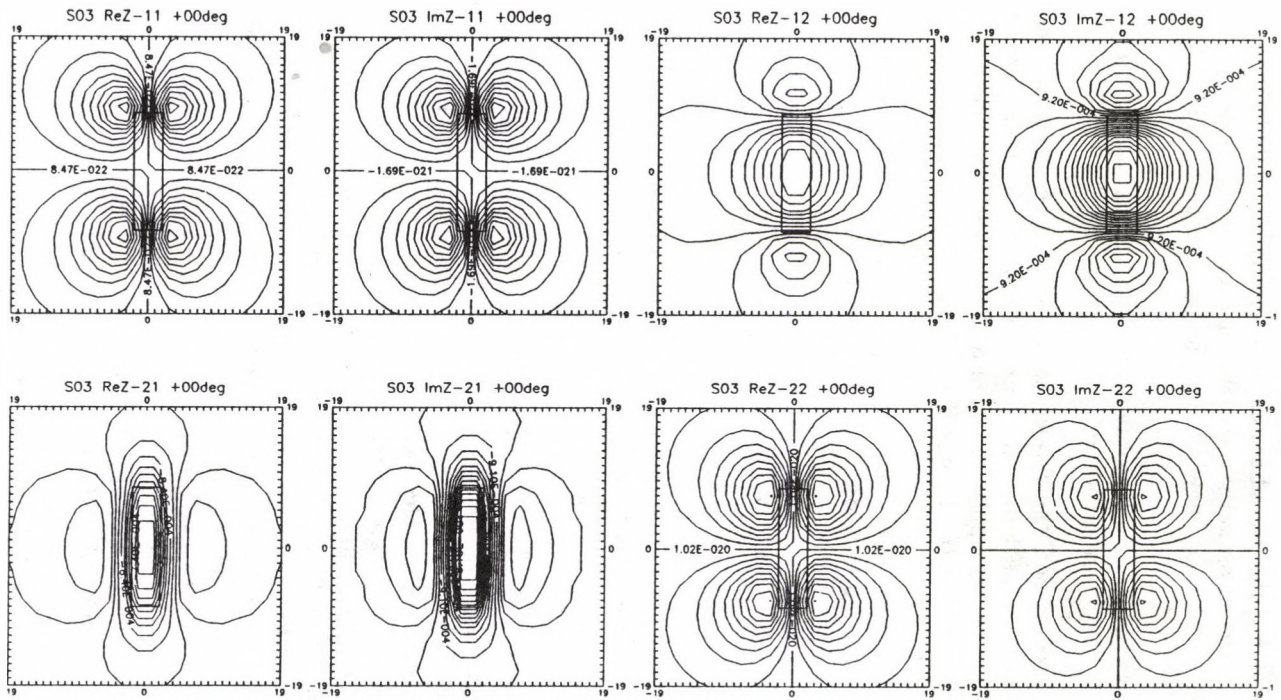

Fig. 5. Real and imaginary parts of the four (Z_{xx}, Z_{xy}, Z_{yx} and Z_{yy}) impedance elements over model S at a period of 225 s

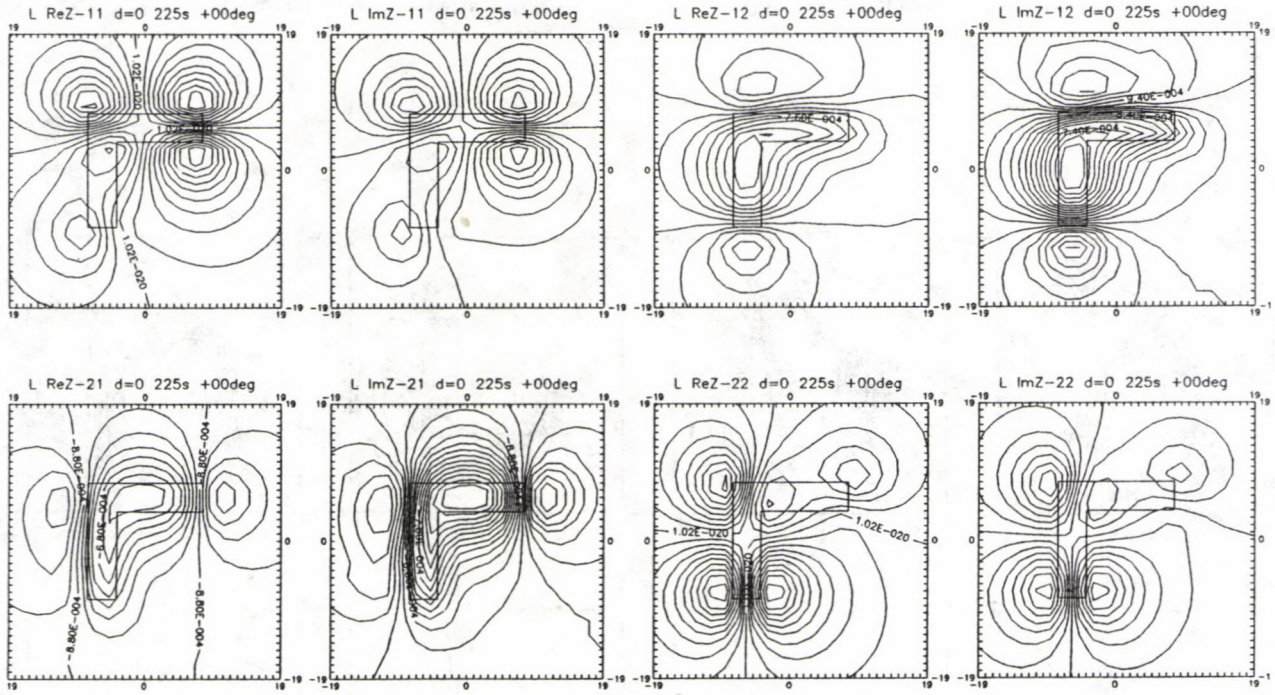

Fig. 6. Real and imaginary parts of the four (Z_{xx}, Z_{xy}, Z_{yx} and Z_{yy}) impedance elements over model L at a period of 225 s

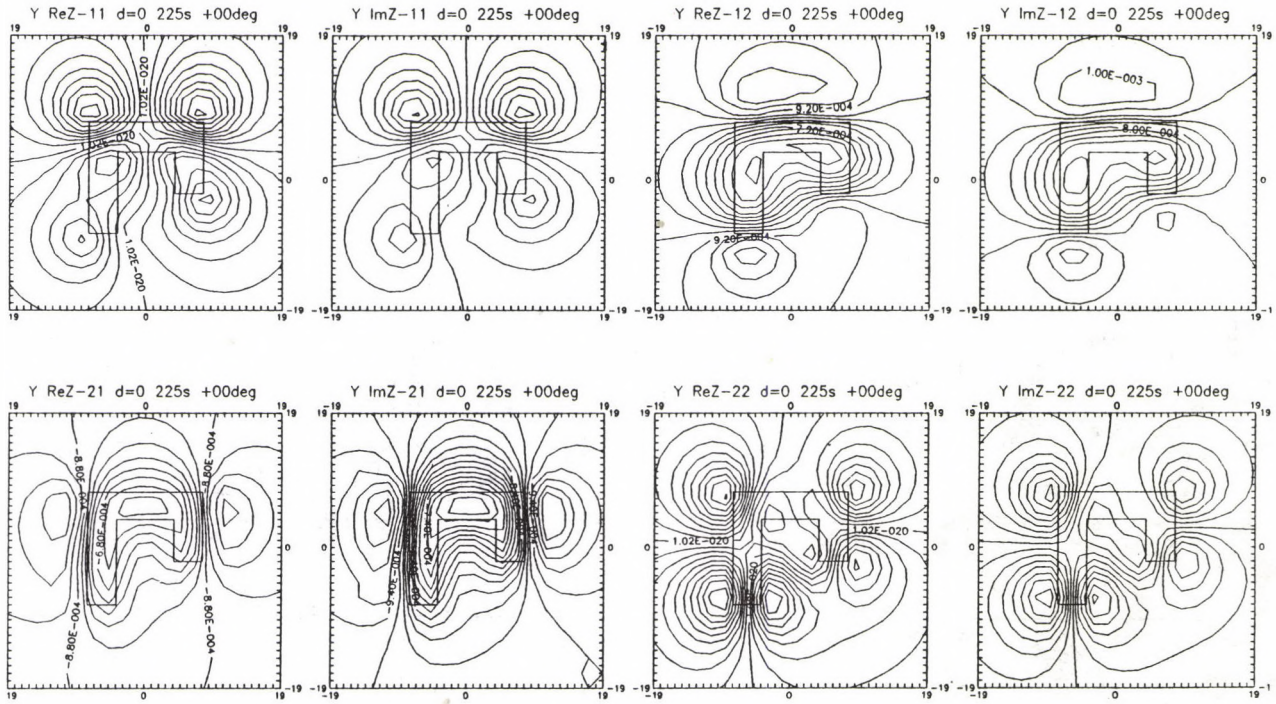

Fig. 7. Real and imaginary parts of the four (Z_{xx}, Z_{xy}, Z_{yx} and Z_{yy}) impedance elements over model Y at a period of 225 s

direction of the inducing field, and the geometry of the underlying structure is supposed to be more or less readible in the invariant maps. A review on the invariants can be found in Vozoff (1991). We will consider here the following quantities

$$c_1 = Z_{xx} + Z_{yy} \tag{13}$$

$$c_2 = Z_{xy} - Z_{yx} \tag{14}$$

$$c_3 = Z_{xy}Z_{yx} - Z_{xx}Z_{yy} \tag{15}$$

$$\alpha = \frac{c_1}{c_2} = \frac{Z_{xx} + Z_{yy}}{Z_{xy} - Z_{yx}} . \tag{16}$$

c_1 can be also called as spur; c_3 is the determinant of the four-element tensor $\mathbf{Z}$; hereafter called as Berdichevsky invariant following Hobbs (1993). α is the skew and it is considered by many authors as a 3-D indicator (e.g. Reddy et al. 1977, Jupp and Vozoff 1972, Ting and Hohmann 1981). Substituting the approximate expressions of the tensor elements (see relations (7) to (10)) in relation (16), it comes

$$\alpha \sim \frac{e_x H_y - e_y H_x}{E_x H_x - E_y H_y} \tag{17}$$

When values of H_x and H_y over all the area under study are in the same range, Eq. (17) can be reduced to:

$$\alpha \sim \frac{e_x + e_y}{E_x - E_y} \tag{18}$$

Over nearly-isometric (where x and y dimensions are nearly the same) 3-D models, Eq. (18) provides a good approximation of α. Since E_x and E_y have opposite signs and are generally the same order of magnitude, the skew is controlled first of all by the combined effect of the additional electric components (remind that these components are along the inducing magnetic field direction). Figures 8a, 9a and 10a show both real and imaginary parts of c_1, c_2, c_3, and skew over the three models we consider by using the exact formulas; Figs 8b, 9b and 10b show the corresponding electric field-based approximations. The comparison of these figures illustrates that the electric field-based approximations are relevant in the situations we consider.

The tensor invariant c_1 and the skew seem to be 'corner' indicators for the rectangular model. Since the exact and approximated maps have very similar shapes, their physical meaning can be inferred from the behaviour of the electric field components only. According to Eq. (13), c_1 is the sum of the two additional electric field components in the two polarizations. When E is parallel to the longer sides, the e-peaks are higher than in the other polarization. Since in the sum of the two additional components the effect of the higher peaks dominate, the c_1 parameter is principally sensitive to current channelling effects along the longer sides of the model (see Fig. 11). This is the reason for the slight shifts from the corners towards the longer sides. The number of e-peaks increases with increasing model complexity (see Figs 2-4), resulting in a complicated, quasi-random shape of c_1 and skew over geometrically complex models. The skew therefore may be useless, even misleading for qualitative interpretation over geometrically complex models.

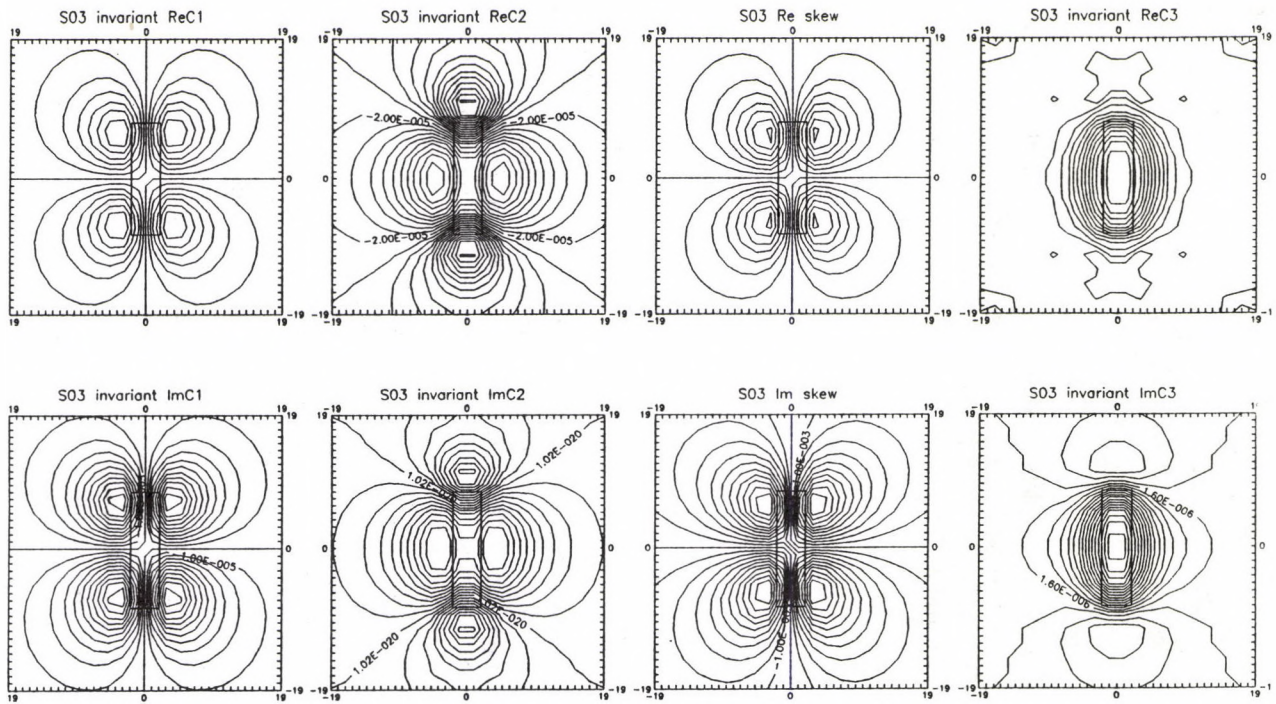

Fig. 8. a. Real and imaginary parts of c_1, c_2, c_3 invariants and the skew over model S at a period of 225 s by using exact formula

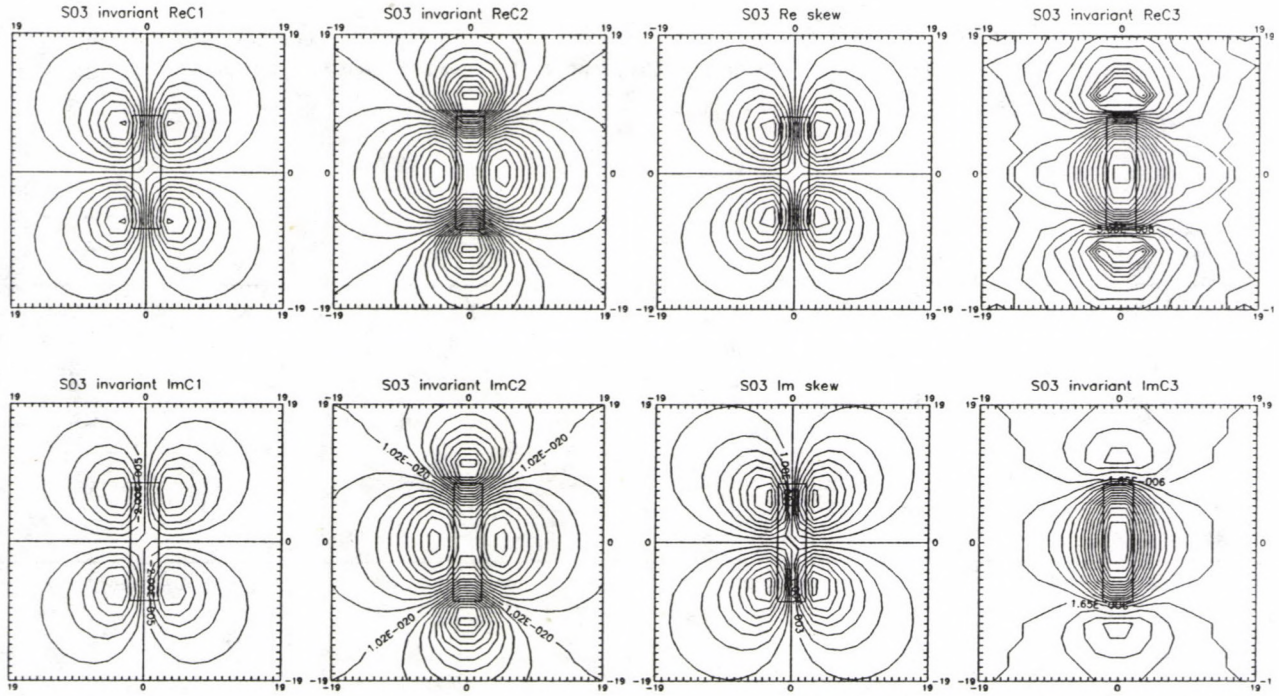

Fig. 8. b. Real and imaginary parts of c_1, c_2, c_3 invariants and the skew over model S at a period of 225 s by using approximating formula based on the electric field components only

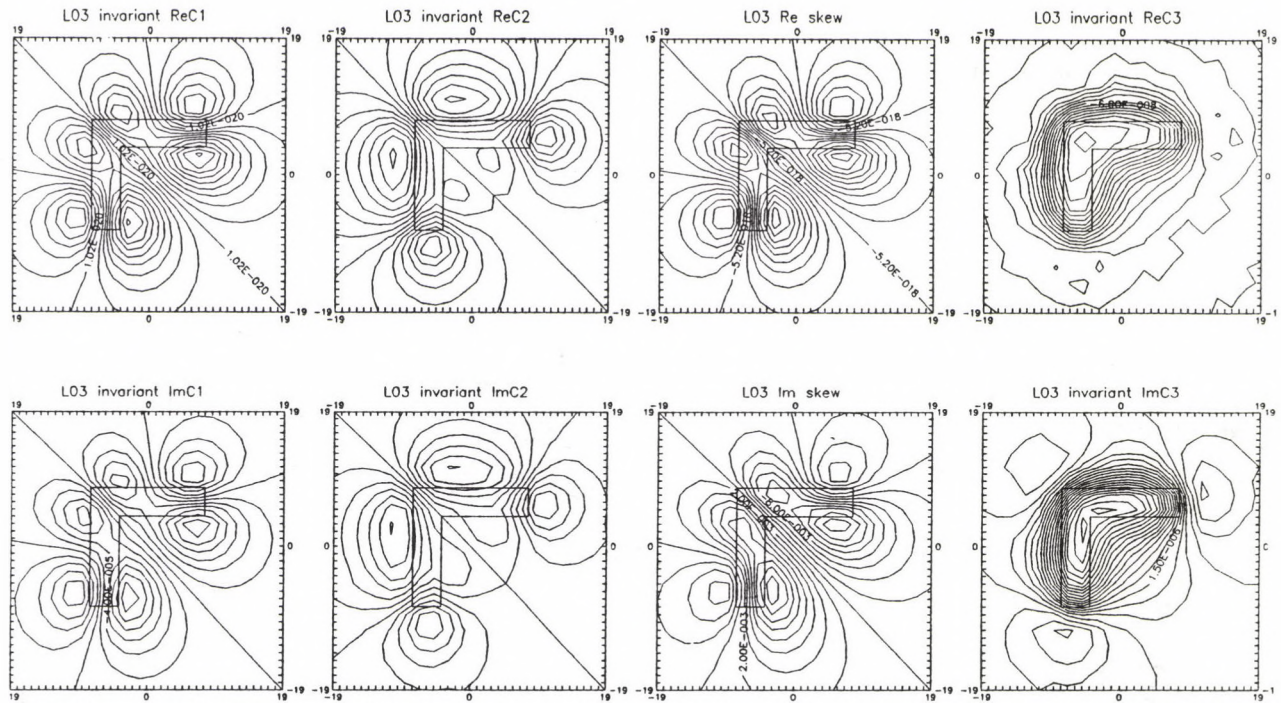

Fig. 9. a. Real and imaginary parts of c_1, c_2, c_3 invariants and the skew over model L at a period of 225 s by using exact formula

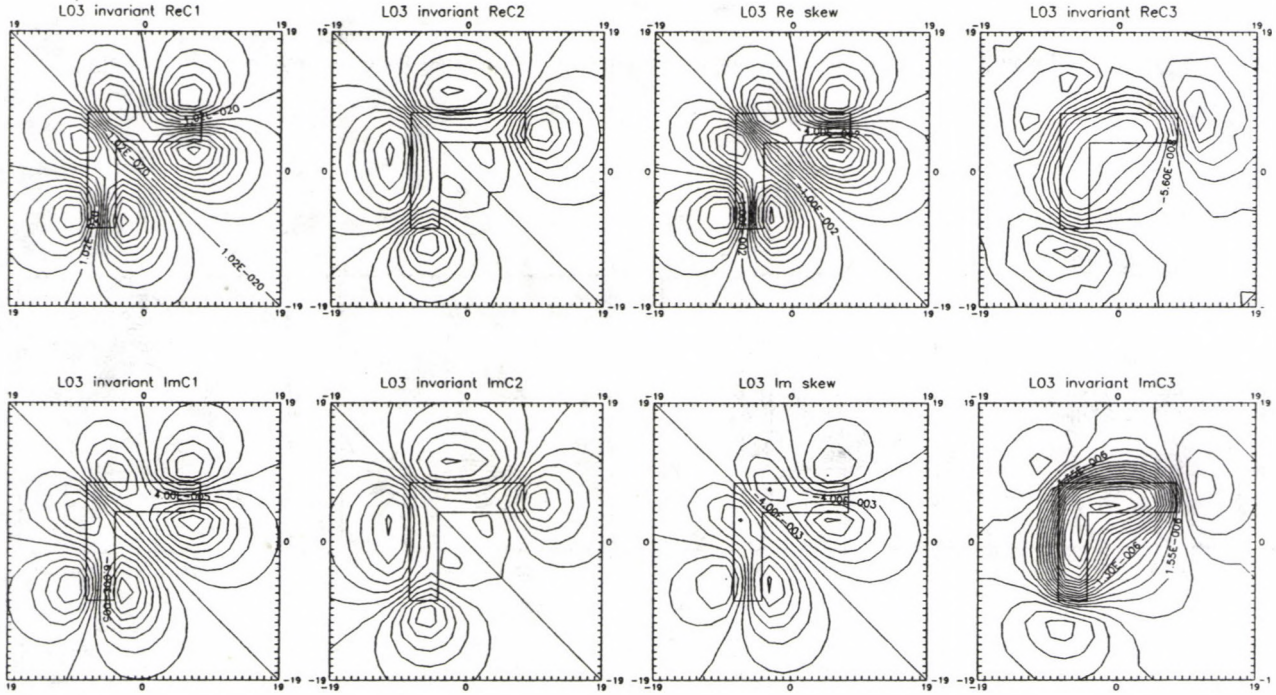

Fig. 9. b. Real and imaginary parts of c_1, c_2, c_3 invariants and the skew over model L at a period of 225 s by using approximating formula based on the electric field components only

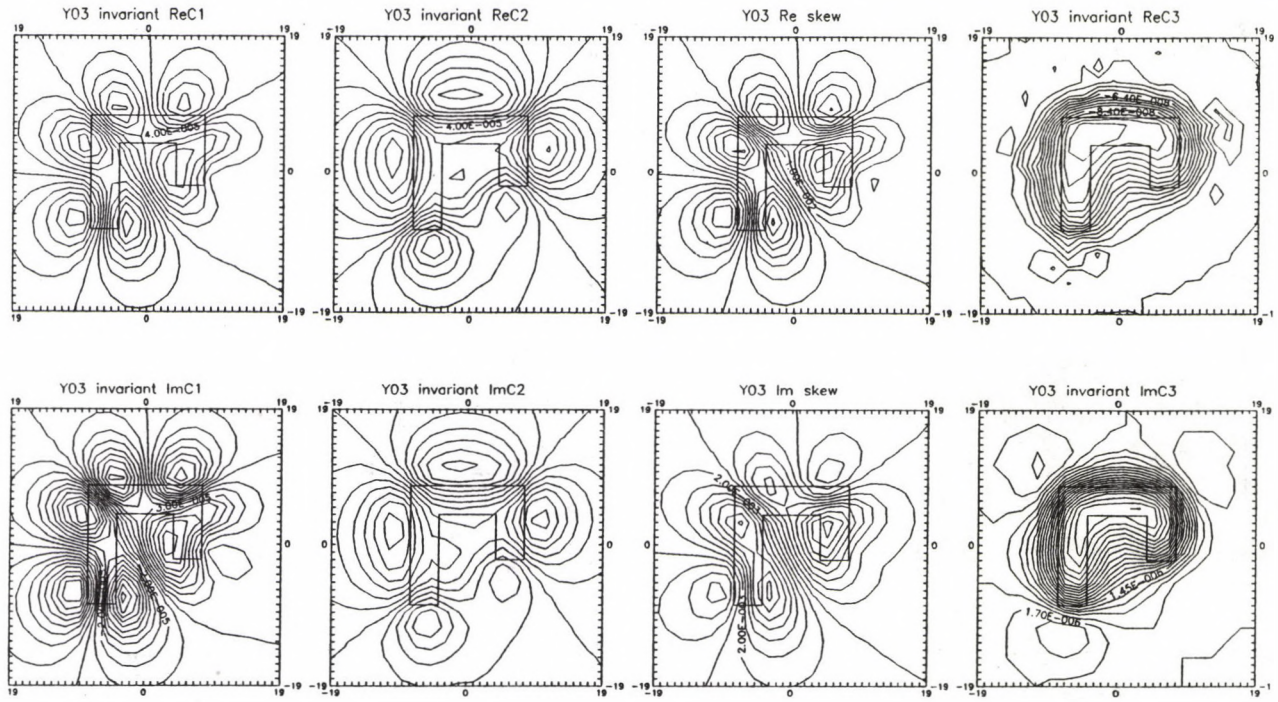

Fig. 10. a. Real and imaginary parts of c_1, c_2, c_3 invariants and the skew over model Y at a period of 225 s by using exact formula

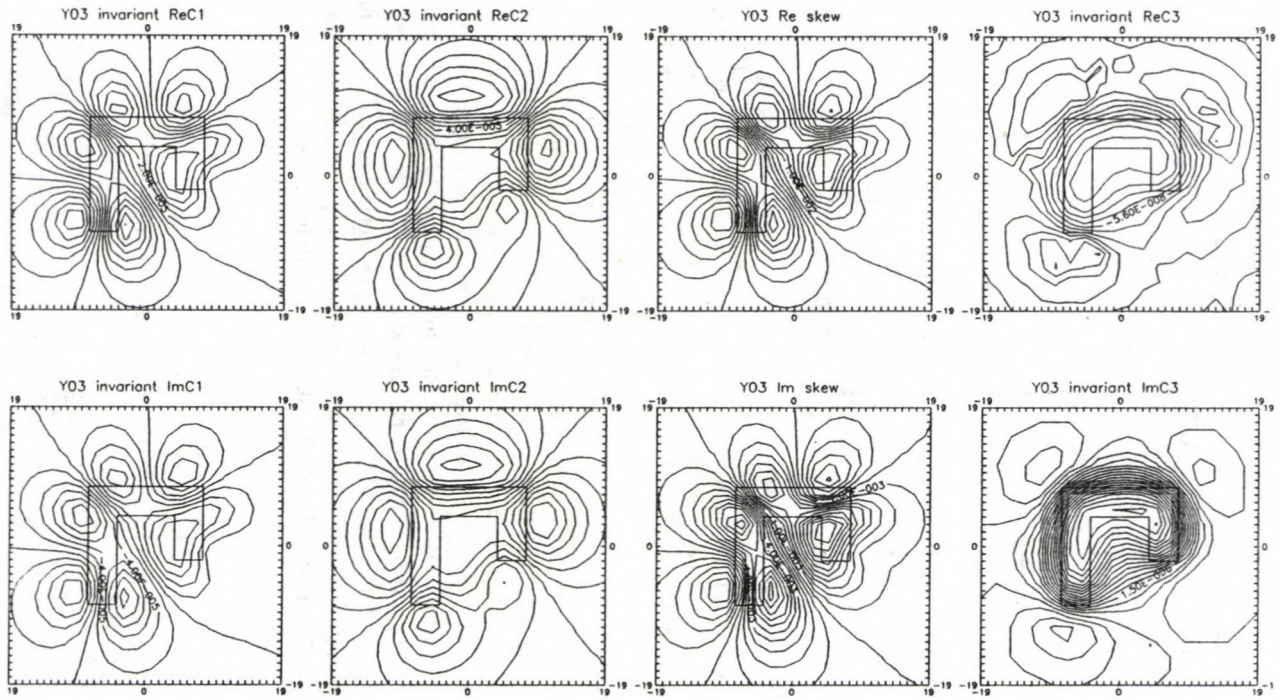

Fig. 10. b. Real and imaginary parts of c_1, c_2, c_3 invariants and the skew over model Y at a period of 225 s by using approximating formula based on the electric field components only

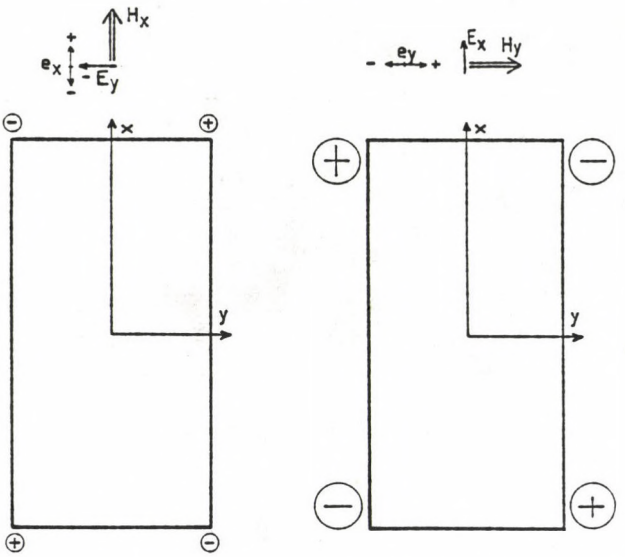

Fig. 11. The positive and negative peaks in the additional electric field components, representing secondary current channelling effects over a three-dimensional rectangular model at a period of 225 s. The current channelling inside the model is more important when the main electric field is parallel to its elongation as it demonstrated by larger circles. Therefore the extremum of the tensor invariant c_1 and in the skew c_1/c_2 the maximum will be shifted from the corners towards the longer sides

On the contrary, the c_2 and c_3 tensor invariants appear to be better 3-D indicators. For the three models we consider, c_2 seems to be a reliable 'side' indicator, while c_3 (the Berdichevsky invariant) fairly well depicts the geometry of the conductive structure. It is remarkable that the more complex the model geometry is, the better are the c_3 indications, since the 'side' maximum areas around the more complex models are more distributed. In addition, the electric field approximation of c_3 has significant 'side' maximum zones, calling attention to the not negligible smoothing effect in the impedance due to the horizontal magnetic components. The Berdichevsky invariant appears to be the best indicator for a qualitative interpretation over complex 3-D structures (Berdichevsky and Dmitriev 1976).

The ellipticity

The ellipticity (usually denoted by $\beta(\Theta)$) does depend on the field direction (Θ). Its definition is

$$\beta(\Theta) = \frac{Z_{11}(\Theta) - Z_{22}(\Theta)}{Z_{12}(\Theta) + Z_{21}(\Theta)} \tag{19}$$

where Θ is the (clockwise from North) direction of the inducing magnetic field; Z_{11}, Z_{12}, Z_{21} and Z_{22} are the rotated impedance elements. According to Eqs(7) to (10),

β can be estimated by:

$$\beta(\Theta) \sim \frac{e_x(\Theta) - e_y(\Theta)}{E_x(\Theta) + E_y(\Theta)}. \tag{20}$$

The rotated complex β parameter, for neither of the three models seems to be useful. These figures, which are not shown here clearly indicate that β is completely useless. The main problem with the β parameter is that it has curious maximum zones, corresponding to location where its denominator $(Z_{xy}+Z_{yx})$ becomes close to zero. For a rectangular model these β-maxima are along outward spreading curved lines originating from corners. In case of the L-shaped model the denominator is zero along the symmetry axis. Finally, it is worth noticing that for the U-shaped model (the most complex one, having no symmetry axis) β has no visible meaning for any rotation angle.

Discussion

The availability of a powerful 3-D thin sheet numerical algorithm made possible to analyse the behaviour of different EM parameters over 3-D structures for periods ranging from 14 to 3600 seconds. In the paper only some basic MT formulae are used, and the complicated derivations are intentially avoided. Instead of them the demonstration was given by means of maps and their brief descriptions.

We first consider the morphologic characteristics of the main and additional electromagnetic field components over geometrically complex high-conductivity 3-D models.

The geometric distribution of the impedance tensor elements, the tensor invariants, the skew and the ellipticity were also discussed. We showed that the morphological characteristics of most 3-D interpretation parameters are essentially the same as those of some combination of the electric components. The maps we present clearly show that the antidiagonal terms of the tensor are relevant side indicators while the diagonal terms bear almost no useful information. The tensor invariant c_1, the skew and the ellipticity are proved to be useless to depict the geometry of complex 3-D high conductivity structures. On the contrary, c_2 seems to be a reliable 'side' indicator, while the Berdichevsky invariant (the tensor invariant c_3) over such complex models becomes even better than over rectangular ones.

The results we present in this paper illustrate what can be the contribution of numerical studies to EM interpretation, provided efficient numerical tools are available.

In order to make a synthesis of our results, let us consider the case of a geophysical exploration over the U-shaped structure. Consider first the better situation, when an evenly spaced network (say one station at the center of each cell) is available. We therefore know the maps of the observed components (Fig. 4), the maps of the tensor impedance elements (Fig. 7), and those of the c_1, c_2, c_3 skew and ellipticity.

The ellipticity provides almost no information, and we will forget it immediately. In addition, Fig. 10 shows that neither the c_1 parameter, nor the skew provide any useful informations. On the contrary the Berdichevsky invariant depicts the

main features of the geometry of the structure. Some estimation about the outer limits of the conductive structure is also provided by the c_2 parameters. However the indications about the U shaped geometry of the structure are given by the antidiagonal elements of the tensor, and by the vertical component z (i.e. the tipper components): the portions of the zero lines in areas of important lateral gradient depict the outer edges of the structure in the case of the antidiagonal terms of the tensor, and the axes of the structure in the case of the tipper (see Figs 4 and 7).

The main electric and magnetic components may help in confirming the conclusions drawn from the antidiagonal terms of the tensor and the tipper components. The additional electric and magnetic fields do not provide further information.

Consider now the situation which prevails in practice, namely an ill conditioned network. In this case, the smoothness of the variations of the indicator is of basic importance, because it governs the relevancy of the interpolation which will be necessarily made. Short spatial wavelengths will induce aliasing, and in practice lead to interpolated maps without any practical meaning. The diagonal term of the tensor, the skew and the c_1 parameter are therefore to be rejected. On the contrary, the tipper components, the Berdichevsky invariant and the antidiagonal terms of the tensor have fairly smooth variations, and they will still provide an accurate description of the structure, although less precise because of the larger spacing between the stations.

Therefore it clearly appears that, in any situation, the more reliable description of a conductive structure is obtained by using:

- the Berdichevsky invariant which depicts the main lateral geometrical trends;
- the antidiagonal terms of the tensor which indicate the outer limits of the structure. Note that these terms may be favourably replaced by the corresponding electric components also in case of strong magnetic effect of the currents flowing in the structure;
- the tipper components which delineate the symmetry axes of the structure.

In opposition, the ellipticity, as well as the diagonal term of the tensor and the related indicators, the skew and the spur, should be avoided. (The Berdichevsky-invariant is not only able to separate shallow and deep anomalies (as it was shown in Ranganayaki 1984), but it is the best in lateral terms as well.)

Acknowledgement

This work was partly supported by a French-Hungarian (CNRS–Hungarian Academy of Sciences) program and partly by the Hungarian Research Found (No 1173).

References

Ádám A 1992: *J. of China Univ. of Geosciences*, 3, 92–104.

Ádám A, Szarka L, Verő J, Wallner Á, Gutdeutsch R 1986: *Physics of the Earth and Planetary Interiors*, 42, 165–177.

Berdichevsky M N, Dmitriev V I 1976: In: Ádám A (Ed.) Geoelectric and Geothermal Studies, Akadémiai Kiadó, Budapest, 165–221.
Cagniard L 1953: *Geophysics*, 18, 605–635.
Counil J L, Le Mouel J L, Menvielle M 1986: *Annales Geophysicae*, 4, 115–125.
Dawson T W, Weaver J T, Raval L 1982: *Geophys. J.R. Astr.Soc.*, 69, 209–234.
Edwards R N, Eggers D E 1982: *Geophysics*, 47, 1204–1214.
Hobbs B 1993: *Surveys in Geophysics*, 13, 489–515.
Jupp L B, Vozoff K 1972: *Geophysics*, 41, 325–328.
Larsen J C 1975: *Geophys. J. R. Astr. Soc.*, 68, 575–587.
Le Mouel J L, Menvielle M 1982: *Geophys. J. R. Astr. Soc.*,
Mareschal M, Jouanne V, Menvielle M, Chouteau M, Grandis H, Tarits P 1992: *Phys. Earth Planet. Interiors*, 74, 191–197.
Menvielle M 1987: *Survey in Geophys.*, 9, 319–348.
Park S K 1985: *Geophysics*, 50, 785–797.
Ranganayaki R P 1984: *Geophysics*, 49, 1730–1748.
Reddy I K, Rankin D, Phillips R J 1977: *Geophys. J. R. Astr. Soc.*, 51, 313–325.
Tarits P 1989: Contribution des sondages électromagnétiques profonds dans l'étude du Manteau Supérieur terrestre: These d'Etat, Paris VII, 333 pp.
Ting S C, Hohmann G W 1981: *Geophysics*, 46, 182–197.
Vasseur G, Weidelt P 1977: *Geophys. J. R. Astr. Soc.*, 51, 669–690.
Vozoff K ed. 1986: Magnetotelluric Methods: Geophys. Reprint, Ser. No. 5, Soc. Expl. Geophys.
Vozoff K 1991: In: Electromagnetic methods in applied geophysics (ed. M Nabighian) Vol. 2, Application, Part B, 641–711.
Wannamaker P E 1990: *Geophysics*, 55, 372–375.
Wannamaker P E, Hohmann G W, Ward S H 1984: *Geophysics*, 49, 1517–1533.
Weidelt P 1975: *Journal of Geophysics*, 41, 85–109.
Zhang P, Pedersen L B, Mareschal M, Chouteau M 1993: *Geophys. J. Int.*, 113, 693–700.

Acta Geod. Geoph. Hung., Vol. 29(1–2), pp. 107–124 (1994)

A THIN SHEET NUMERICAL STUDY OF THE ELECTROMAGNETIC FIELD OVER GEOMETRICALLY COMPLEX HIGH CONDUCTIVITY STRUCTURES: DEPTH AND LATERAL CHARACTERISTICS OF DIFFERENT APPARENT RESISTIVITY DEFINITIONS

L SZARKA[1], A ÁDÁM[1], M MENVIELLE[2], P TARITS[3]

[Manuscript received February 15, 1994]

In this paper, belonging to a series of papers dealing with thin sheet studies of the electromagnetic field over geometrically complex models, a systematic study about different apparent resistivity definitions (summarized by Spies and Eggers 1986) in 1-D and 3-D environments at the surface and in the depth is given. The 3-D heterogeneities are high-conductivity crustal thin sheet structures having different geometrical complexity in a homogeneous lithosphere, underlain by a deep high-conductivity asthenosphere. We have found a big and meaningful difference between the apparent resistivity definitions determined at different depth levels, while the lateral characteristics of these resistivity definitions were found to be more or less the same. The conventional Cagniard apparent resistivities seem to be in general acceptable, but in some special studies the use of ϱ_{ReZ} (apparent resistivity computed from the real part of the complex impedance) might give better results.

Keywords: apparent resistivity; magnetotellurics; numerical modelling; thin sheet; three-dimensional models

1. Introduction

Several years ago Spies and Eggers (1986) compared different definitions of magnetotelluric apparent resistivities. They expected to find an ideal apparent resistivity curve, having perfect asymptotes, and quick and oscillationfree transitions at the shortest periods. In this sense the best apparent resistivity was found to be that which is based on the real part of the impedance, and one of the worst was that which is determined from the imaginary part of the impedance. Cagniard's classical apparent resistivity we generally use is the arithmetic mean of these two apparent resistivities.

We investigate in this paper how these different apparent resistivities behave in presence of 3-D inhomogeneities. The aim of this study is to help in deciding which

[1]Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences, H–9401 Sopron, POB 5, Hungary

[2]Laboratoire de Physique de la Terre et des Planetes (URA 1369), Bat. 504 – Université Paris Sud, F–91405 Orsay Cedex, France

[3]Université de Bretagne Occidentale Laboratoire de Geosciences Marines (URA 1278), 6 Avenue Le Gorgeu, F–29200 Brest, France

parameter is to be used for data interpretation, given the existing data (number of stations, frequency range, data quality, *etc.*) and the *a priori* knowledge about the structure under study. In many situations, the different parameters are not in practice equivalent, because of their different response to conductivity heterogeneities. It is in particular the case when making preliminary data interpretation with non-sophisticated imaging tools.

The models considered are described in Section 2 and the algorithms used for computation are presented in Section 3. The results are discussed in Section 4. Finally illustrations of the possible contribution of our results to field measurements are given in Section 5.

2. The numerical modelling

We considered four different models corresponding to typical geophysical situations:

Model 1. A two layer model with a resistive upper layer standing for the crust over a conductive half space simulating a conductive crustal basement ($\varrho_1 = 40\ \Omega$m, $h_1 = 4$ km, $\varrho_2 = 1\ \Omega$m);

Model 2. A two layer model with a resistive upper layer standing for the lithosphere ($\varrho_1 = 40\ \Omega$m, $h_1 = 65$ km) over a conductive half space simulating a conductive asthenosphere ($\varrho_2 = 1\ \Omega$m);

Model 3. A four-layer half space, based on the two previous models ($\varrho_1 = 40\ \Omega$m, $h_1 = 4$ km, $\varrho_2 = 0.4\ \Omega$m, $h_2 = 0.4$ km, $\varrho_3 = 40\ \Omega$m, $h_3 = 65$ km, $\varrho_4 = 1\ \Omega$m);

Model 4. Model No. 4 was a 3-D one, based on 1-D model No. 3, but instead of the third layer one of the 3-D thin sheet heterogeneities was placed. (The set of 3-D thin sheet heterogeneities is shown in Fig. 1.)

We studied the behaviour of different resistivities at different apparent depths for the three 1-D models we consider (models 1, 2 and 3), and in the vicinity of 3-D thin sheet structures (model 4).

We used the Vasseur and Weidelt (1977) and Tarits (1989) numerical thin sheet algorithms. These programs allow to determine the electromagnetic field not only at the surface (Vasseur and Weidelt algorithm), but at arbitrary depths, too.

The electromagnetic field was modelled over a 40 km $\times$ 40 km wide area using a 20 $\times$ 20 mesh of 2 km $\times$ 2 km cells (see Fig. 1), and computed at four different depths: at the surface (level 0), at the depth of 2 km (level 1), at the depth of 3 km (level 3), and at the depth of 3.5 km (level 4). Five different periods, covering the range from the short period limitation of the thin sheet approximation to the dominant appearance of the asthenosphere were used: 14 s, 56.25 s, 225 s, 900 s and 3600 s. The impedance of the corresponding 1-D models was also computed. More details about the numerical modelling problems is discussed by Szarka et al. (1994).

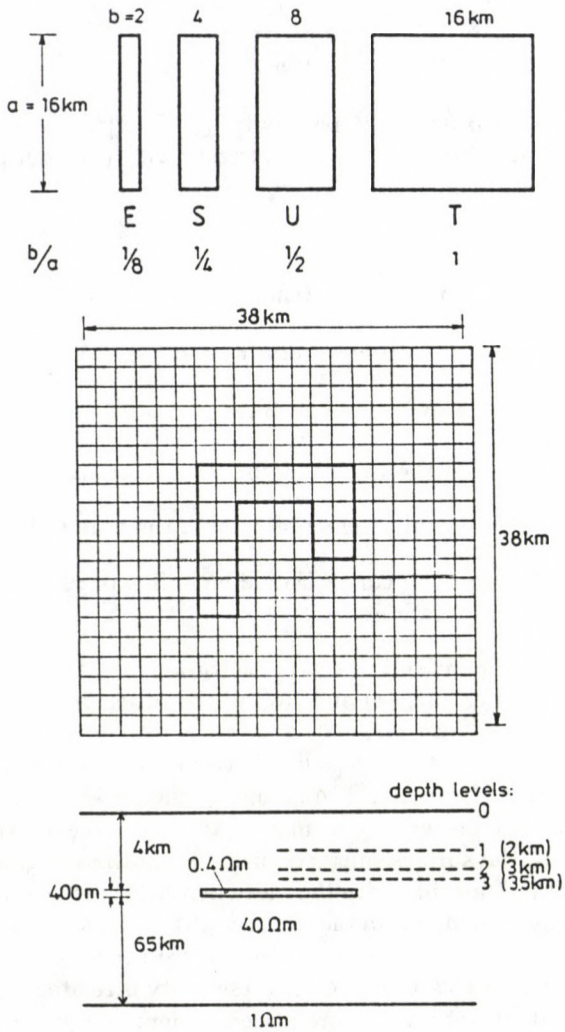

Fig. 1. a) Plan views of 3-D thin sheet models used in these studies. E, S, U and T denote different rectangular models. b) Plan view of the thin sheet computation mesh and a characteristic model cross section showing all resistivity and geometric parameters and also the investigated depths

3. The apparent resistivity definitions

We consider hereafter the five different MT apparent resistivity definitions studied by Spies and Eggers (1986):

$$\varrho_1 = \varrho_{ReZ} = 2|ReZ|^2/\omega\mu \tag{1}$$

$$\varrho_2 = \varrho_{ImZ} = 2|ImZ|^2/\omega\mu \tag{2}$$

$$\varrho_3 = \varrho_{|Z|} = (\varrho_{ReZ} + \varrho_{ImZ})/2 \tag{3}$$

$$\varrho_4 = \sqrt{(\varrho_{ReZ} \cdot \varrho_{ImZ})} \tag{4}$$

$$\varrho_5 = 1/2 \cdot \sqrt{(\varrho_{ReZ}^2 + \varrho_{ImZ}^2)}\,. \tag{5}$$

ϱ_3 is the classical Cagniard apparent resistivity, $\varrho_3 = |Z|^2/\omega\mu$. It is known that ϱ_{ReZ} is nothing else but Schmucker's apparent resistivity (Schmucker 1970). ϱ_{ReZ} and ϱ_{ImZ} can be derived from $\varrho_{|Z|}$ as follows:

$$\varrho_{ReZ} = 2\cos^2\varphi \cdot \varrho_{|Z|} \tag{6}$$

$$\varrho_{ImZ} = 2\sin^2\varphi \cdot \varrho_{|Z|} \tag{7}$$

where φ is the argument of Z. It comes from (6) and (7) that

$$\varrho_{ImZ}/\varrho_{ReZ} = \tan^2\varphi\,. \tag{8}$$

Spies and Eggers(1986) tested the MT sounding curves for the following criteria:

— the asymptotic values should well approximate the true resistivities;

— the changes on the sounding curves should be well expressed, should be quick, and possibly free of oscillations.

ϱ_{ReZ} was found to be relatively the best and ϱ_{ImZ} was found to be definitely the worst. According to Eqs (3), (4) and (5), all the other definitions (ϱ_3, ϱ_4 and ϱ_5) are some means of ϱ_{ReZ} and ϱ_{ImZ}.

Schmucker (1970) evidenced the close link between ImZ and as he called it 'the center of currents depth'. Later on, Szarka and Fischer (1989, 1991) showed that the real part of the impedance gives information about the mean depth of currents flowing out-of-phase to the surface magnetic field, while the imaginary part of the impedance is in linear relationship with the mean depth of currents flowing in-phase with the surface magnetic field. From the formulas (3)–(5) it follows that ϱ_3, ϱ_4 and ϱ_5 describe some alternatively defined mean depths of subsurface currents flowing in the earth. For example, the classical Cagniard resistivity is related to the arithmetic mean of 'in-phase' and 'out-of-phase' current-centre depths.

In the following, we will consider the two basic apparent resistivity definitions ϱ_{ReZ}, ϱ_{ImZ}, and for comparison the Cagniard apparent resistivity $\varrho_{|Z|}$. In a multi-dimensional situation ϱ_{xy}, ϱ_{yx} and even $\varrho_{\text{inv}} = \sqrt{\varrho_{xy} \cdot \varrho_{yx}}$ (the invariant apparent resistivity, see Berdichevsky (1968) for details) are also introduced. They can be computed from either the real part, or the imaginary part, or the modulus of the corresponding complex impedance element.

4. The results

Before looking at the behaviour of the apparent resistivities close to 3-D structure, it is worth recalling some properties of the evolution with depth of the apparent resistivities in 1-D situation.

1-D situations

Let us consider the three different cases mentioned in Section 3 to illustrate what happens in 1-D situations. The computation was made for the same period range (14–3600 s) as in Paper 1, Szarka et al. (1994). The observation depth ranges from 0 m (surface) to 3500 m.

Two-layer models

In case of the first two-layer crustal model (Model 1), the difference between ϱ_{ReZ} and ϱ_{ImZ} is very big in the whole period range, because ϱ_{ReZ} reaches much earlier (that is at much shorter periods) the high-conductivity basement (see Fig. 2).

For the asthenosphere model (Model 2) at all investigation depths the same results were obtained, which are very close to the surface values. This is due to the relatively small subsurface observation depths compared to the asthenosphere depth (see Fig. 3).

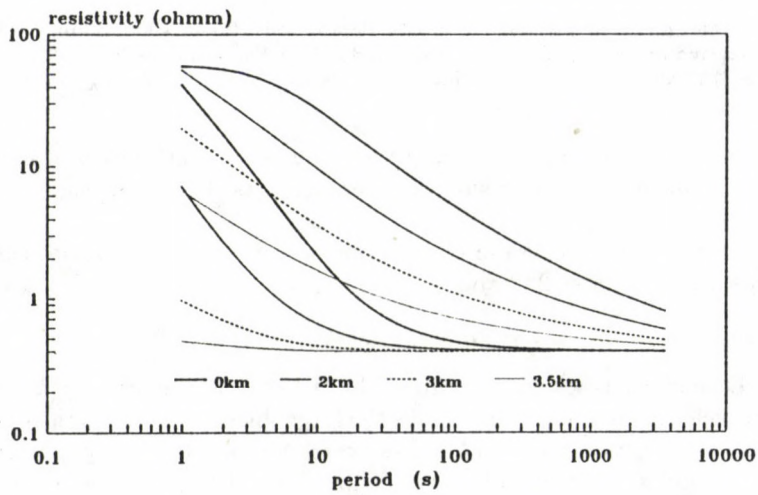

Fig. 2. ϱ_{ReZ} and ϱ_{ImZ} at four different crustal depth levels (0 km, 2 km, 3 km, 3.5 km) over a simple two-layered model, representing a high-conductivity basement at a depth of 4 km. The ϱ_{ReZ} curves are those which decrease more steeply and have lower resistivity values than the ϱ_{ImZ} ones

Four-layer model

In case of the four-layer Model 3 the crustal effect disappears at long periods both in ϱ_{ReZ} and ϱ_{ImZ}, but it takes place at much shorter periods in case of ϱ_{ReZ} (Fig. 4a) than in case of ϱ_{ImZ} (Fig. 4b). (The ϱ_{ReZ} apparent resistivity curves at different depths are nearly the same at longer than T $\approx$ 10 s periods. Such a period limit exists for ϱ_{ImZ} only at periods T $\approx$ 100 s.) In the long period--range, where the apparent resistivity values are the same at all investigated depths, exclusively the asthenospheric effects dominate in the resistivity values. In the

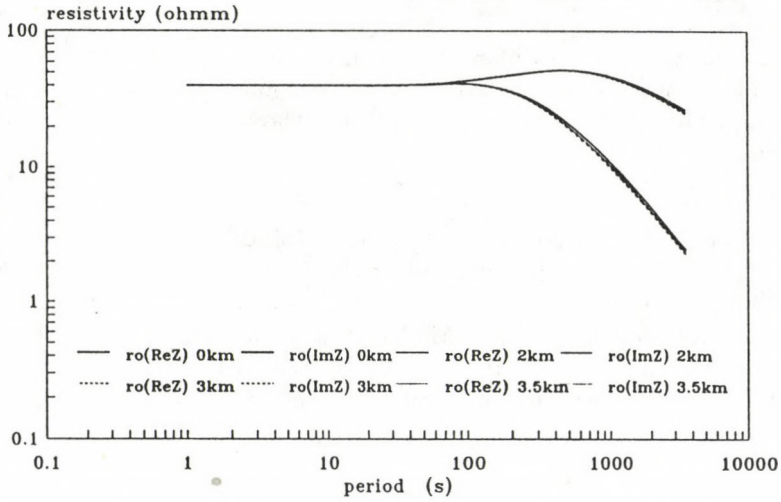

Fig. 3. ϱ_{ReZ} and ϱ_{ImZ} at four different crustal depth levels (0 km, 2 km, 3 km, 3.5 km) over a simple two-layered model, representing a high-conductivity basement (asthenosphere) at a depth of about (69.4) 70 km. The ϱ_{ReZ} curves have lower resistivity values than ϱ_{ImZ} ones only at long periods

10 s–100 s period range ϱ_{ReZ} is already the same at all depth levels, whilst ϱ_{ImZ} is still strongly influenced by near-surface layers and has therefore a significant depth dependence.

It must be remarked that the conclusions are drawn from a visual comparison of sounding curves in Figs 2, 3 and 4.

Relation between 1-D curves at different depths

The 1-D characteristics are determined by the key-parameters $\lambda_1/(h-z)$ (where λ_1 is the wavelength in the first layer, h is the layer thickness and z is the observation depth), and s (reflection coefficient). The expression of ϱ_{ReZ} and ϱ_{ImZ} as function of $\lambda_1/(h-z)$ and s are given in the Appendix. The interrelation between ReZ and ImZ (which is very similar to that between ϱ_{ReZ} and ϱ_{ImZ}) is shown in Fig. 5 for a $\sigma_2 = 100\sigma_1$ half-space. According to the Appendix for a certain conductivity contrast S the resistivity and phase values are determined exclusively by $\lambda_1/(h-z)$. It means that completely the same ϱ and ϕ conditions prevail for all T and z values (period and depth) characterized by the same $\lambda_1/(h-z)$. The period shift with depth is determined by the constancy of $\lambda_1/(h-z)$ (Szarka and Fischer 1991).

From Fig. 5 it can be seen that the change in ReZ is much less than in ImZ when the $\lambda_1/(h-z)$ parameter is large enough (in Fig. 5 when it is greater than about 10). This is in perfect accordance with the asymptotes in ϱ_{ReZ} appearing earlier at long periods.

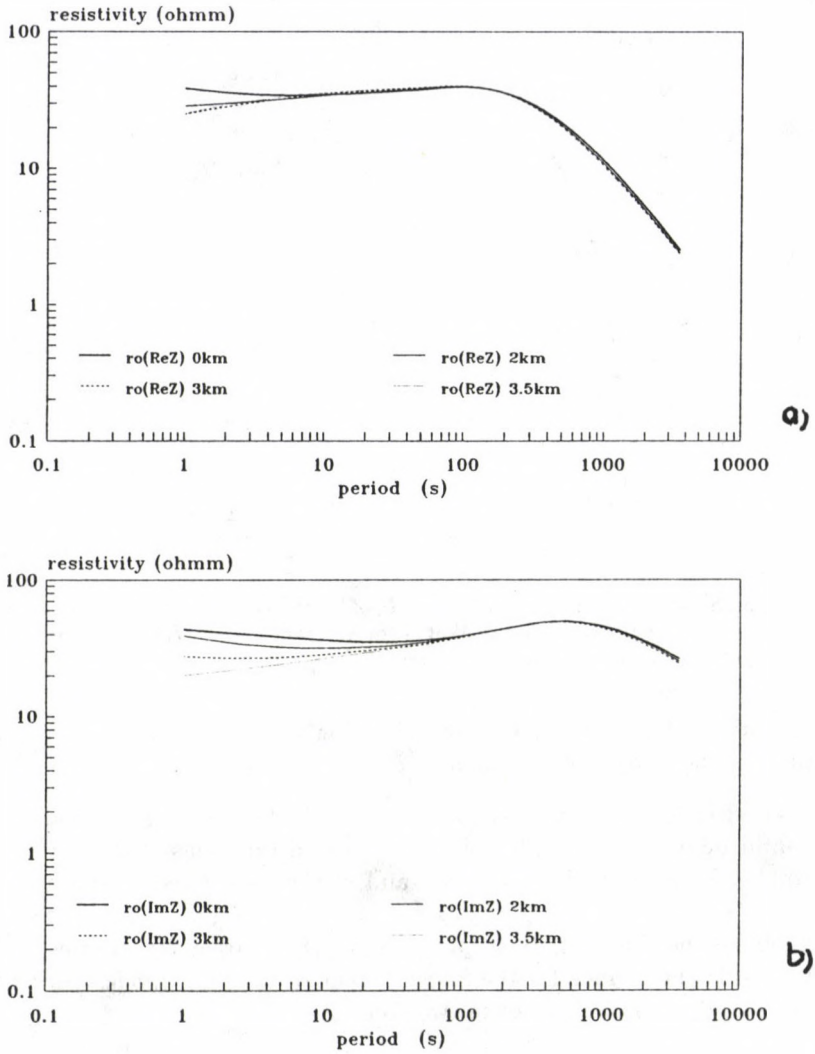

Fig. 4. ϱ_{ReZ} and ϱ_{ImZ} at four different crustal depth levels (0 km, 2 km, 3 km, 3.5 km) over a four-layered model, representing a high-conductivity crustal layer at 4 km and a high-conductivity basement (asthenosphere) at a depth of about 70 km. a) ϱ_{ReZ} curves, b) ϱ_{ImZ} curves

3-D SITUATIONS

3-D characteristics of ϱ_{ReZ} and ϱ_{ImZ} were studied for models described in Fig. 1.

Far from the three-dimensional structures, 1-D relations dominate, as expected. Coming closer to the heterogeneities, the resulting anomaly depends on the direction of polarization. At any depth, two different, well-known effects appear:

1. an apparent resistivity decrease roughly over the central part of the models;

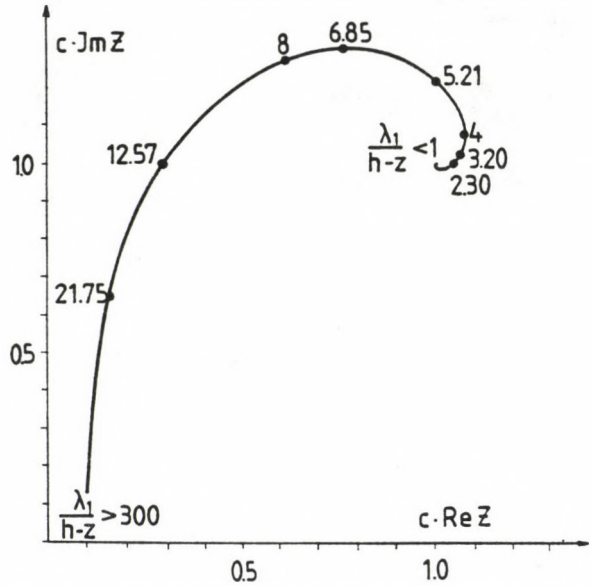

Fig. 5. Relation between dimensionless ReZ and ImZ on the surface and inside the first layer of a $\sigma_2 = \sigma_1$ two-layered half-space, as a function of the key parameter $\lambda_1/(h-z)$. The numbers — in terms of $\lambda_1/(h-z)$ — denote different features on the apparent resistivity and phase sounding curves

2. a resistivity increase ('spatial overshooting') just outside the main current inflow-outflow edges of the model.

The former effect is always more significant than the latter (both in areal extension and in amplitude) but both effects are complicated functions of the direction and the period of the field on the one hand, and of the model geometry on the other hand.

We mapped the three different (ϱ_1, ϱ_2 and ϱ_3) apparent resistivities which we consider at different periods for the four different rectangular models, described in Fig. 1. The study was carried out in two steps:

1. A morphological comparison of ϱ_{xy}, ϱ_{yx} and of the invariant apparent resistivity ϱ_{inv} maps
2. Study of 3-D effects in selected sites. Regarding the apparent resistivity maps over the models, both the minimum-rho (ϱ_{min}) and maximum-rho sites (ϱ_{max}) are discussed.

Comparison of ϱ_{xy} ϱ_{yx} and ϱ_{inv} apparent resistivity maps by using different apparent resistivity definitions

Figures 6, 7, 8 and 9 show ϱ_{xy}, ϱ_{yx}, and ϱ_{inv} apparent resistivity maps corresponding to the three different definitions of the apparent resistivity (ϱ_1, ϱ_2 and ϱ_3) we consider. The apparent resistivity values are normalized to the corresponding

1-D values, far from the inhomogeneities. Figures 6, 7, 8 and 9 correspond to 0 (surface), 2, 3, and 3.5 km observation depth, respectively.

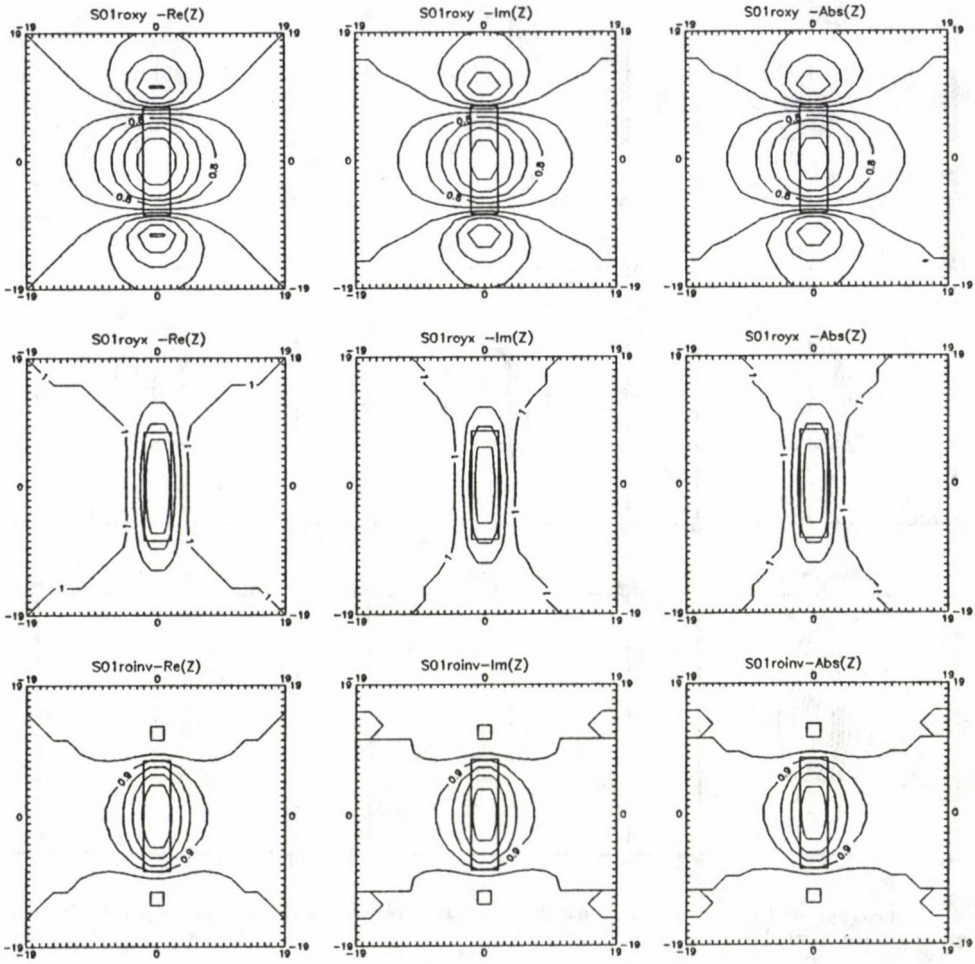

Fig. 6. $\varrho_{xy}, \varrho_{yx}$ and ϱ_{inv} maps by using three different resistivity definitions for model S on the surface at a period of 3600 s

The periods presented here were selected in such a way that the 1-D key parameter, $\lambda_1/(h-z)$ is the same for each map. The periods, depths, and the corresponding $\lambda_1/(h-z)$ values are summarized in Table I.

These figures show that at the surface the areal distribution of ϱ_{ReZ} and ϱ_{ImZ} are practically the same. This is confirmed by other maps we have drawn (not published in this paper). This is due to the large value of $\lambda_1/(h-z)$, and to the corresponding small phase anomaly.

When going deeper and deeper the difference between ϱ_{ReZ} and ϱ_{ImZ} maps —

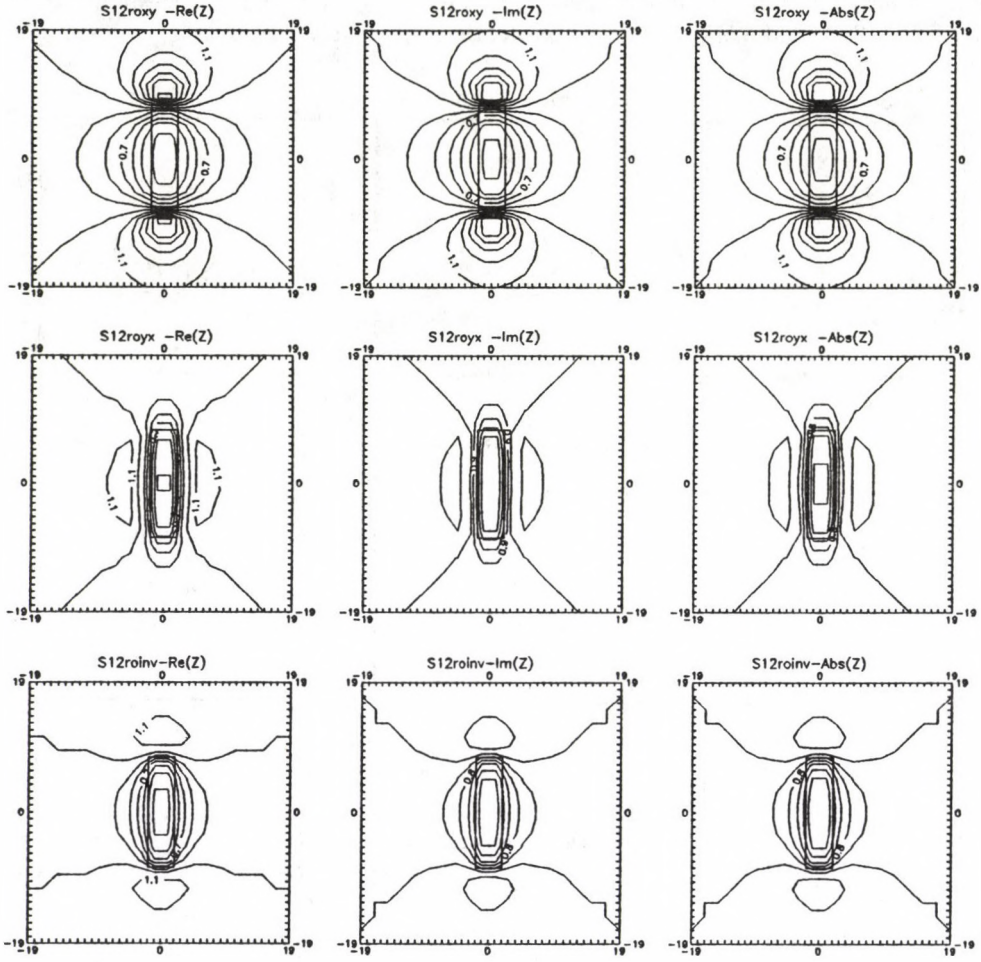

Fig. 7. $\varrho_{xy}, \varrho_{yx}$ and ϱ_{inv} maps by using three different resistivity definitions for model S at a depth of 2 km, period 900 s

first of all in the center parts and in the slight $\varrho_{\max}$ sites — becomes more and more significant. The central anomaly and the side effects are more expressed in both field directions in the ϱ_{ReZ} maps than in the ϱ_{ImZ} ones. It is interesting to notice that at the same time the difference in terms of 'spatial overshootings' between the different apparent resistivity definitions completely disappears in the ϱ_{inv} maps.

Invariant apparent resistivity maps at selected ($\varrho_{\max}$ and $\varrho_{\min}$) sites

In this section we only consider the invariant apparent resistivities. They are computed from orientation-free invariant tensor elements, and normalized by the 1-D values far from the 3-D model. From Figs (6–9) it can be seen that the relative apparent resistivity decrease observed above 3-D crustal inhomogeneities is quite

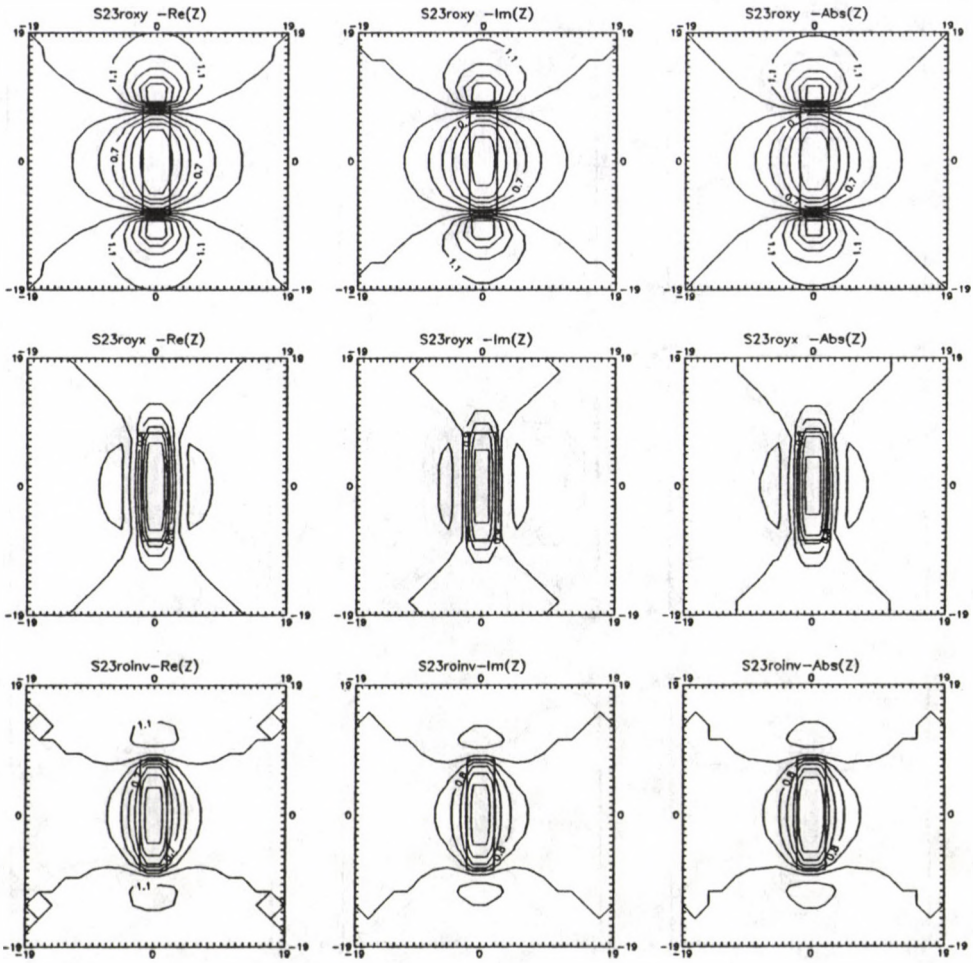

Fig. 8. $\varrho_{xy}, \varrho_{yx}$ and ϱ_{inv} maps by using three different resistivity definitions for model S at a depth of 3 km, period 225 s

considerable for all apparent resistivity definitions and the highest effect (the most isolines) can be seen in ϱ_{ReZ}. At the same time, all invariant apparent resistivity maps give nearly the same 3-D effect outside the 3-D heterogeneities.

The overall 3-D characteristics is shown in Fig. 10 which was compiled from invariant $\varrho_{\min}$ and invariant $\varrho_{\max}$ values (always normalized by the corresponding 1-D values) for the four different models, at a period of 225 s. In this figure both surface (continuous line) and subsurface relative resistivities (dotted lines) are shown. This figure shows that:

1. at any depth, it is always ϱ_{ReZ} which has the largest decrease, and ϱ_{ImZ} which has the smallest decrease over high conductivity models;

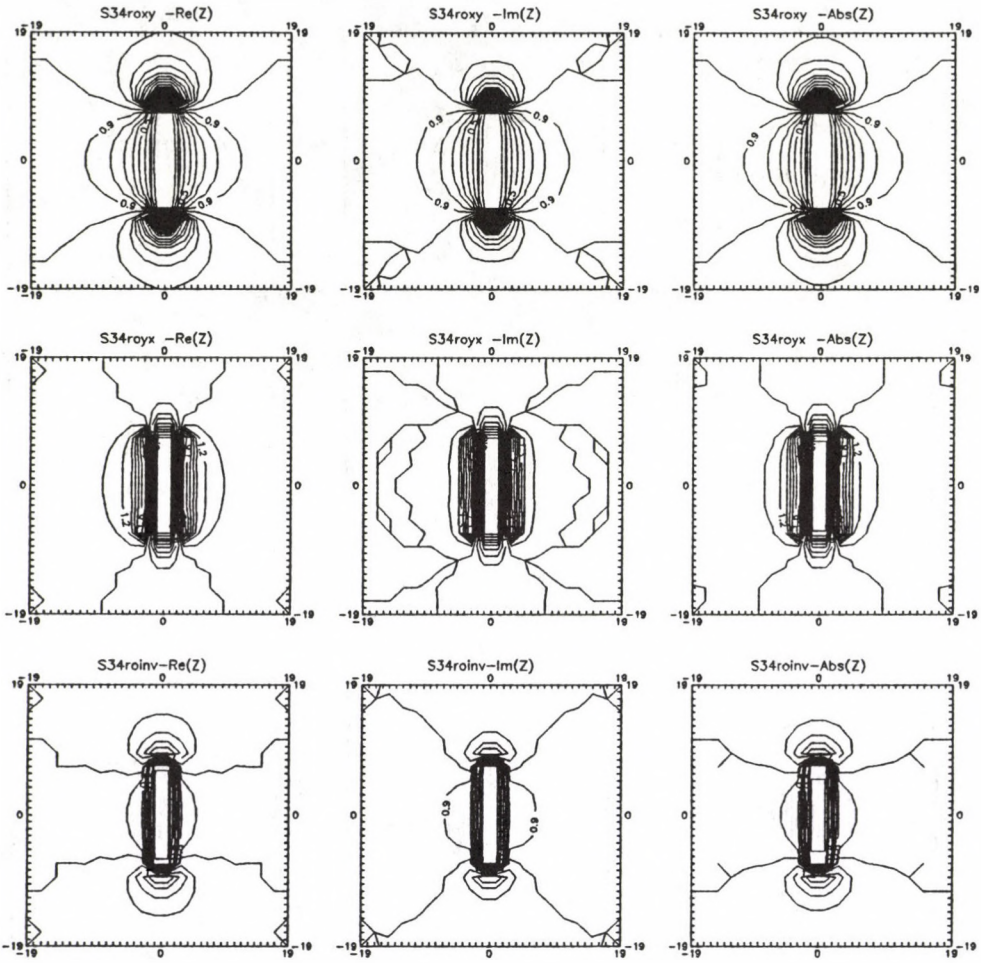

Fig. 9. $\varrho_{xy}, \varrho_{yx}$ and ϱ_{inv} maps by using three different resistivity definitions for model S at a depth of 3.5 km, period 56.25 s

2. the difference between the three invariant apparent resistivities due to the 3-D models is insignificant (practically less than 1 percent) off the high-conductivity structure. Remember that for this second case $\varrho_{xy(ReZ)}$ and $\varrho_{yx(ReZ)}$ are always more distorted than $\varrho_{xy(ImZ)}$ and $\varrho_{yx(ImZ)}$.

At depth, the resistivity increase outside a more elongated model (that is $a \gg b$) is definitely higher than for a more isometric one (that is $b \sim a$). It is due to the more concentrated charge-accumulation effect along the shorter ends, acting as main current inflow-outflow sites. These results are confirmed by the other maps we computed, but not published here.

Table I. Collection of different models shown in Figs (6–9) having the same $\lambda_1/(h-z)$ values

figure	model name	depth	period	$\lambda_1/(h-z)$
6	S01	0	3600	300
7	S12	2000	900	300
8	S23	3000	225	300
9	S34	3500	56.25	300

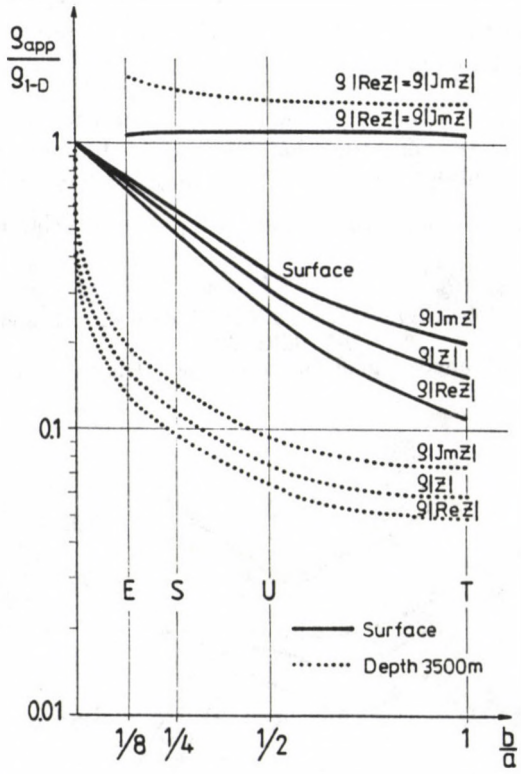

Fig. 10. Effect of 3-D thin sheet models on invariant ϱ_{ReZ}, ϱ_{ImZ} and $\varrho_{|Z|}$ resistivity values at 225 s over their centre ($\varrho_{\rm app}/\varrho_{1-D} < 1$ values) and outside them ($\varrho_{\rm app}/\varrho_{1-D} > 1$ values) at the surface (thick lines) and at a depth by using 3.5 km (dotted lines), as a funtion of the elongation parameter (b/a, where $a = 16$ km) of the model

5. Field example

It will be demonstrated by two field examples that the use of ϱ_{ReZ} does help in the interpretation of sounding curves and in the recognition of 3-D distortions.

Two sets of broadband MT sounding curves have been chosen from the Pannonian basin to illustrate the numerical calculations by field examples:

1. MT site *"Kisigmánd"*, from the Transdanubian area (Western Hungary). It probably lies at the rim of the Transdanubian crustal conductor due to strongly anisotropic graphitic formations (Ádám and Varga 1990).

   The sounding curves (Fig. 11) calculated by using Eqs (1), (2) and (3) are separated differently from each other in the case of the $\varrho_{\max}$ and $\varrho_{\min}$ curves. (In this section $\varrho_{\max}$ and $\varrho_{\min}$ are defined in 2-D terms.) Much clearer difference appears in the case of the $\varrho_{\min}$ curves. The indications of the conductivity anomalies both in the crust, and in the upper mantle more definitely appear in ϱ_{ReZ} curves. According to statements based on numerical methods, the resistivity of the conductor is better approximated by this curve.

   It is interesting to note the "collapse" of the $\varrho_{\max}$ curves is certainly due to the conducting block boundary.

2. *"Turkeve-Csodaballa"* is a deep sounding point in the Great Hungarian Plain, covered by thick sediments and characterized by shallow conducting asthenosphere due to the high heat flow of the Pannonian basin (Ádám et al. 1989).

   The extreme curves are here also shown separately (Fig. 12a and b).

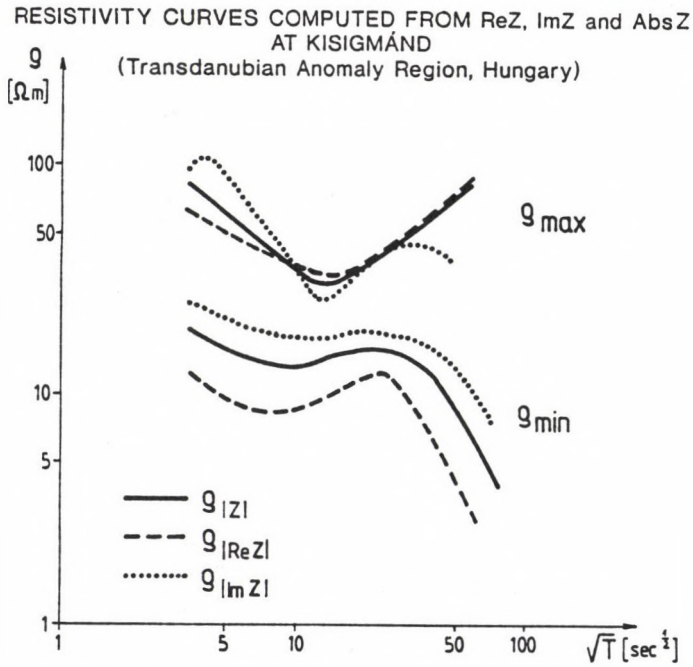

Fig. 11. ϱ_{ReZ}, ϱ_{ImZ} and $\varrho_{|Z|}$ extreme MT sounding curves at "Kisigmánd" measured in the area of the Transdanubian upper crustal conductivity anomaly

At the first glance the ϱ_{ReZ} curves differ from the traditional $\varrho_{\min}$ and $\varrho_{\max}$ (determined from $\varrho_{|Z|}$) curves on basis of the following important characteristics:

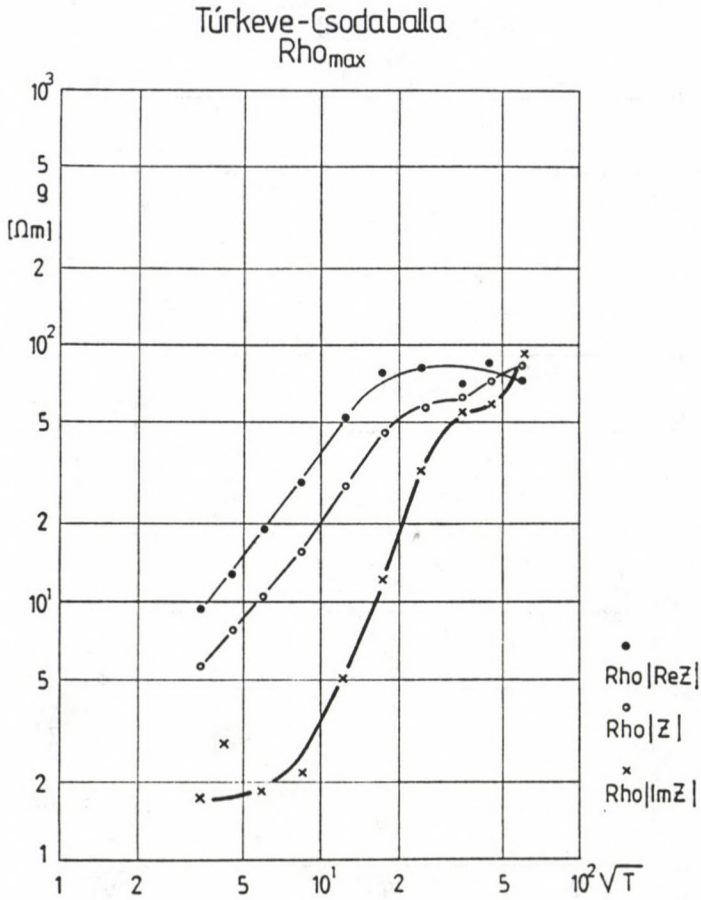

Fig. 12. a. The same curves as in Fig. 11 for Turkeve-Csodaballa (Great Hungarian Plain) in a sedimentary basin with shallow (~ 55 km) asthenosphere: ϱ_{max} curves

— their apparent resistivity level is the highest among them according to the highly resistive basement of the sedimentary basin;

— the indication of the asthenospheric conductor is much stronger than in the case of ϱ_{ImZ} and $\varrho_{|Z|}$ curves.

These field examples — nevertheless in a more complicate way — prove that attention should be paid to the information provided by the comparison of ϱ_{ReZ}, $\varrho_{|Z|}$ and ϱ_{ImZ} curves.

Conclusion

In this paper a brief summary is given about the main depth- and areal characteristics of ϱ_{ReZ}, ϱ_{ImZ} and $\varrho_{|Z|}$.

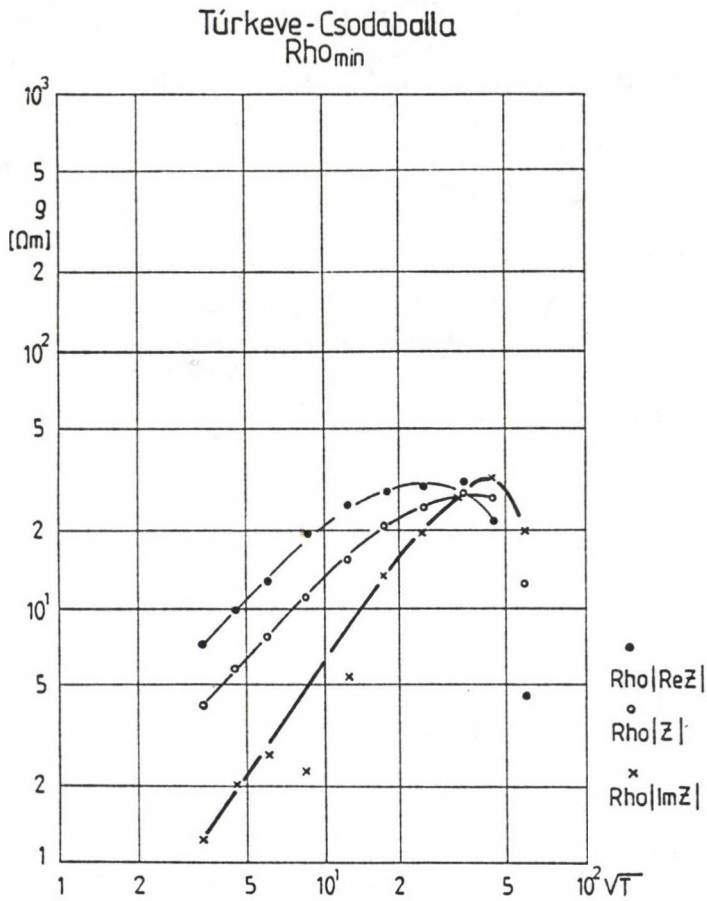

Fig. 12. b. The same curves as in Fig. 11 for Turkeve-Csodaballa (Great Hungarian Plain) in a sedimentary basin with shallow (~ 55 km) asthenosphere: $\varrho_{\min}$ curves

In 1-D when the observation depth is comparable to the depth of the investigated layer boundary (in case of Model 1), the difference between ϱ_{ReZ} and ϱ_{ImZ} at a depth is much greater than at the surface. In case of a very deep conducting layer (Model 2), when the layer boundary is far below the deepest observation depth, all near-surface effects will disappear in ϱ_{ReZ} at much shorter periods than in ϱ_{ImZ}.

In 3-D situations, the central anomaly and the spatial overshooting effects are more expressed, and more increasing with depth on ϱ_{xy} and ϱ_{yx} apparent resistivities maps based on ReZ than on the other maps. By using invariant resistivity maps the difference between the different apparent resistivity definitions (1), (2) and (3) in the overshooting areas perfectly disappears, and in the central parts is significantly reduced.

Asthenospheric effects can be studied by using ϱ_{ReZ} at much shorter periods than by using any other apparent resistivity definition. Unfortunately it is just ϱ_{ReZ}

which is distorted by 3-D crustal structures by the largest amount. This distortion can be reduced by using invariant resistivity maps but cannot be perfectly avoided in the vicinity of 3-D crustal structures.

We saw that in spite of their different relation to the subsurface current systems there is no striking difference between different apparent resistivity definitions in 3-D. We should not forget that outside 3-D crustal models it is always the surrounding 1-D medium which makes the difference between the levels of ϱ_{ReZ} and ϱ_{ImZ} and not the 3-D crustal structure itself.

According to the field examples the use of ϱ_{ReZ} may help in two respects:

1. A mutual comparison of ϱ_{ReZ}, ϱ_{ImZ} and $\varrho_{|Z|}$ gives more information than any apparent resistivity definition alone;

2. The greatest investigation depth can be reached by using ϱ_{ReZ} which might be important in case of period limitations.

Appendix

It is well known that in 1-D the subsurface impedance can be easily calculated by neglecting the layers above the observation level. The key parameters are $\lambda_1/(h-z)$ (where λ_1 is the wavelength in the first layer, and z is the observation depth in the first layer) on the one hand, and the conductivity 'reflection' coefficient s on the other hand. Since in many-layered MT there are as many $\lambda_i/(h-z)$ values as the number of layers, the situation will be illustrated by the simple two-layered (ϱ_1, h_1, ϱ_2) half-space. The λ_1, α and s key parameters are then expressed by:

$$\lambda_1 = \frac{2\pi}{\sqrt{\frac{\omega\mu\sigma_1}{2}}} \qquad \alpha = \frac{4\pi(h-z)}{\lambda_1} \qquad s = \frac{\sqrt{\sigma_1}-\sqrt{\sigma_2}}{\sqrt{\sigma_1}+\sqrt{\sigma_2}}$$

By using simple exponential and trigonometric functions of $\lambda_1/(h-z)$ and s, ϱ_{Rez}, ϱ_{ImZ} and the phase can be easily written in terms of α and s:

$$\begin{aligned}
\varrho_{ReZ} &= \varrho_1 Re^2\left(Z_1\sqrt{\frac{2\sigma_1}{\omega\mu}}\right)\\
\varrho_{ImZ} &= \varrho_1 Im^2\left(Z_1\sqrt{\frac{2\sigma_1}{\omega\mu}}\right)\\
\text{phase} &= \arctan\frac{ImZ_1}{ReZ_1}
\end{aligned}$$

where

$$ReZ_1\sqrt{\frac{2\sigma_1}{\omega\mu}} = \frac{e^{2\alpha}+2se^{\alpha}\sin\alpha - s^2}{e^{2\alpha}-2se^{\alpha}\cos\alpha + s^2}$$

$$ImZ_1\sqrt{\frac{2\sigma_1}{\omega\mu}} = \frac{e^{2\alpha}-2se^{\alpha}\sin\alpha - s^2}{e^{2\alpha}-2se^{\alpha}\cos\alpha + s^2}$$

For a certain conductivity contrast s, ϱ_{ReZ}, ϱ_{ImZ} and the phase are determined exclusively by α.

References

Ádám A, Landy K, Nagy Z 1989: *Tectonophysics*, 164, 361–368.

Ádám A, Varga G 1990: *Phys. Earth. Planet. Int.*, 60, 80–88.

Berdichevsky M N 1968: Electrical prospecting with the method of magnetotelluric profiling (in Russian). Nedra, Moscow

Schmucker U 1970: Anomalies of geomagnetic variations in the Southwestern United States. Bulletin of the Scripps Institution of Oceanography. University of California, San Diego, Vol. 13, p. 69.

Spies B R, Eggers D E 1986: *Geophysics*, 51, 1462–1471.

Szarka L, Fischer G 1989: *Geophysical Transactions*, 35, 157–172.

Szarka L, Fischer G 1991: *Geophysical Transactions*, 37, 25–38.

Szarka L, Menvielle M, Tarits P, Ádám A 1994: A numerical study of the electromagnetic field over geometrically complex high-conductivity structures: The field components and their relation with some 3-D MT interpretation parameters. *Acta Geod. Geoph. Hung.* (present issue)

Tarits P 1989: Contribution des sondages électromagnétiques profonds dans l'étude du Manteau Supérieur terrestre: These d'Etat, Paris VII, 333 pp.

Vasseur G, Weidelt P 1977: *Geophys. J. R. Astr. Soc.*, 51, 669–690.

Acta Geod. Geoph. Hung., Vol. 29(1–2), pp. 125–138 (1994)

A THIN SHEET NUMERICAL STUDY OF THE ELECTROMAGNETIC FIELD OVER GEOMETRICALLY COMPLEX HIGH CONDUCTIVITY STRUCTURES: ON THE CURRENT CHANNELLING IN HIGH CONDUCTIVITY 3-D MODELS

L SZARKA[1], M MENVIELLE[2], P TARITS[3], A ÁDÁM[1]

[Manuscript received February 15, 1994]

The present paper is a part of a series of papers dealing with thin sheet modelling studies of the electromagnetic field over geometrically complex models. It is shown by a figure compiled from numerous model computations, how the observed current direction (determined by connecting the centers of the resistivity overshooting areas over the opposite model sides) over elongated rectangular models follows the rotation of the external inducing field. A nonlinear relationship was obtained, characterized by preferred current directions pointing parallel to the longer model sides (in case of very narrow models) or to the diagonals (in case of square-like models). The $H_z = 0$ lines closely follow the internal current direction for any rotation angle. This close relationship between the $H_z = 0$ lines (giving information about the main current direction) and the connecting line of the resistivity overshooting areas are preserved in case of any, in geometric meaning much more complicated 3-D models. In this way the $H_z = 0$ lines (H_z is known as a typical "charge-free" 2-D indicator) can be interpreted in 3-D as lines, pointing toward the accumulation centers of surface charges.

The fact that the internal current direction seems to be fastened to the longer sides, emphasises the importance of current channelling and degrades the validity of H-polarization-like approximations over elongated dyke-like models.

Keywords: current channelling; magnetotellurics; numerical modelling; thin sheet; three-dimensional models

Introduction

Although 3-D MT anomalies in basic field directions are known from different numerical model computations, practically no attention has been paid to some special characteristics of the rotated maps. Sometimes — largely due to uncleared sign- and related phase problems — the rotated maps might even be handled erronously. By studying rotated resistivity-, phase and tipper maps, in this paper remarkable

[1]Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences, H–9401 Sopron, POB 5, Hungary

[2]Laboratoire de Physique de la Terre et des Planetes (URA 1369), Bat. 504 – Université Paris Sud, F–91405 Orsay Cedex, France

[3]Université de Bretagne Occidentale Laboratoire de Geosciences Marines (URA 1278), 6 Avenue Le Gorgeu, F–29200 Brest, France

1217-8977/94/$ 4.00

relations between 3-D resistivity- and H_z (or tipper element) maps will be pointed out. The observed relations help us in checking the correctness of rotation, and — by means of a clear physical insight — they may also help in the MT interpretation over geometrically complicated 3-D structures.

Two main sets of models were concerned. Both sets represent possible upper crustal structures from the Pannonian Basin (Ádám 1992). The resistivity and depth values are in conformity with those in our earlier papers (Szarka et al. 1994a, 1994b). All parameters (resistivities, depths, grid dimensions and plan views) are shown in Fig. 1.

The first set contained different elongated rectangles, denoted by E, S, U and T. In the second set the geometrical complexity varied from a simple rectangle (S) via an L-shaped model (L) to an asymmetric U-shaped form (Y).

In our first paper (Szarka et al. 1994a) the electromagnetic field components, several 3-D magnetotelluric parameters (invariants, skew, ellipticity, *etc.*) were discussed. In the second paper (Szarka et al. 1994b) the characteristics of different resistivity definitions at the surface and at depth, and near 3-D thin sheet inhomogeneities, based on thin sheet numerical computation were studied.

The thin sheet algorithm we used is based on results by Vasseur and Weidelt (1977); and was developed and the program itself was made by Tarits (1989). This technique is a very efficient numerical modelling method, offering almost a unique tool to study the electromagnetic field over 3-D structures with small vertical over horizontal dimension ratio.

After summarizing the rotation equations for the resistivity and the tipper, a series of rotated resistivity- and tipper maps will be shown over rectangular thin sheet models. The main current direction within the rectangular models will be defined by the connecting line between the so called overshooting (or "side") anomalies, appearing outside the models. After demonstrating a close relation between the angle of rotation and the observed current direction, it will be pointed out that the $H_z = 0$ lines should cross these overshooting regions in case of any complicated model, at any rotation angle.

Rotation definitions

In the paper Cagniard's apparent resistivity definition is consequently used. The x axis of the coordinate system points to North, y points to East. ϱ_{xy}, ϱ_{yx}, φ_{xy} and φ_{yx} mean resistivities and phases, belonging to the two different basic field directions. We follow notations suggested by Hobbs (1993).

It is known that ϱ_{xy} and ϱ_{yx} resistivity maps (which correspond to the two basic — perpendicular to each other — field directions) over such high conductivity rectangular models have significant minimum. The minimum zones are elongated parallel to the external magnetic field. In these ϱ_{xy} and ϱ_{yx} maps, outside the models, at the two opposite ends two slight resistivity enhancements, appearing as "overshooting areas" can be seen as well. (The overshooting areas are direct consequences of current inflow and outflow regions, giving information about the

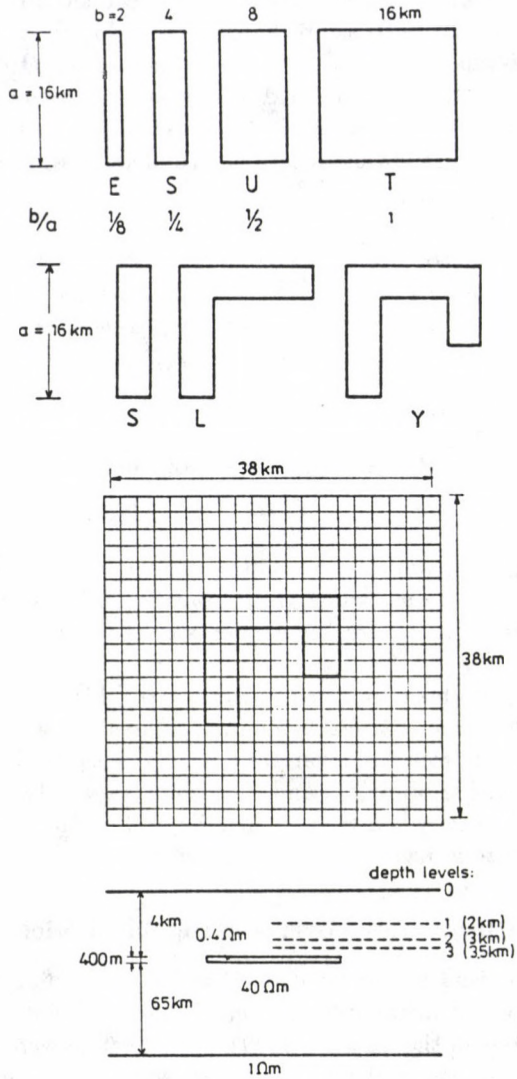

Fig. 1. Top: Plan views of 3-D thin sheet models used in this study. E, S, U and T denote different rectangular models; S, L and Y mean a model set of increasing geometrical complexity. Bottom: Plan view of the thin sheet computation grid and a characteristic model cross section showing all resistivity and geometric parameters

charge accumulation sites at the model edges. The buildup of the central minimum and the side maxima slightly depend on the period.)

H_z is known as a typical so-called 'E-polarization' indicator: it appears in 2-D, when the electric field, parallel to the 'strike' has lateral variation in the so called 'dip' direction. This simple 2-D rule can be extended for appropriate regions of elongated 3-D structures. Since H_z is usually normalized by the horizontal magnetic

field, instead of H_z here the tipper is used. (The denominator, the horizontal magnetic field is considered also as an 'E-polarization' indicator.)

In order to study what happens, when the inducing magnetic field is not parallel to the symmetry axes, the inducing field was rotated in many directions, denoted by Θ, where Θ is a clockwise-directed, North(x)-based rotation angle. For the impedance elements the following rotation formulae were applied (Berdichevsky 1968):

$$Z_{11} = Z_{xx}\cos\Theta\cos\Theta + Z_{xy}\sin\Theta\cos\Theta + Z_{yx}\sin\Theta\cos\Theta + Z_{yy}\sin\Theta\sin\Theta \quad (1)$$
$$Z_{12} = -Z_{xx}\sin\Theta\cos\Theta + Z_{xy}\cos\Theta\cos\Theta - Z_{yx}\sin\Theta\sin\Theta + Z_{yy}\sin\Theta\cos\Theta \quad (2)$$
$$Z_{21} = -Z_{xx}\sin\Theta\cos\Theta - Z_{xy}\sin\Theta\sin\Theta + Z_{yx}\cos\Theta\cos\Theta + Z_{yy}\sin\Theta\cos\Theta \quad (3)$$
$$Z_{22} = Z_{xx}\sin\Theta\sin\Theta - Z_{xy}\sin\Theta\cos\Theta - Z_{yx}\sin\Theta\cos\Theta + Z_{yy}\cos\Theta\cos\Theta \quad (4)$$

and for the tipper elements:

$$Z_1 = Z_x\cos\Theta + Z_y\sin\Theta \quad (5)$$
$$Z_2 = -Z_x\sin\Theta + Z_y\cos\Theta \quad (6)$$

where Z_{xx}, Z_{xy}, Z_{yx}, Z_{yy} are elements of the impedance tensor expressed in terms of the two basic polarizations, while Z_{11}, Z_{12}, Z_{21}, Z_{22} are the rotated tensor elements. Z_x, Z_y and Z_1, Z_2 are the elements of the magnetic transfer function before and after rotation.

For the simplicity, in this paper, parameters, derived from the impedance tensor element Z_{12} and the magnetic transfer function element Z_2 will only be dealt with. (Z_{21} can be easily given from the invariant $c_2 = Z_{12} - Z_{21}$ and Z_{12}; Z_1 can also be obtained from Z_2. Furthermore, Z_{12} and Z_2 are normalized by the same horizontal magnetic field component: when $\Theta = 0$, then $Z_{12} = Z_{xy}$ and $Z_2 = Z_y$. Z_{21} and Z_1 have another normalizing magnetic field component.)

Observations over rectangular models

The inducing EM field was rotated by $\Theta = 5 \cdot n$ ($n = 0, 1, 2, \ldots$) degrees over four differently elongated rectangular models denoted by E, S, U and T shown in Fig. 1. Several periods in the T = 14–3600 s period range were applied.

In Fig. 2 rotated $\varrho_{12}, \varphi_{12}$ and ImZ_2 maps are shown over one of the rectangular models, in case of rotation angles Θ = 15, 30, 45, 60 and 75 degrees. The actual period of the MT field is 225 s. (The areal distribution of the real and imaginary parts of the tipper elements are nearly the same, so Z_2 is well represented by its imaginary part.)

a) Impedance tensor element maps

In all the rotated resistivity- and phase maps we determined the direction of the connecting line between the center of the two resistivity enhancements (or "overshooting areas") outside the model. This direction approximates the overall current direction within the model, since these overshooting areas are direct consequences

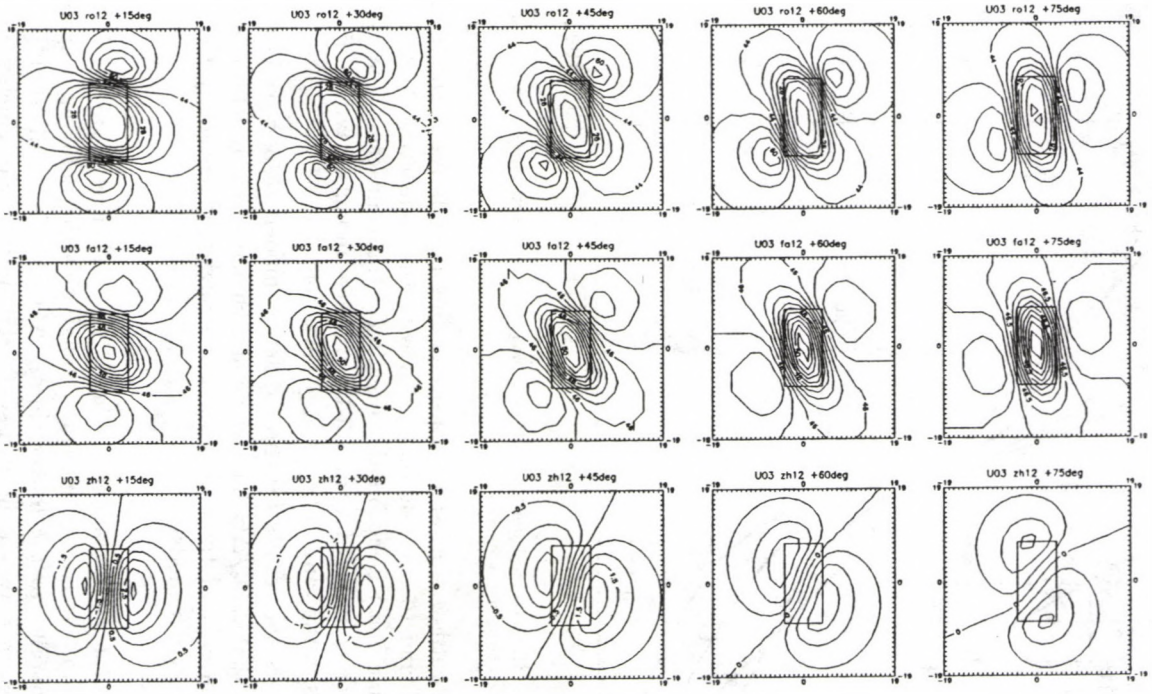

Fig. 2. Rotated resistivity-, phase- and tipper element (ϱ_{12}, φ_{12} and Z_2) maps over model U. Rotation directions are determined by $\Theta = 15, 30, 45, 60$ and 75 degrees

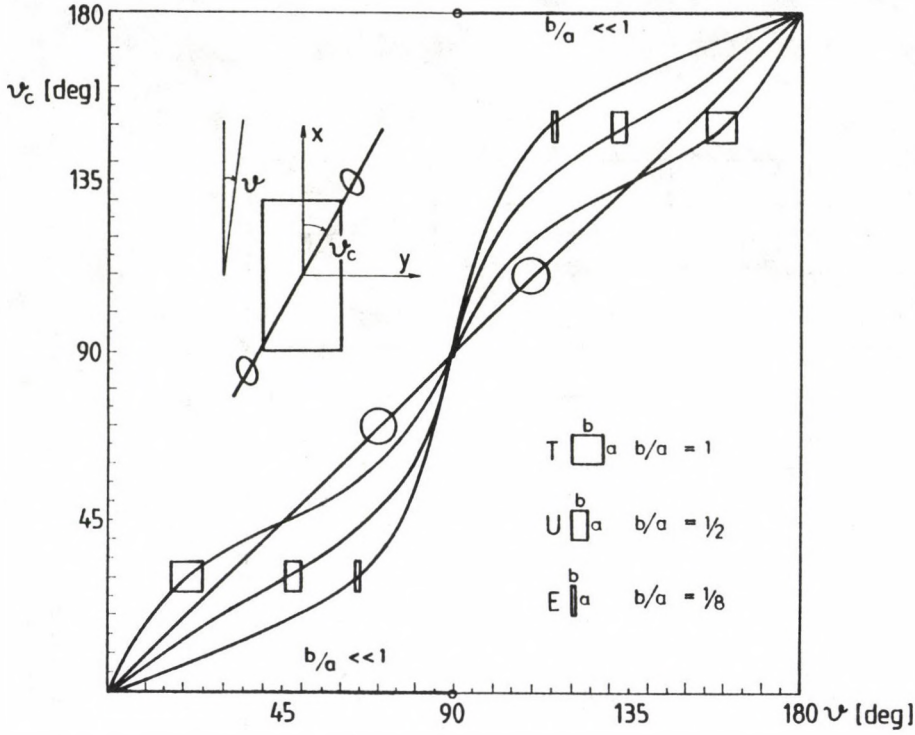

Fig. 3. Relation between the external field direction Θ and the observed current direction Θ_c within differently elongated thin sheet models

of current inflow and outflow sites at the corresponding boundaries of the model. (This direction was found to coincide quite well with the connecting line of the corresponding phase minima.) Denoting the direction of this connecting line by Θ_c, a close relation between Θ and Θ_c for these 3-D thin sheet models, having different elongation parameters in the range $1/8 \leq b/a \leq 1$ was found.

Figure 3 summarizes the results for three different models ($b/a = 1$, $1/2$ and $1/8$). It can be seen that $\Delta\Theta_c/\Delta\Theta$ has systematic changes as a function of the external field direction Θ; and these changes are strongly influenced by the b/a parameter of the rectangular models.

— *The most elongated model* (model E, $b/a = 1/8$). When Θ is small (that is the direction of the electric component belonging to the inducing magnetic field is nearly parallel to the longer sides), $\Delta\Theta_c/\Delta\Theta$ is small, that is the connecting direction between the two slight resistivity enhancements rotate very slowly. (The two resistivity enhancements seem to be fastened to the two shorter ends.) When the external electric field is becoming more and more perpendicular to the longer side, $\Delta\Theta_c/\Delta\Theta$ becomes higher, that is the change in the observed current direction will be extremely accelerated.

— *The square model* T (where $b/a = 1$) is perfectly antisymmetric to the $\Theta = \Theta_c$ line and if $\Theta = 45°$, then $\Theta_c = 45°$, too. It demonstrates that the most preferred current directions within the model are the two diagonal directions.

— The systematic slope change in $\Theta_c = f(\Theta)$ curve belonging to the model $b/a = 1/2$ confirms that the degree of current deviation within the model strongly depends on the external field direction. This curve lies between those belonging to $b/a = 1$ and $b/a = 1/8$. The most preferred current direction is not exactly the direction of the diagonals (it would be arctan 1/2), the preferred zone involves a wide range around the longer side of the model.

These relations can be easily extrapolated for circular- and very long and narrow ($b \ll a$) models as follows:

— In case of a perfectly circular model $\Theta_c = \Theta$ for any Θ, and the resulting curve would be a straight line, also indicated in Fig. 3, too.

— In case of very long ($b \ll a$) models the current would be perfectly channelled along the longer side. It means that over extremely elongated structures (when the current inflow and outflow boundaries are very far from each other) 'H-polarization' effects practically do not exist and always 'E-polarization' effects are dominant. This simple conclusion may degrade the validity of interpretation, based on 2-D H-polarization relations, since according to Fig. 3, H-polarization over dyke-like 2-D structures practically does not exist.

We mention that at any studied period, from both the resistivity and the phase distributions (for all resistivity definitions), a more or less similar relationship was observed. One exception, having no importance was found in case of the very long period (close to 0) phase values, when all curves seemed to be shifted toward the line $\Theta_c = \Theta$, which would correspond to the circular model.

b) Tipper maps

A similar feature was observed in the rotated tipper element maps: the Z_2 anomalies seem to be fastened much better to the longer sides than to the shorter ones.

Having made similar direction-determinations for the same models, a similar relation to that shown already in Fig. 3 was obtained: the $H_z = 0$ lines and the Θ_c directions were found to run very close to each other.

In Fig. 4 — in an unusual way — both the $H_z = 0$ lines and the slight resistivity enhancements are indicated for a rectangular model, where $b/a = 1/2$. Figure 4a demonstrates the situation at the surface, while in Fig. 4b the situation at depth is shown. It can be seen that the $H_z = 0$ lines coincide quite well with the connecting lines between the resistivity enhancements at any rotation angle.

About the theory behind

It is obvious that the two slight overshooting areas surrounding the high-conductivity models are apparently connected to the main current inflow- and outflow sites

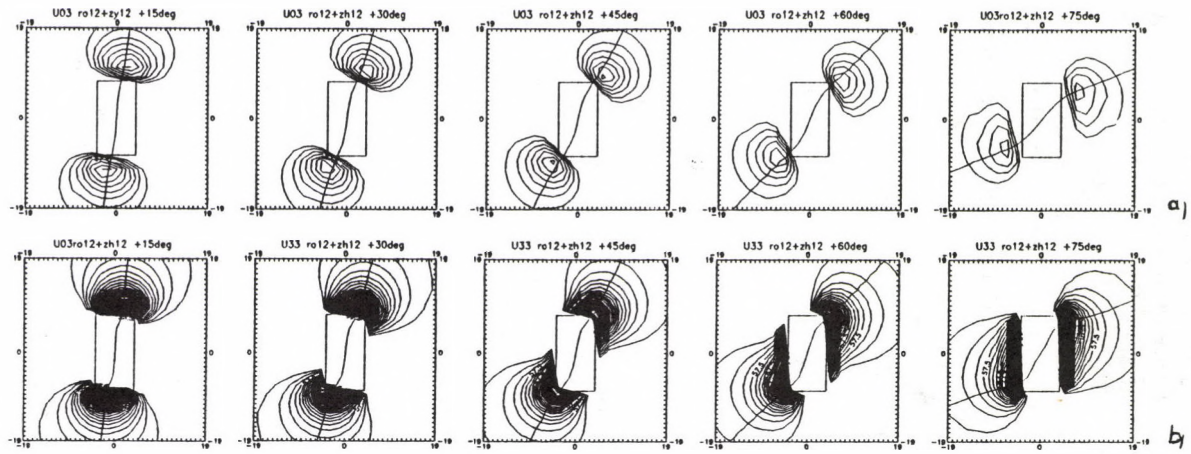

Fig. 4. Relation between the $H_z = 0$ lines and the connecting direction of the two resistivity overshooting areas (represented by isolines) over model U at several rotation degrees. a) $H_z = 0$ lines and resistivity overshooting isolines at the surface. b) $H_z = 0$ lines and overshooting isolines at a half-depth of model U (the model depth is 4 km, the observation depth is 2 km)

of 3-D models, which seem to be fastened to the shorter ends, or to the opposite corner regions of the rectangular models.

The main current inflow and outflow sites in a model are at the same time the main charge-accumulation sites, since the role of charges, appearing at resistivity interfaces is just to assure the continuity of currents at resistivity boundaries.

This charge accumulation takes place in a similar way as in the pure electrostatics: the induced charges appear in such a way that the total electromagnetic energy is minimum. Consequently the distance between the accumulated positive and negative charges on a conducting rectangle should be maximum. The electrostatic analogy can be regarded as a 0th approximation of the real situation.

In case of finite resistivity contrast, the effect of the resulting current flow should also be taken into account. According to Ohm's law, the current prefers to flow within the better conducting medium as long as "possible". So this phenomenon is not against the electrostatic charge accumulation process: in case of direct current flow the positive and negative surface charges prefer to be accumulated at the "possibly" longest distance from each other. It is known that over a straight horizontal current flow H_z is 0. Figure 4 shows that the $H_z = 0$ lines run along the centre zone of the main current path within the structure where the length of the current flow is "the possible longest".

In magnetotellurics the electromagnetic induction should also be taken into account, of course. The computed maps at different periods indicate that the whole situation only slightly changes due to the induction. According to our results and other well-known reports (e.g. Jones 1988) the contribution of current channelling in MT anomalies is very important.

Justification of the energy minimum requirement by more complex models

We did not use any mathematical formulae to demonstrate this energy-minimum requirement. It would be impossible to find close relations and such formulae perhaps would not solve any practical problem. Instead of using mathematical formulae, we try illustrate the correctness of this hypothesis on further (in geometric meaning more complex) thin sheet models.

In Fig. 5a ϱ_{12} resistivity maps (based on impedance element Z_{12} by using Cagniard's definition) are shown for 12 different inducing field directions ($\Theta = 0°, 15°, 30°, \ldots 165°$). In these maps special attention should be paid to the sometimes slowly, sometimes quickly changing resistivity overshooting areas, which may be divided or unified at the two shorter ends, depending on the actual value of Θ. In Fig. 5b H_z anomaly maps are shown over the same model. Their most characteristic features are the $H_z = 0$ lines, following what is taking place with the current flow. In Fig. 5c the close relation between the $H_z = 0$ lines and the positions of the resistivity overshooting areas is demonstrated by putting together some resistivity isolines and the $H_z = 0$ lines. The $H_z = 0$ lines cross the most significant $\varrho_{\max}$ sites in case of any rotation angle.

This close relationship between $H_z = 0$ lines and the resistivity overshooting

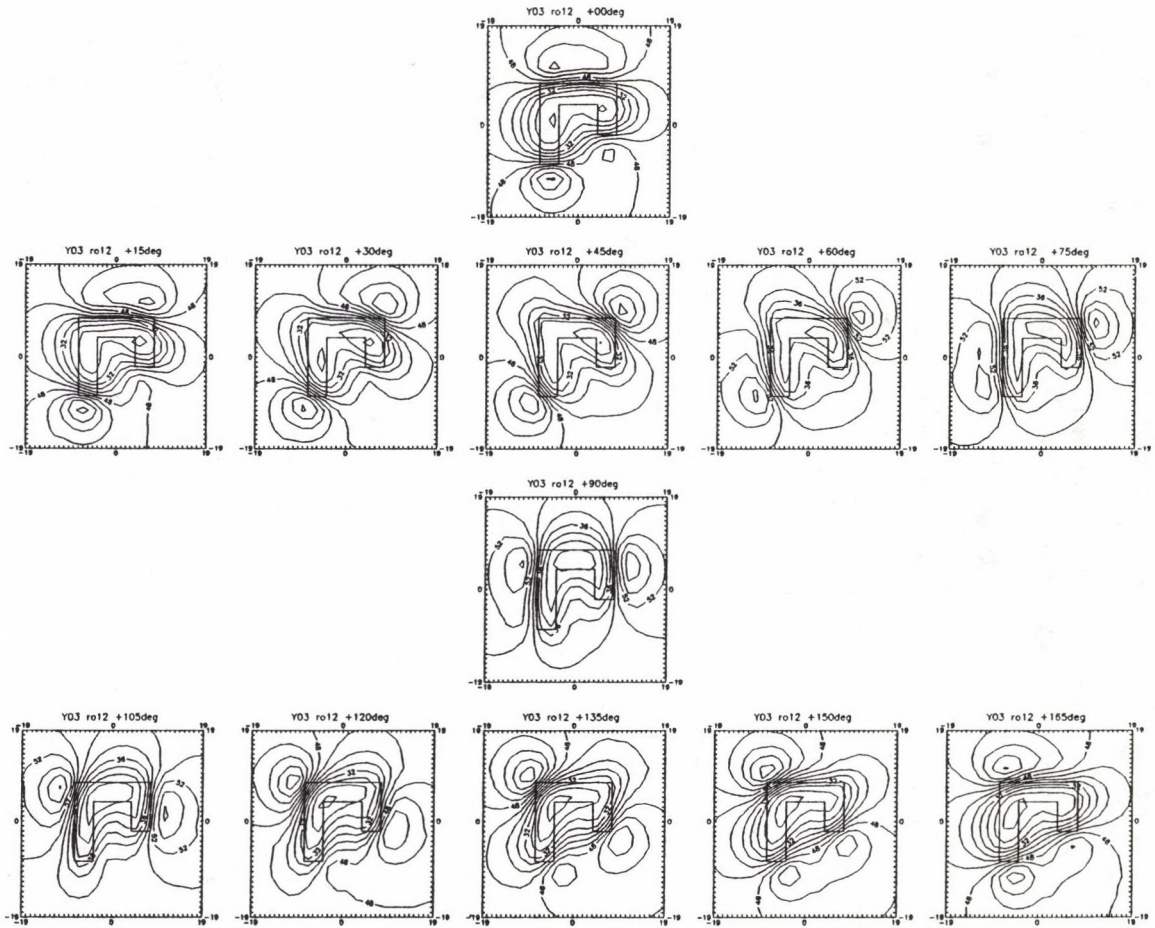

Fig. 5. a. Relation between the $H_z = 0$ lines and the connecting direction of the two resistivity overshooting areas over (in geometrical meaning) a more complex model (model Y). Rotation degrees are from 0° to 165° by step of 15°. Rotated ϱ_{12} maps in the $0° - 165°$ range

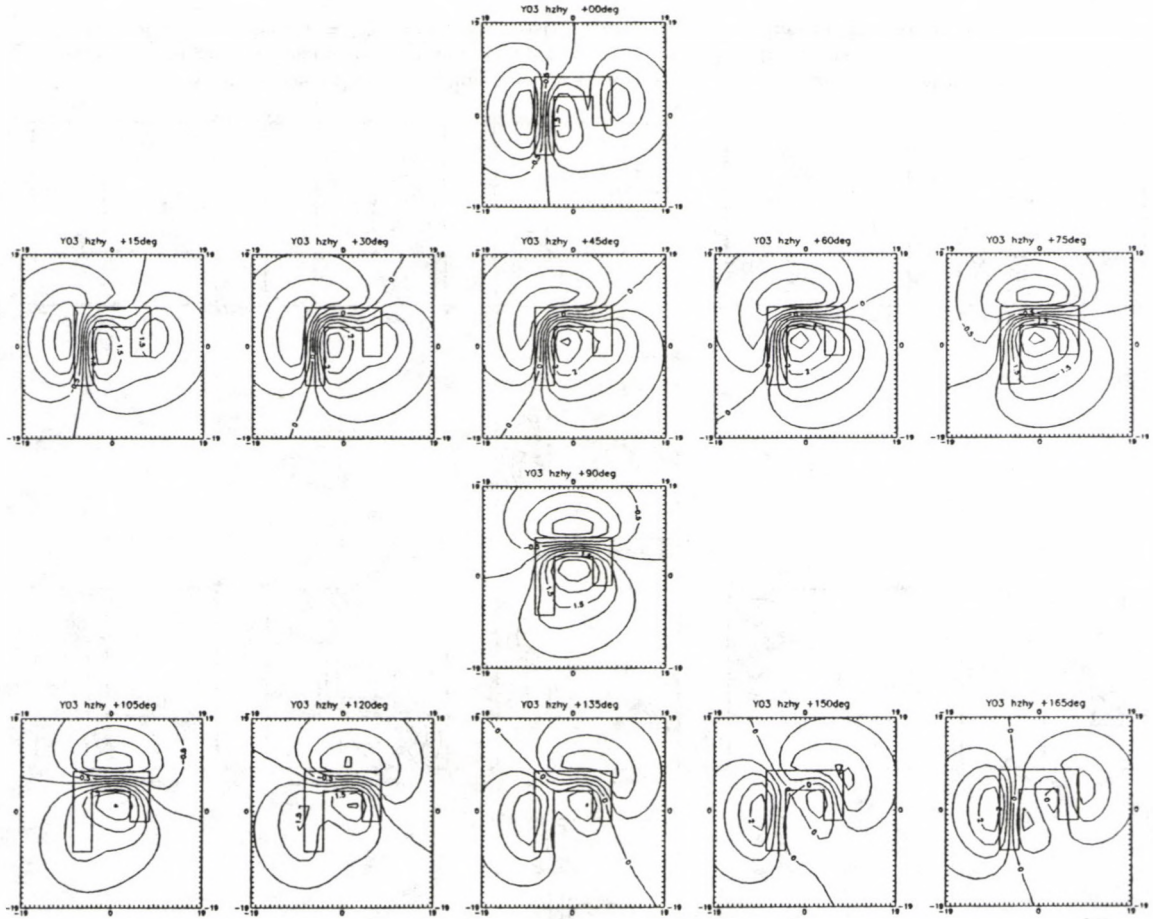

Fig. 5. b. Relation between the $H_z = 0$ lines and the connecting direction of the two resistivity overshooting areas over (in geometrical meaning) a more complex model (model Y). Rotation degrees are from 0° to 165° by step of 15°. b) Rotated $Im(Z_2)$ maps in the $0° - 165°$ range

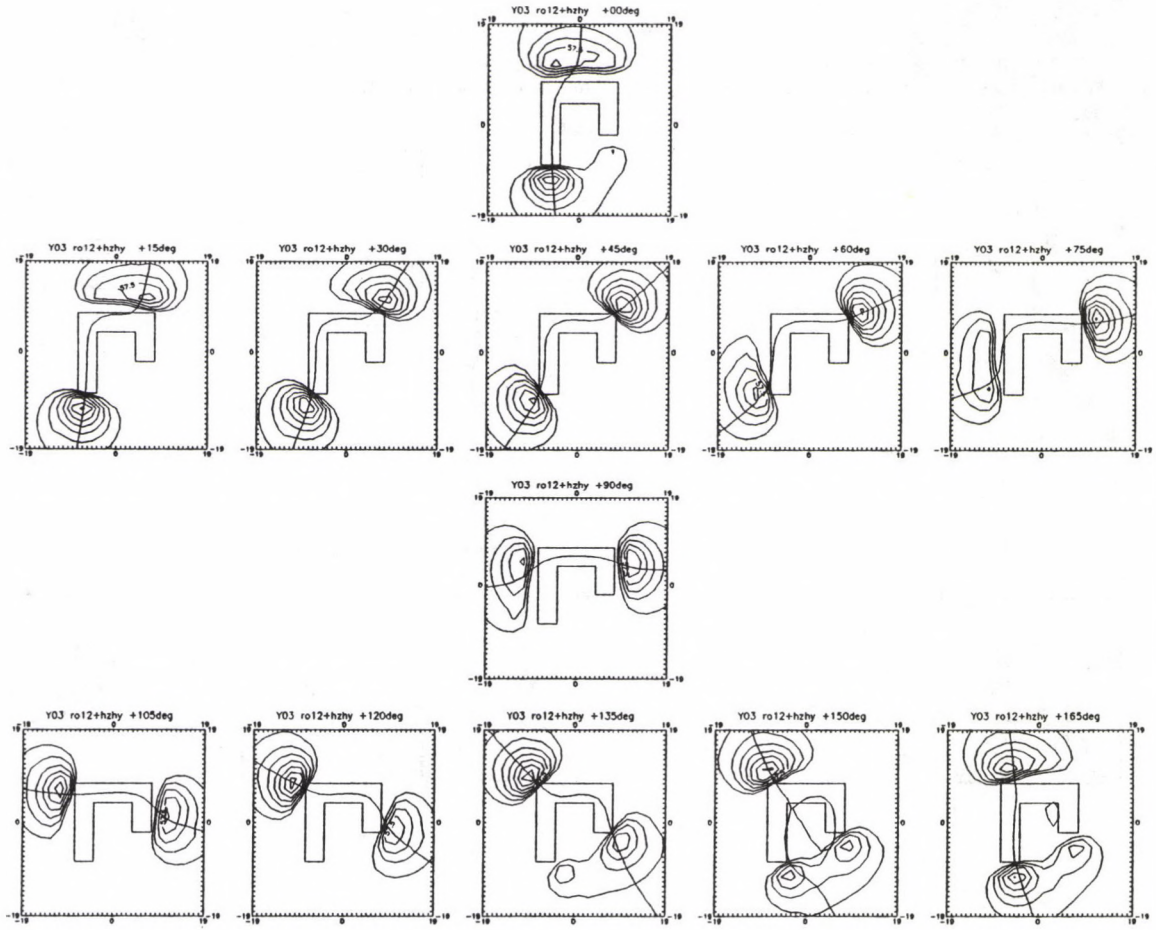

Fig. 5. c. Relation between the $H_z = 0$ lines and the connecting direction of the two resistivity overshooting areas over (in geometrical meaning) a more complex model (model Y). Rotation degrees are from 0° to 165° by step of 15°. c) The $H_z = 0$ lines and the resistivity overshooting isolines in the rotated maps

areas may give a valuable thumb-rule in MT interpretation of complicated 3-D structures.

Relation between resistivity maps and polar diagrams

The polar diagrams summarize mathematical values of some MT parameters (e.g. resistivities) for all possible inducing field directions. They contain pure mathematical information, regardless if a field direction can be observed in reality or not.

It was found that in case of inducing field directions parallel to the longer sides or to the diagonals the current direction changes very slowly; and when the inducing field is at right angles to the longer sides, the current direction changes extremely fast, and in case of very long models ($b/a \ll 1$) practically no current flow, perpendicular to the local strike direction, is available. It means that at the longer sides of elongated 3-D structures the "E-polarization" is a stable and the H-polarization is an "instable" situation.

Summary

In this paper the wandering of the observed current direction within 3-D models as a function of the external field direction in resistivity maps over rectangular and more complicated high-conductivitiy models was discussed.

Some basic relations (some of them can be applied as thumb rules) have been found as follows:

— A close relation was found between the external field direction and the current direction within the structure and discussed for different rectangular thin sheet models. It was found that the internal current direction seems to be fastened along the longer sides ($b/a \ll 1$) or to the diagonals ($b/a \approx 1$).

— For the same 3-D models the $H_z = 0$ lines and the connecting line of the slight resistivity enhancements should largely coincide, since both of them reflect the same phenomenon: the internal current flow between the surface charge accumulation sites at the two opposite model sides.

  In this way the $H_z = 0$ line (H_z is often regarded as indicator for the so-called 'charge-free' E-polarization), has become in 3-D situations an indicator of charge-accumulation. There is no contradiction with 2-D: in 'E-polarization' the $H_z = 0$ lines point toward the infinitely distant charges.

— All observed phenomena are preserved in case of complex models as well. It means that the close relation found between the $H_z = 0$ lines and the connecting line between the overshooting areas around the model may help in checking the corrections of the rotated MT maps.

— A direct conclusion of Fig. 3 is, that over extremely elongated models (like 2-D dykes) H-polarization does not exist, consequently all interpretation results obtained in such situations are questionable.

References

Ádám A 1992: *J. of China University of Geosciences*, 3, 92–104.

Berdichevsky M N 1968: Electric exploration by magnetotelluric profiling (in Russian). Nedra, Moscow

Hobbs B 1993: *Surveys in Geophysics*, 13, 489–515.

Jones A G 1988: *Geophysics*, 53, 967–978.

Szarka L, Menvielle M, Tarits P, Ádám A 1994a: *Acta Geod. Geoph. Hung.* (present issue)

Szarka L, Ádám A, Menvielle M, Tarits P 1994b: *Acta Geod. Geoph. Hung.* (present issue)

Tarits P 1989: Contribution des sondages électromagnétiques profonds dans l'étude du Manteau Supérieur terrestre: These d'Etat, Paris VII, 333 pp.

Vasseur G, Weidelt P 1977: *Geophys. J. R. Astr. Soc.*, 51, 669–690.

Acta Geod. Geoph. Hung., Vol. 29(1–2), pp. 139–148 (1994)

COMPARISON OF SOME UPPER-ATMOSPHERIC MODELS OF MARS

M J ILL[1]

[Manuscript received March 4, 1994]

Lifetimes of a fictitious artificial satellite of Mars were calculated to verify atmospheric models. In the computations first the density profiles obtained during the landing of Viking-1 and Viking-2 were used and then the density profiles given by other 6 different atmospheric models. Only 1 of the 6 models gave practically the same lifetime as observed in the case of Viking-1. However, none of the 6 models gave acceptable lifetimes compared with the lifetime using Viking-2 data. Consequently, these models still need important improvements.

Using Viking-2 data the obtained lifetime is 5.7 times longer than the one obtained using the data of the Viking-1 entry. A detailed analysis shows that the differences between the two profiles are significant and not negligible. The density profiles of both Viking landers contain local maxima and minima. The origin of these variations is unknown.

Keywords: artificial Satellites; atmosphere; dynamics; Mars; orbits

Introduction

Nowadays we hear more and more about space missions related to planet Mars. As a consequence of former missions, experts grew richer in knowledge concerning Mars, *e.g.* there exist already several models making efforts to describe some parts of Mars' atmosphere. However, examining closely these models it becomes evident that sometimes they differ to a high degree from each other (at some altitudes the differences between the vertical density profiles exceed 100 %!).

The structure of Mars' atmosphere was measured *in situ* by instruments (accelerometers, mass spectrometers) on board the two Viking landers. Among others, it was also possible to deduce two vertical density profiles from these measurements. This fact gave us the idea to compare the density profiles given by different models with those deduced from Viking measurements. The ratios of measured and calculated densities give varying pictures along the vertical profiles, but in this manner it is not easy to arrive to practical conclusions.

Nevertheless, it is well-known that the density of the atmosphere plays a decisive role among the factors having an influence on the lifetime of a satellite. Therefore the accuracy of a given model can be checked if we calculate the lifetime of a fictitious satellite of Mars by using first the Viking measurements and then the density profile of the model. The comparison of the obtained lifetimes clearly shows to which degree the model departs from the real atmospheric conditions measured during the entries of the Vikings.

[1]Konkoly Observatory of the Hungarian Academy of Sciences, Box 67, H-1525 Budapest, Hungary

1217-8977/94/$ 4.00

Method

In our computations first we used the density profiles measured during the landing of Viking-1 and Viking-2. The data of the Mars missions were taken from Seiff and Kirk (1977). Since in the case of Viking-2 we did not find density values for altitudes higher than 175 km, we chose $h_p = 175$ km as pericenter height and $h_a = 1000$ km as apocenter height of the orbit. Further initial orbital elements of this orbit are: orbital period $P_0 = 127.362$ min, semi-major axis $a_0 = 3980.901$ km and excentricity $e_0 = 0.10362$.

Our method of computation is as follows. From the initial values a_0 and e_0 we compute their changes Δa_0 and Δe_0 during one revolution. In this way we obtain from $a_1 = a_0 + \Delta a_0$ and $e_1 = e_0 + \Delta e_0$ a new pericenter height and a new orbital period P_1. After this, using a_1 and e_1 we compute again their changes Δa_1 and Δe_1 during one revolution, whereby we obtain further $a_2 = a_1 + \Delta a_1$ and $e_2 = e_1 + \Delta e_1$ determining a new pericenter height and the new orbital period P_2 *etc.* We stop the computations when the change of pericenter height during one revolution becomes larger than 40 km. In this case the satellite performs generally not more than 1 revolution, since this limiting case occurs around $h_p = 100$ km. On the base of the foregoings the lifetime of the satellite is the sum of the successive orbital periods: $T = P_0 + P_1 + P_2 + \ldots$

We compute the Δa and $\Delta(a \cdot e) = \Delta x$ values using the formulae published by King-Hele (1964):

$$\begin{aligned}\Delta a = \; & -(2\pi H/x)^{1/2} F a^2 D_p \left[1 + 2e + 3e^2/2 + H/8x + 9H^2/128x^2 - \right. \\ & \left. - 3H/4a + 3M/8(1 + 2e + 5H/8x + 105H^2/128x^2)\right]\end{aligned}$$

$$\begin{aligned}\Delta x = \; & -(2\pi H/x)^{1/2} F a^2 D_p \left[1 + 2e + 3e^2/2 - 3H/8x - 15H^2/128x^2 - \right. \\ & \left. - 3H/4a + 3M/8(1 + 2e - 15H/8x - 175H^2/128x^2)\right]\end{aligned}$$

where

a, e	=	orbital elements
D_p	=	density at altitude h_p
H	=	density scale height at h_p
M	=	gradient of H.

These formulae are valid when the scale height H varies with altitude, in the case if $e \leq 0.2$ and $B = a \cdot e/H \geq 3$.

Let the mass of our fictitious satellite be $m = 100$ kg, and its effective cross-section $A = 1$ m^2! Let us suppose that the value of the drag coefficient is $C_D = 2.0$, so in our case the value of the aerodynamic coefficient is $F = A \cdot C_D/m = 0.02$ m^2/kg.

In our computations we used densities derived from accelerometer and upper-atmospheric spectrometer data collected during the landing of the Viking probes. The data were derived and published in tabular form by Seiff and Kirk (1977). In order to facilitate the computations we applied a least squares fit to the density

data. Hereby we obtained the best fits by determining a logarithmic curve for the densities corresponding to altitudes not higher than the approximate turbopause height (120–130 km), and a second one for higher altitudes.

In Table I the densities measured by Viking-1 in the 175–96 km altitude range are presented, as well as the densities given by our logarithmic curves together with their differences from the observed data (O–C). In the case of Viking-1 the mean difference is ±2.96 % (with a maximum value of 8.34 % in the vicinity of the turbopause height).

Table I. Observed and calculated densitites

Altitude (km)	Density (kg/m^3) Viking-1	Calculated	O–C %
96	.2880E–06	.2755E–06	4.35
100	.1670E–06	.1730E–06	–3.58
104	.1060E–06	.1076E–06	–1.52
108	.6590E–07	.6631E–07	–0.62
112	.3950E–07	.4048E–07	–2.48
116	.2420E–07	.2448E–07	–1.16
120	.1600E–07	.1467E–07	8.34
130	.3800E–08	.3910E–08	–2.89
135	.1590E–08	.1464E–08	7.91
140	.7250E–09	.7476E–09	–3.12
145	.4100E–09	.3849E–09	6.13
150	.2410E–09	.2368E–09	2.09
155	.1480E–09	.1468E–09	0.79
160	.9350E–10	.9401E–10	–0.55
165	.6270E–10	.6127E–10	2.28
170	.4210E–10	.4207E–10	0.07
175	.2770E–10	.2875E–10	–3.81

$.2880E–06 = .2880 \cdot 10^{-6}$

In Table II the same data corresponding to Viking-2 are shown. Here we have a mean O–C of ±3.31 %. As we shall see later, the value measured at 144 km is a strongly irregular point of the density profile. If we omit his single point, the mean O–C decreases to ±2.7 %.

To use the formulae mentioned above we also need the density scale height H. Its value was calculated for the whole altitude range in steps of 4–5 km, using the well-known formula: $H = -\Delta h_p / \ln(D_2/D_1)$, where Δh_p is the difference in altitude and D_1, D_2 are the corresponding densities. The calculated scale heights were fitted by curves of the second degree.

Results

Our computations are performed according to the above mentioned method and supposing the initial conditions (h_p = 175 km, h_a = 1000 km). Using the density profile derived from Viking-1 measurements the pericenter height of our fictitious satellite decreased to 100 km after 26028 revolutions, corresponding to a lifetime of

Table II. Observed and calculated densitites

Altitude (km)	Density (kg/m^3) Viking-2	Calculated	O–C %
96	.2300E–06	.2291E–06	0.38
100	.1420E–06	.1472E–06	–3.68
104	.9360E–07	.9105E–07	2.72
108	.5620E–07	.5419E–07	3.58
112	.3080E–07	.3104E–07	–0.78
116	.1690E–07	.1711E–07	–1.24
120	.8860E–08	.9077E–08	–2.45
128	.2310E–08	.2277E–08	1.41
132	.1430E–08	.1397E–08	2.30
136	.8650E–09	.8558E–09	1.06
140	.4900E–09	.5213E–09	–6.39
144	.2750E–09	.3158E–09	–14.84
148	.1810E–09	.1902E–09	–5.11
152	.1130E–09	.1140E–09	–0.87
156	.6820E–10	.6791E–10	0.43
160	.4150E–10	.4023E–10	3.05
162	.3230E–10	.3091E–10	4.32
166	.1900E–10	.1816E–10	4.42
170	.1070E–10	.1061E–10	0.83
174	.5800E–11	.6166E–11	–6.32

$.2300E–06 = .2300 \cdot 10^{-06}$

2221.5 (terrestrial) days. But in turn, if we used the Viking-2 densities keeping all other conditions unchanged, the satellite was able to perform 149212 revolutions, *i.e.* its lifetime increased to 12769.6 days. We consider it very surprising that the use of two density profiles supposed to be only slightly different, results in such different lifetimes. For that very reason we considered interesting to determine the lifetime of the same satellite (maintaining the same initial conditions), but using density profiles of different models of Mars' atmosphere and to compare these lifetimes with those obtained using Viking-data. In these computations we used the density profiles of the NOMINAL, MINIMUM and MAXIMUM MODELS published by Sehnal (1990), as well as the NOMINAL, COLD and WARM MODELS published by Moroz et al. (1991). (We have no other model at our disposal for altitudes above 100 km). The obtained lifetimes are given in Table III.

Discussion

Comparing the lifetimes given in Table III it can be easily established that the lifetime calculated by the Viking-1 data is fairly well approximated by the Nominal Model of Sehnal. The difference is only 29.6 days or 340 revolutions, amounting to 1.3 %. Among the Moroz et al. models it is likewise the Nominal Model, which gave the best result. However, in this case the difference is already 492.4 days or 5789 revolutions. This difference amounts to 22.2 % and it can be considered as

Table III. Computed lifetimes of a fictitious satellite using different density profiles

Density profile	Lifetime in Days	Years	Number of rev.
Viking-1	2221.5	6.08	26028
Viking-2	12769.6	34.96	149212
Sehnal-Nom.	2191.9	6.00	25688
Sehnal-Max.	810.1	2.22	9494
Sehnal-Min.	6874.5	19.10	81739
Moroz-Nom.	1729.1	4.73	20239
Moroz-Warm	392.0	1.07	4594
Moroz-Cold	15708.3	43.01	184047

important (1.35 years!). This large deviation may be explained by the differences in the temperature profiles. In the Nominal Model (Moroz et al. 1991) the surface temperature is 210 K, at 130 km it decreases to 182 K and at 200 km it increases up to 211 K. The corresponding values of the Viking-1 profile are: 238 K, 120 K and 102 K, respectively.

The lifetime calculated using Viking-2 data is "best" approximated by the Cold Model of Moroz et al. (1991), but even in this case the difference is very important: the model overestimates the lifetime by 2938.7 days or 34855 revolutions. Using the Minimum Model of Sehnal the difference is still larger: it underestimates the lifetime by 5795.1 days or 67479 revolutions. These deviations show that above 100 km these model-densities differ significantly from those measured by Viking-2.

The fact that in the case of Viking-2 data we obtained a lifetime 5.7 times longer than by using the data of Viking-1, drove us to compare in detail the two density profiles. The densities published in Tables IV and V of Seiff and Kirk 1977 render a comparison in the 28–175 km range possible.

It can be established at first sight that the densities D_2 derived from Viking-2 measurements are always smaller than the corresponding values D_1 of Viking-1 (a single exception is at 60 km), but the measure of the difference varies with altitude. In Fig. 1 the ratio D_2/D_1 of the observed densities versus altitude is plotted. From this curve it appears that the ratio decreases with increasing altitude, but the most remarkable feature is the wave structure all along the curve (below the turbopause as well as above it).

Naturally, the large and variable difference between the Viking-1 and Viking-2 densities can also clearly be recognized by comparing the number of revolutions performed (or the lifetime) so long as the pericenter decreased by *e.g.* 5 km. In Table IV the number of revolutions performed in a given altitude range are given for both density profiles together with their ratios. These ratios vary with altitude as expected, namely they decrease in the intervals where the density ratios D_2/D_1 increase (see Fig. 1). The most conspicuous differences appear, of course, in the

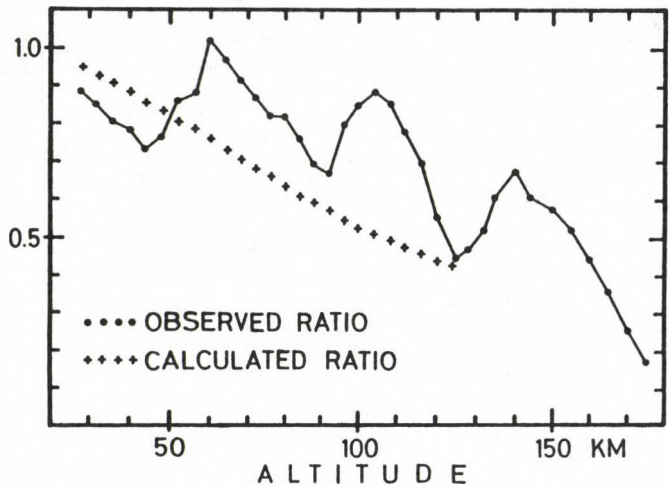

Fig. 1. Ratio D_2/D_1 of densities derived from measurements of Viking-1 and Viking-2 (dots). Crosses show the same ratio, calculated starting from the observed surface pressures and temperatures

175–150 km range where the differences between the two observed temperature profiles amount up to 67 K.

At the landing time of the Viking probes the surface pressures and temperatures were 7.62 and 7.81 mbar and 238 K and 226 K, respectively. In interpreting the accelerometer data hydrostatic equilibrium has been assumed by Seiff and Kirk (1977). Using the same hypothesis and starting from the observed surface pressures and temperatures we calculated the two "theoretical" density profiles up to 124 km, the supposed altitude of turbopause (Stewart 1987). The ratios of these calculated densities are marked by "+" crosses in Fig. 1. It is apparent that the general trend of the "observed" ratios is the same as that of the "calculated" ones, but the calculated curve is nearly linear and has no wave structure at all. Since we would like to know the cause of the wave structure of the curve obtained by using the Viking-data, it seems to be logical to search for answer in the difference in the landing conditions of the two vehicles.

It is well-known that Viking-2 landed 45 days after Viking-1, therefore the question arises whether the discovered differences between the two density profiles originate from a seasonal effect. Taking into account that the 45 day interval is shorter than 1/3 of the shortest Martian season, it is reasonable to suppose that this short time-difference cannot cause large changes in the density profiles. However, at the computation of our density profiles we started with the measured surface data, so the profiles contain also the seasonal change, but the curve has no wave structure. The fact that the difference in chemical composition of the atmosphere during the two landings was small, speaks also against the decisive role of a seasonal effect. On the basis of all these we consider the seasonal effect as inadequate to explain the wave structure of the curve.

Table IV. Number of revolutions performed in a given altitude interval

Alt. range (km)	Viking-1 R1	Viking-2 R2	Ratio R2/R1
175–170	10320	92500	8.96
170–165	6640	34830	5.25
165–160	4030	13186	3.27
160–155	2355	5196	2.21
155–150	1320	2066	1.57
150–145	710	878	1.24
145–140	368	301	0.82
140–135	184	145	0.79
135–130	72	66	0.92
130–125	17	27	1.59
125–120	7	10	1.43
120–100	5	7	1.40

Viking-2 landed 6 hours earlier (in Martian local time) than Viking-1. According to several authors (*e.g.* Stewart 1987, Moroz et al. 1991) on the surface of Mars there exists a diurnal effect, the amplitude of which decreases rapidly with altitude and above 100 km it is already negligible. Therefore the wave structure cannot be explained as the consequence of a diurnal effect.

Neither the 10.7 cm solar flux can be taken into account because its intensity was 69.4 units in the first case and 75.7 units at the second entry. This difference is too small to explain the fact that Viking-2 densities at 175 km were roughly 5 times smaller than the corresponding values of Viking-1.

Finally, we remark that the two entries were separated by 178° in longitude and 25° in latitude. Although the latter causes certainly some differences in temperature and pressure (taken into account by using the measured surface data), we can disregard it as the cause of the wave structure.

In summary, taking into account all known differences between the landing conditions of the two vehicles, we have no explanation for the observed large variations of the density ratios in Fig. 1. Accordingly, our final conclusion is that the wave structure can most probable be attributed to developing of internal gravity waves. This idea seems to be supported by Fig. 2 which shows the ratio of the observed density D_1 (Viking-1) and the corresponding density given by the Nominal Model of the COSPAR Mars Reference Atmosphere. The same ratio for Viking-2 is also given in Fig. 2. Both normalized curves show a considerable wave structure. The local density maxima and minima of the normalized curves of Fig. 2 permit the determination of the apparent wavelengths. The altitude differences between adjacent density extrema altitudes can be accepted as apparent vertical half-wavelengths. In Table V are given the altitudes of extrema for both vehicles, as well as the corresponding apparent vertical half-wavelengths.

Newton et al. (1969) observed local variations in the terrestrial atmospheric density, confirming that waves propagate in the neutral atmosphere. They in-

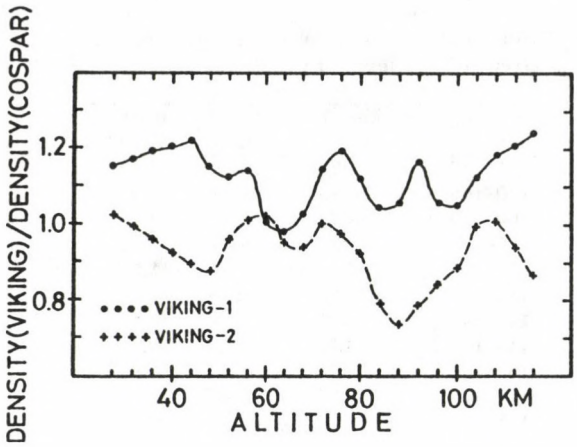

Fig. 2. Ratio of Viking-density to Cospar model-density versus altitude

terpreted these waves as internal gravity waves. The observed apparent vertical half-wavelengths increased with altitude from 1 km at 286 km altitude to 70 km at 510 km altitude, in agreement with theoretical considerations of Hines (1960).

We note that according to the data in Table V our half-wavelengths are significantly longer than those observed by Newton et al. (1969) and the increase of wavelength with altitude is noticable only in the case of Viking-2. It will be interesting to analyse the expected properties of internal gravity waves in the Martian atmosphere.

Conclusions

The fact that the lifetime calculated using the Viking-2 densities is 5.7 times longer than using Viking-1 data exhibits that the density profiles obtained at the two entries differ by an important degree, especially at altitudes above 120 km. Our analysis shows that the differences can be interpreted as internal gravity waves originating at several altitudes with different amplitudes during the two entries, but in addition to this a latitudinal effect can also play a role. It is interesting to remark that during both entries the extrema of density appeared at nearly the same altitudes (See Table V). We consider it as an accidental coincidence.

The lifetimes calculated by using several models differ considerably from one another. Only 1 from 6 models gave an acceptable result. It is evident that these models still need important improvements, because in their actual form they describe Mars' atmosphere, especially the upper-atmosphere, only to a very limited degree.

Acknowledgements

The author thanks I Almár, P Bencze, E Illés-Almár and T Pálvölgyi for their critical remarks and helpful discussions.

This work was supported by OTKA I-1813.

Table V. Local density extrema altitudes and corresponding apparent vertical half-wavelengths with their mean altitudes

									Viking-1							
Extrema alt.	(km)	−44		−50		+56		−64		+76		−86		+92		−100
Apparent $\lambda/2$	(km)		6		6		8		12		10		6		8	
Mean alt.	(km)		47		53		60		70		81		89		96	
									Viking-2							
Extrema alt.	(km)			−48		+58		−66		+74		−88		+106		
Apparent $\lambda/2$	(km)				10		8		8		14		18			
Mean alt.	(km)				53		62		70		81		97			

"+" denotes maximum, "–" denotes minimum

References

Hansen W, Hubbard W, James P, Kieffer T, Leovy C, Martin T, Miner E, Owen T, Pollack J, Seiff A, Stewart A 1978: The COSPAR Mars Reference Atmosphere, J. P. L., Pasadena

Hines C O 1960: *Can. J. Phys.*, 38, 1441–1481.

King-Hele D 1964: Theory of Satellite Orbits in an Atmosphere. Butterworths, London

Moroz V I, Kerzhanovich V V, Krasnopolsky V A 1991: *Kosm. Issl.*, 29, 3–84.

Newton G P, Pelz D T, Volland H 1969: *J. Geoph. Res.*, 74, 183–196.

Sehnal L 1990: *Bull. Astron. Inst. Czechosl.*, 41, 108–115.

Seiff A, Kirk D B 1977: *J. Geoph. Res.*, 82, 4364–4378.

Stewart A I F 1987: Revised Time Dependent Model of the Martian Atmosphere for Use in Orbit Lifetime and Sustenance Studies. JPL PO NQ-802429 Final Report

Acta Geod. Geoph. Hung., Vol. 29(1–2), pp. 149–159 (1994)

IS THERE ANY RELATION BETWEEN THE EARTHQUAKES AND GRAPHITIC CONDUCTORS IN THE UPPER CRUST? – A HYPOTHESIS –

A ÁDÁM[1]

[Manuscript received March 10, 1994]

There is an indication in the area of the Transdanubian electrical conductivity anomaly (CA) that some earthquakes may be generated in the depth range of the conductors. It seems that "low viscosity graphitic formations", parallel to fractures, influence earthquake generation in the crust (tectonic earthquakes) and at the same time attenuate the earthquake waves by their lateral extension and so decrease the seismic hazard. On the basis of these phenomena the graphitic origin of the CA's may be deduced for an area which has not been drilled. This hypothetical relation should be proved by a statistically large enough data set.

Keywords: conductivity anomaly; graphitic formation; low viscosity; Pannonian basin; seismicity; tectonic earthquake

1. Upper crustal conductor in NW Transdanubia (Hungary)

In NW Transdanubia a large conductivity anomaly (TCA) was detected by related telluric and magnetotelluric soundings in the early sixties (Ádám and Verő 1964, Takács 1968, Ádám 1985). The TCA lies between two great tectonic lines, i.e., the Balaton (S) and the Rába line (N) in the eastern part of the Bakony-Drauzug independent geologic unit (BDU, Kázmér and Kovács 1985) where similar narrow conductivity anomalies have been delineated (Ádám et al. 1990).

The TCA consists of 10–15 km wide stripes of high conductivity at two different depth ranges (at 5–7 km and 12–14 km) parallel to the characteristic longitudinal (NE-SW) fractures/strike slips of the Pannonian basin (Fig. 1). The ratio of the extreme resistivity values (Rhomax/Rhomin), i.e., the simple anisotropy is high. (In some cases it is greater than 1000.) Its conductance value ($S = h/\varrho =$ thickness/average resistivity) reaches 10–20 thousand Siemens.

As mentioned above, this anomaly is strongly related to the tectonics of the area. Based on their extreme parameters it has been concluded that it is caused by graphitic schists/black shales which outcrop in the western part of BDU in the Gail Valley Alps (Ádám et al. 1990). The low shear strength (low viscosity) of black shale focuses the tectonic deformation within these units and leads to a

[1]Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences, H-9401 Sopron, POB 5, Hungary

1217-8977/94/$ 4.00

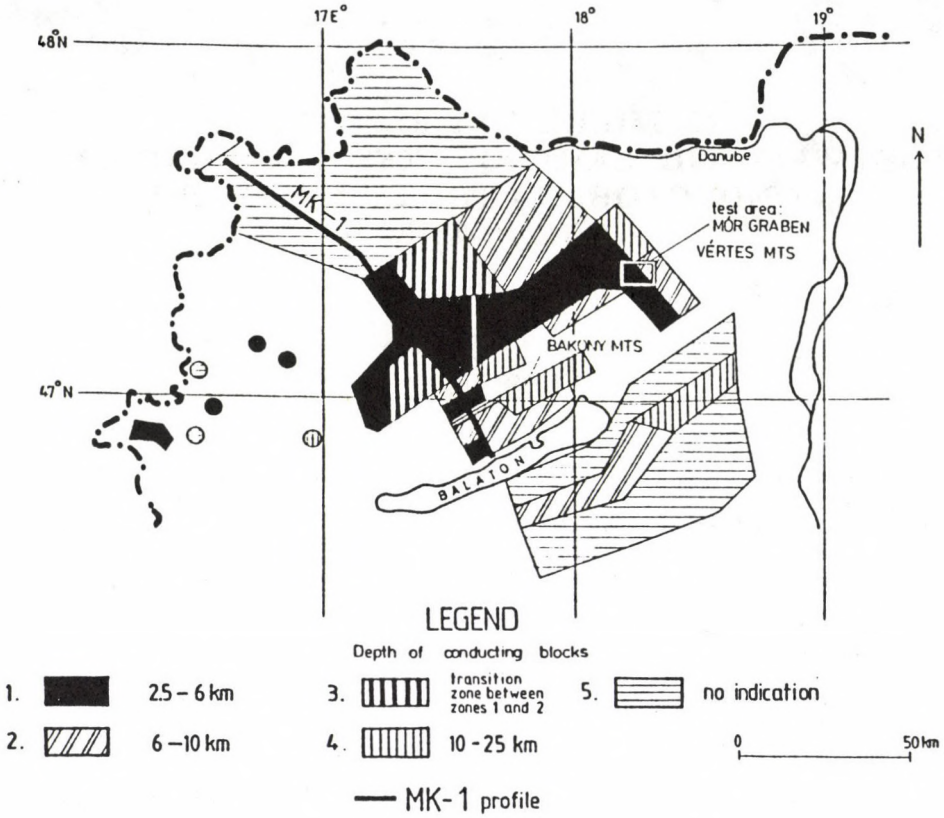

Fig. 1. Map of the occurrence and depth distribution of CA-s in NW Transdanubia (Ádám and Varga 1990)

preferential accumulation of carbon along the shear zones, as stated by the Ensenada EM Induction Workshop in 1990 (Hjelt 1990).

The relation of this graphitic anomaly to the earthquakes of the area has already been studied in the seventies by Ádám (1976). Zsíros and his colleagues' (1983, 1985) very detailed investigation on the seismicity of Hungary and new MT soundings in the area of the Transdanubian CA enabled us to look for closer connection of these phenomena.

2. Connection between depth distribution of the earthquakes and the upper crustal CA in Transdanubia

Zsíros and his co-workers from the Seismological Department of the Geodetic and Geophysical Research Institute published several diagrams on the focal depth distribution of earthquakes (Zsíros 1983, Zsíros and Tóth 1984, Zsíros et al. 1989) determined by Kövesligethy's (1907) formula (or by its simplified version) based on the isoseismals. 73 percent of the seismic events occurred in the depth range

3 to 12 km (Zsíros et al. 1989), i.e., in the depth range of the Transdanubian conductivity anomaly (Fig. 2).[2]

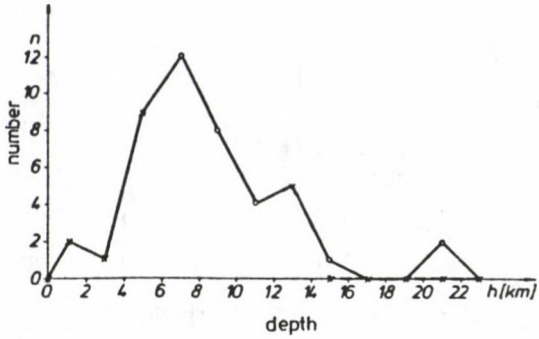

Fig. 2. Frequency of the depth values of CA-s in NW Transdanubia (Ádám 1981)

It would be interesting to compare the relation of the depths of these quite different physical phenomena in well determined earthquake epicenters lying in the main stripes of the Transdanubian CA of different depths clearly indicated by Wiese induction arrows, too (Fig. 3) in the area of the TCA.

Therefore isoseismal maps of two earthquakes with well determined foci have been selected: one has been generated near the shallower conducting stripe, the other one near the deeper one of the TCA (See Table I).

Table I.

Site	Time	Focal depth [km]	Magnitude (M)	Epicentral intensity $I_{o\ epi}$
Ukk-Türje	Sept. 13, 1953	5	4.2	6.5
Berhida	Aug. 15, 1985	~ 10* 5 – 9.5**	4.7	6.5

*Zsíros et al. (1989)
**Bondár (1994)

The isoseismals are shown for these earthquakes in Figs 4a and 4b. The agreement between the depth values of the two different phenomena, earthquake and CA, seems to be good, i.e., the earthquakes originated from about the depth of the conductors. Thus the graphitic schist blocks of low viscosity — similarly to the preformed and weak fault and fractured zones (see viscosity and velocity behaviour in the San Andreas fault zone in Fig. 5 after Meissner 1986) — can influence the generation of the earthquakes. (These are the tectonic earthquakes.)

[2]It is worth noting that the lower crustal conductor in the Pannonian basin (Hungary) lies below these depths, at the depth of 18 ± 5.3 km as determined by Ádám et al. (1989), or somewhat shallower (Ádám et al. 1990), i.e., in the depth range without earthquakes, in the ductile zone corresponding to the high heat flow values in the Pannonian Basin ($\geq$ 80 mWm^{-2}).

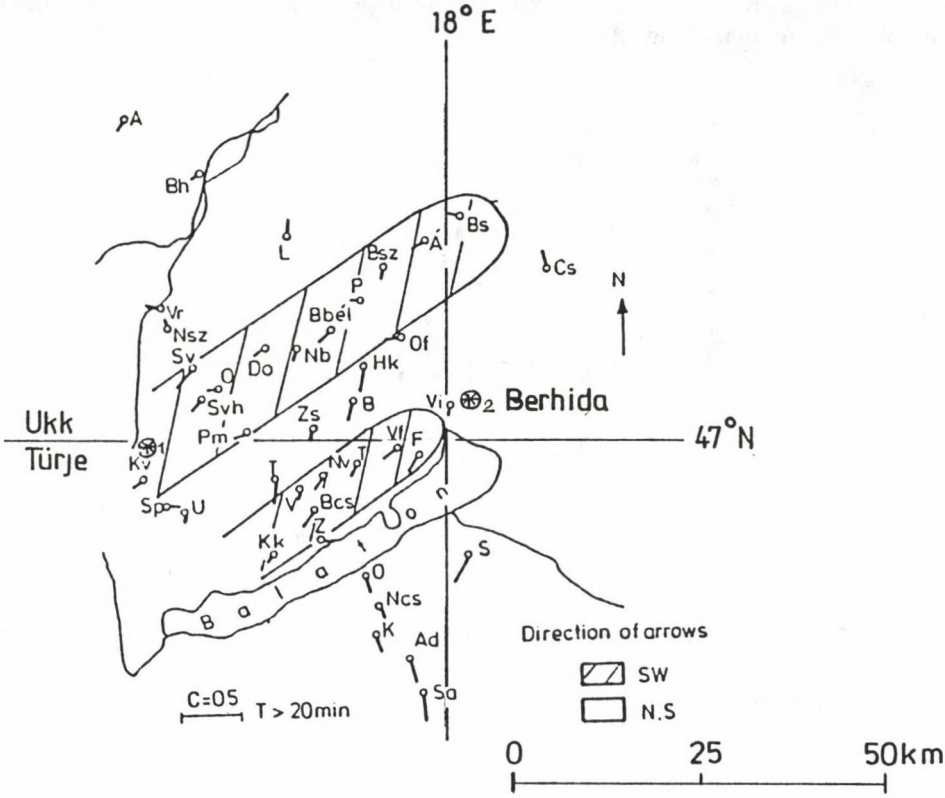

Fig. 3. Wiese arrows in NW Transdanubia (Wallner 1977), $\otimes_1$ Ukk-Türje, $\otimes_2$ Berhida earthquake

We are aware of the weakness of the justification of this relation, nevertheless, we wanted to point out its possibility until we have more, and more significant data from both areas to answer many questions arising in this respect.

3. The attenuation of the seismic waves by graphitic formations in the upper crust

The graphitic blocks (CA) attenuate the seismic waves generated by the earthquakes due to low viscosity (η) as the two parameters are in relation. Meissner (1986) has given the following very simplified formula

$$\ln \eta = 4.4 \ln Q + 22$$

where Q = quality factor, its reciprocal value (Q^{-1}) = attenuation. This does not take into account the dependence on frequency, on the rheological model, *etc.*, nevertheless, it qualitatively calls attention to the existing relation.

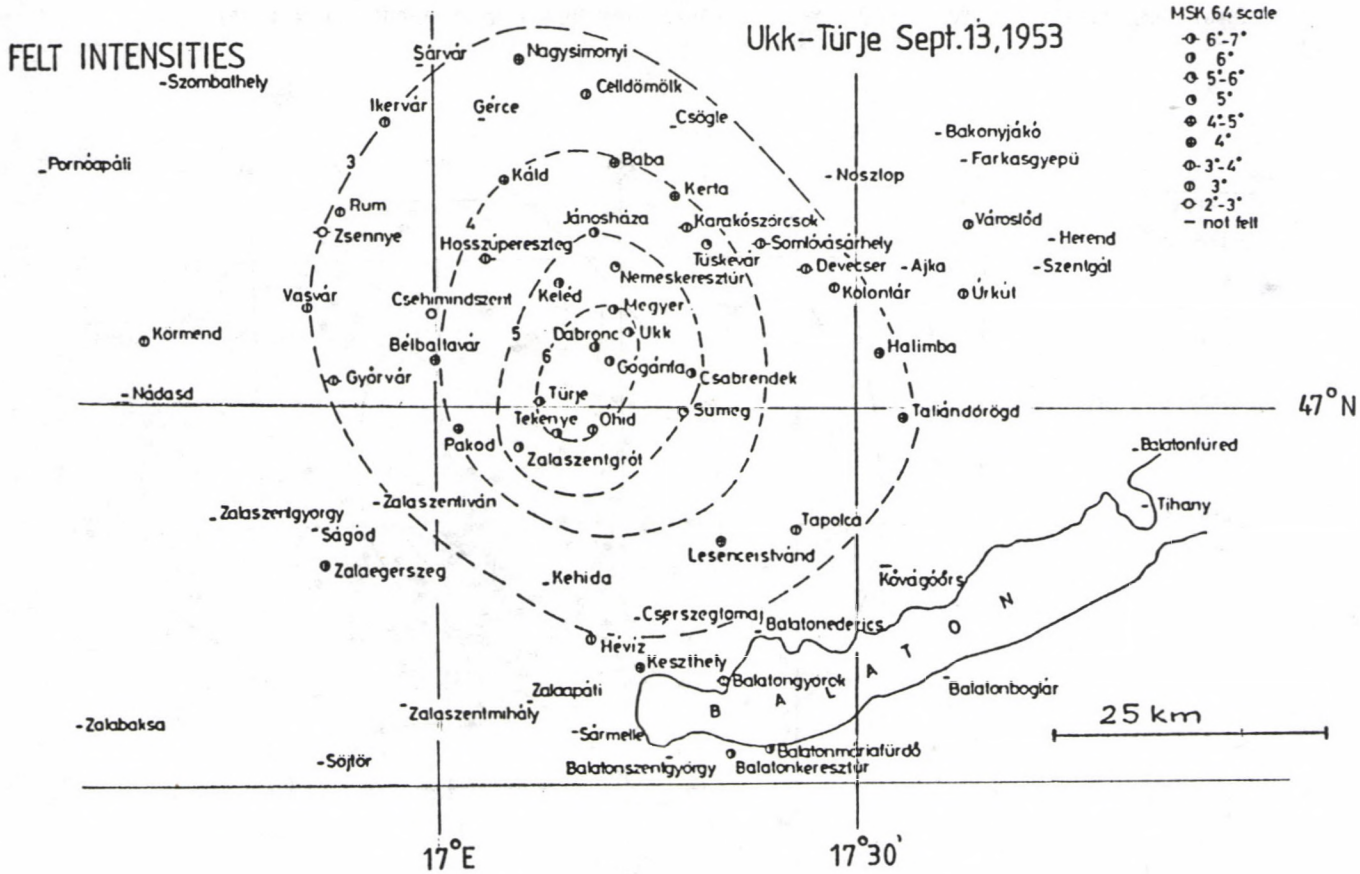

Fig. 4. a. Distribution of the felt intensity of the Ukk-Türje earthquake on September 13, 1953 (Zsíros 1982).

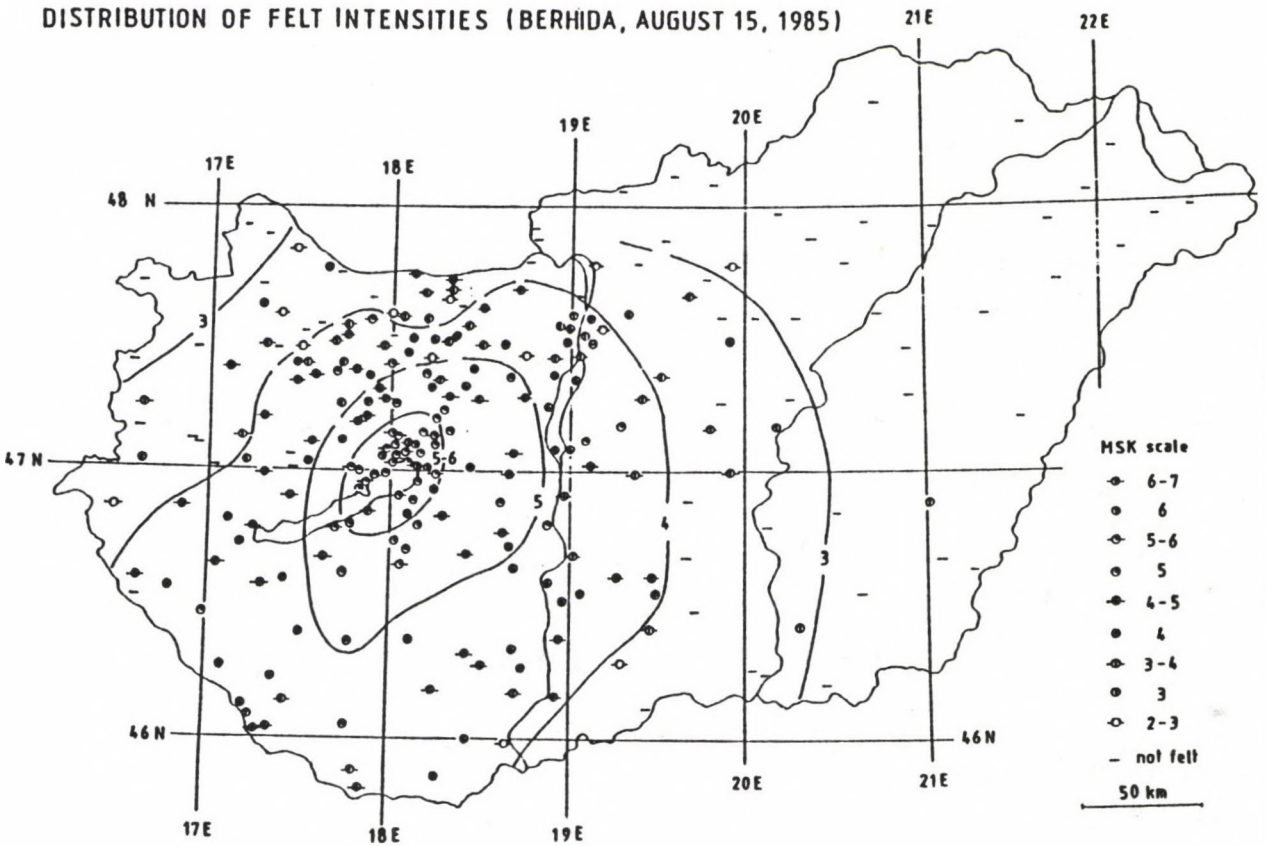

Fig. 4. b. Distribution of the felt intensity of the Berhida earthquake on August 15, 1985 (Zsíros 1989)

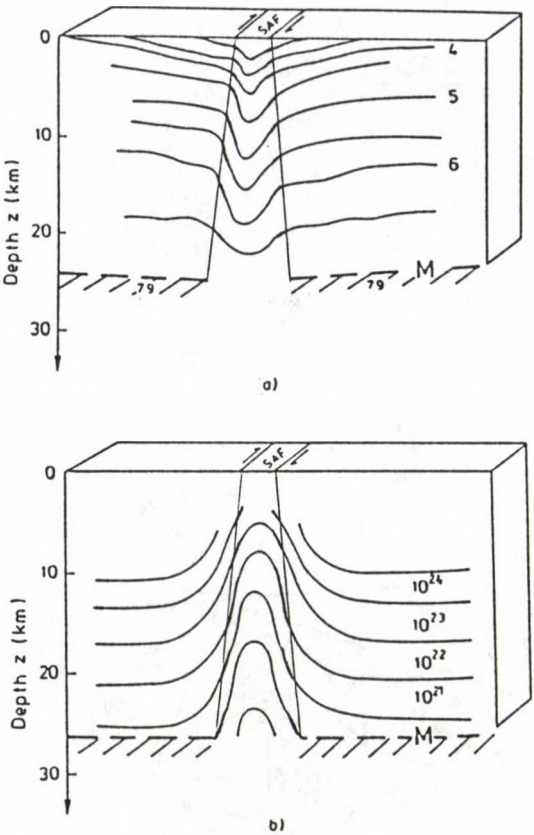

Fig. 5. Velocity and viscosity behaviour in a fault zone area. (a) Velocity isolines around the San Andreas fault, based on seismic reflection and refraction data from Feng and McEvilly (1983) and (b) viscosity isolines taken from Meissner (1986)

Earthquake intensity decreases exponentially with the epicentral distance (independently from the direction):

$$I_k = I_o \exp(-\alpha_k R_k)$$

where R_k denotes the mean distance [km] of the k-th isoseism from the epicentre, α_k the attenuation coefficient of intensity and I_k the intensity of the k-th isoseism. Zsíros (1985) estimated the attenuation coefficient (α_k) for 5 source areas being inside and outside the Pannonian basin (Hungary) (Fig. 6). The arithmetic mean was chosen as representative value ($\bar{\alpha}_k$). Studying the spatial distribution of the attenuation coefficients Zsíros delineated subregions in the greatest source area V including almost the whole area of Hungary. Table II gives the mean territorial attenuation coefficients ($\bar{\alpha}$) of source areas with standard deviations ($\sigma_{\bar{\alpha}}$) and data number (N). All data used in the calculations are shown in Fig. 6.

The greatest attenuation coefficients appear in subarea V/2, i.e., just in the

Table II.

Source area	$\bar{\alpha}$ [km^{-1}]	$\sigma_{\bar{\alpha}}$	N
I	0.012	0.004	20
II	0.012	0.004	19
III	0.016	0.005	6
IV	0.016	0.013	12
V/1	0.023	0.005	5
V/2	0.042	0.029	13
V/3	0.018	0.006	8
V/4	0.014	0.004	5
V/5	0.029	0.011	7
V/6	0.010	0.003	6
V/7	0.016	0.006	5

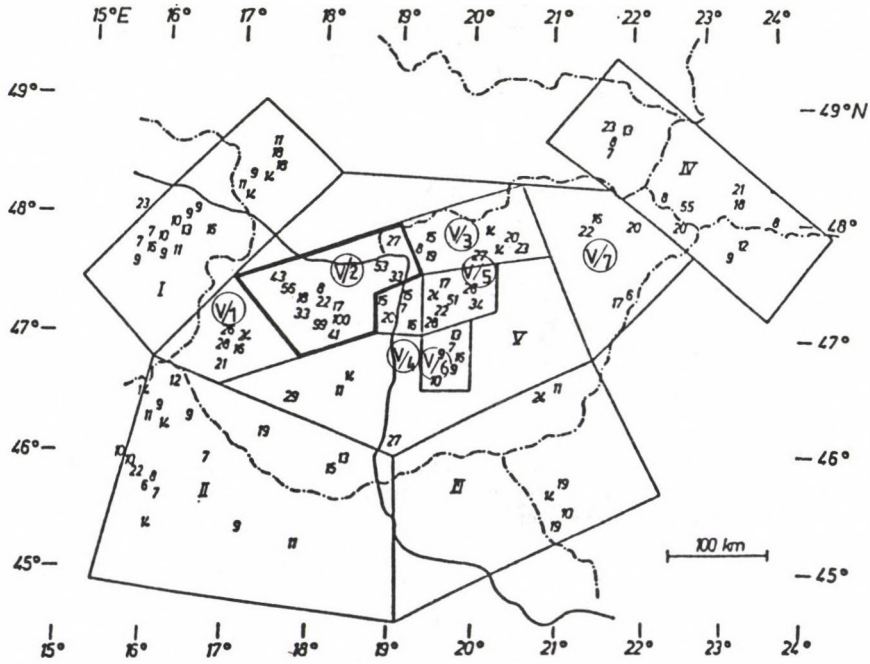

Fig. 6. Attenuation coefficients $\bar{\alpha}$ [10^{-3} km^{-1}] of intensities in the source areas (Zsíros 1985)

area of the Transdanubian CA attributed to highly conducting graphitic blocks. The average is 0.042 km^{-1}, for comparison: the lowest α_k value in Hungary is 0.010 km^{-1}. It is necessary to note that the scatter (and the error, too) of α_k values is the highest in V/2 area because among the greatest values appear also some low values certainly due to the great inhomogeneity of the CA, probably its block structure (Ádám and Varga 1990).

The following two phenomena in the area of the Transdanubian CA are consequences of this high attenuation:

- The isoseismals of the Berhida earthquake are strongly asymmetric (Fig. 4b); the distance between the isoseismals is much smaller in the area of the CA than south of the Lake Balaton, where the attenuation coefficient (α_k) is much smaller.

- The map of maximum felt intensity in Hungary caused by earthquakes generated outside the border, i.e., in the source areas I–IV in Fig. 6 has been constructed by Zsíros and Tóth (1988). As these authors remarked "a large part of Transdanubia is exposed to a considerable hazard. Damaging earthquakes from earlier Yugoslavia (i.e. from south) have effects up to Lake Balaton" (i.e. just to the southern border of the Transdanubian CA). "In the western part of Hungary, Austrian earthquakes denote the main threat but their intensities rapidly decrease with distance from the sources" (especially at the western border of the Transdanubian CA, at the Rába line) (Fig. 7).

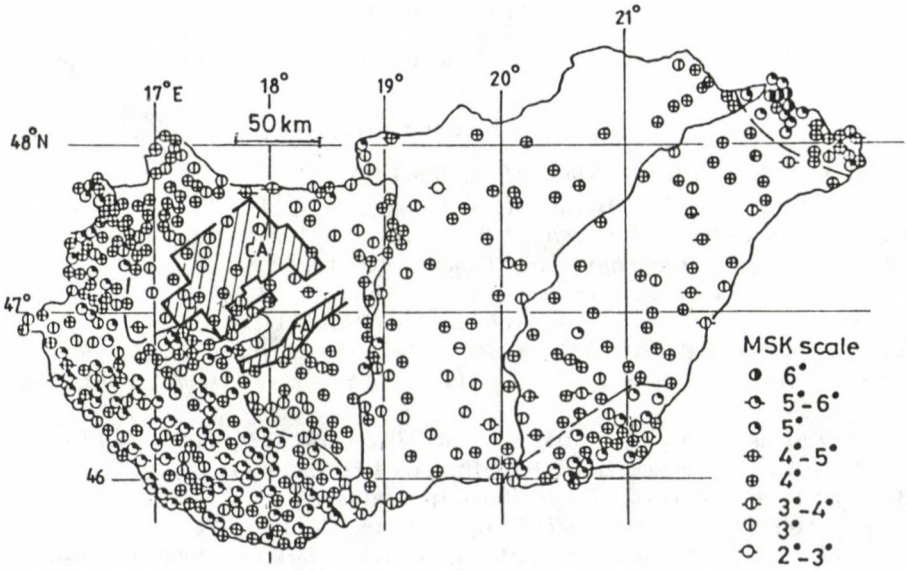

Fig. 7. The largest felt intensity values in Hungary generated by the effects of foreign earthquakes (Zsíros and Tóth 1988)

4. General conclusions from the relation between earthquakes and CA of graphitic origin

As has been shown in case of the Pannonian basin, especially its NW part (Transdanubia), graphitic or C (?) film CA-s may have effects on the depth distribution of earthquakes and on the propagation direction of seismic waves in the brittle upper crust. These effects can be summarized as follows:

- a common depth range of conductivity anomalies and of earthquakes can hint at the graphitic origin of the CA, as such graphitic formations, parallel to fractures and faults (preformed zones) can influence earthquakes with their low viscosities (see Fig. 5)
- earthquakes are more highly attenuated in low viscosity graphitic blocks, therefore their effect extends to much smaller areas
- earthquake mechanism starts at a lower level of stress accumulation, therefore low viscosity graphitic bodies can set an upper limit to earthquake magnitudes.

Such as water in the fracture zone, could also contribute to the attenuation. It is suggested that more study of such attenuation phenomena be initiated, particularly important is a study of the role of graphite.

Acknowledgement

The author is indebted to Prof. Al Duba for the critical review of this paper.

References

Ádám A 1976: *Acta Geod. Geoph. Mont. Hung.*, 11, 277–285.
Ádám A 1981: *Acta Geod. Geoph. Mont. Hung.*, 16, 97–113.
Ádám A 1985: *Acta Geod. Geoph. Mont. Hung.*, 20, 173–182.
Ádám A, Varga G 1990: *Phys. Earth Planet. Inter.*, 60, 80–88.
Ádám A, Verő J 1964: *Acta Techn. Hung.*, 47, 63–77.
Ádám A, Landy K, Nagy Z 1989: *Tectonophysics*, 164, 361–368.
Ádám A, Duma G, Horváth J 1990a: *Phys. Earth Planet. Inter.*, 60, 155–162.
Ádám A, Nagy Z, Nemesi L, Varga G 1990b: *Acta Geod. Geoph. Mont. Hung.*, 25, 291–307.
Ádám O, Haas J, Nemesi L, Tátrai M, Ráner G, Varga G 1984: In: Annual Report of the Eötvös L. Geophysical Institute of Hungary for 1983, 37–44.
Bondár I 1994: *Acta Geod. Geoph. Hung.* (present issue)
Feng R, McEvilly T V 1983: *Bull. Seism. Soc. Am.*, 73, 1701–1720.
Hjelt S E 1990: In: Report on the 10th EM Induction Workshop, 1990. Ensenada, Mexico
Horváth F, Ádám A, Stanley W D 1987: *Rendiconti. Soc. Geol. Italiana*, 9, 123–130.
Kázmér M, Kovács S 1985: *Acta Geod. Hung.*, 20, 71–84.
Kövesligethy R 1907: *Gerlands Beitr. Geophys., Leipzig*, 98, 146–154.
Meissner R 1986: The Continental Crust. A Geophysical Approach. International Geophysics Series. Vol. 34. Academic Press, INC
Takács E 1968: *Acta Geod. Geoph. Mont. Hung.*, 3, 155–160.
Wallner Á 1977: *Acta Geod. Geoph. Mont. Hung.*, 12, 145–150.
Zsíros T 1982: Earthquake in Ukk and Türje (Manuscript in Hungarian)
Zsíros T 1985: *Gerlands Beitr. Geophys., Leipzig*, 94, 111–122.
Zsíros T 1983: *Acta Geod. Geoph. Mont. Hung.*, 18, 119–128.
Zsíros T 1989: Macroseismic observations in Hungary (1983–1988). Seismological Observatory of the Geodetic and Geophys. Res. Ints. of the Hungarian Academy of Sci., Budapest, pp. 67.

Zsíros T, Tóth L 1984: *Acta Geod. Geoph. Mont. Hung.*, 19, 367–381.
Zsíros T, Tóth L 1988: *Acta Geod. Geoph. Mont. Hung.*, 23, 411–415.
Zsíros T, Mónus P, Tóth L 1989: *Gerlands Beitr. Geophys., Leipzig*, 98, 146–154.

Acta Geod. Geoph. Hung., Vol. 29(1–2), pp. 161–177 (1994)

INSTRUMENTS FOR PRECISE DETERMINATION OF HORIZONTAL DEFORMATIONS IN THE PANNONIAN BASIN

GY MENTES[1]

[Manuscript received March 10, 1994]

In this paper the instruments developed for earth tide and deformation measurements at the Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences are described. A long quartz tube extensometer was installed at the Sopron Observatory in 1990 and an another one at the Uran Mining Company in Pécs in 1991. For a higher reliability of the measurements and a better interpretation of the obtained data three short (1m) quartz tube extensometers have been developed and installed at Sopron Observatory for parallel recording with the existing long one. More electronic sensors and their mechanical units were developed to another instruments used at the observatory in Budapest and in the Czech Republic. For better understanding of the effect of the air pressure a microbarograph was developed. In the paper the main aspects of the deformation measurements are given, too.

Keywords: crapoudine; Earth tides; extensometer; magnetostrictive calibration; microbarograph

1. Introduction

The local and global tectonic movements are generally calculated from the residual curves of Earth tide records. For high precision measurement of horizontal displacements extensometers are used. To detect the very slowly varying tectonic phenomena extensometers of very high resolution ($10^{-9} - 10^{-11}$) and stability are needed. Besides of the mechanical stability of the extensometers the requirements against the electronic sensors and amplifiers are very strict because they can cause a very high electrical drift that remains in the residual curve after removing the Earth tide components. To avoid the interpretation of the drift as a horizontal motion at long term measurements, the following conditions should be kept:

1. To record on very stable places, in observatories possibly built into or on the bedrock.
2. To reduce the cavity effect as much as possible.
3. A perfectly stable and rigid attachment of the extensometers to the bedrock.
4. Application of high accuracy sensors with low drift.

[1]Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences, H-9401 Sopron, POB 5, Hungary

1217-8977/94/$ 4.00

5. To use extensometers of very high mechanical stability.

6. To ensure stable environmental conditions at the recording place.

7. Precise measurement of the changes of the environmental parameters (temperature, atmospheric pressure, humidity *etc.*) and taking them into account at the evaluation of the measuring results.

8. High precision calibration of the instruments.

9. Parallel recording by more instruments of different and identical types.

By ensuring the conditions mentioned in the points 1 to 7 the drift of the instruments can be held low and the measured data can be corrected by the measured environmental parameters. The remaining drift can only be diminished by using more instruments for the measurements. The generally used long extensometers (10–30 m) are very expensive and need a large, long underground room for installation, therefore it is a problem to use more parallel instruments.

One of our aims is to develop better sensors and amplifiers which allow to make short (1–2 m) extensometers of the same or higher accuracy than the long ones. The short extensometers enable not only a parallel recording but also a three dimensional measurement in a small recording room.

The main goal of this paper is to summarize our efforts in developing the instruments for the investigation of the movements in the Pannonian Basin. The paper describes the construction of the long and short extensometers, the microbarograph developed at the Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences and deals with the calibration of these instruments.

2. Capacitive transducers for extensometers

Capacitive transducers for horizontal pendulums and gravimeters were developed in the early 70's (Mentes 1981, 1983). The electronics of these sensors was further developed and applied to the extensometers. The principle of the electronics is shown in Fig. 1. The transducer consists of a differential condenser connected into a bride circuit. An oscillator of very stable amplitude supplies the capacitive bridge. The supply voltage of the bridge is peak to peak 25 V. The oscillator has

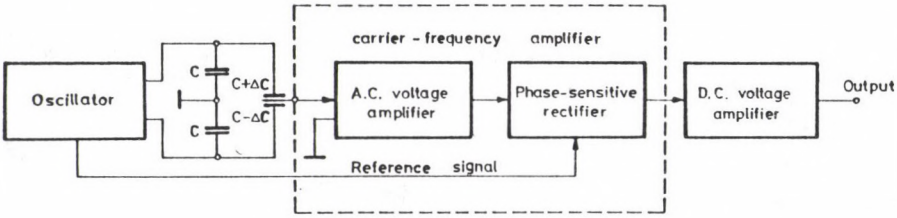

Fig. 1. The principle of the extensometer electronics

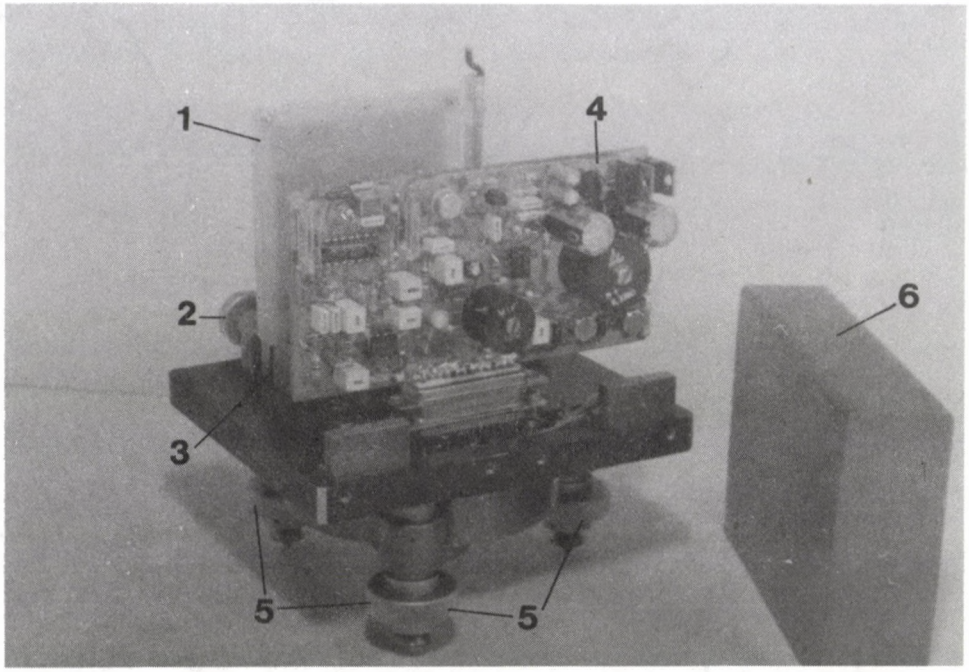

Fig. 2. The sensor unit developed to long extensometers. differential condenser (1), fine adjusting screw of the standing plates of the differential condenser (2), sliding clamp of the standing plates (3), board of electronics (4), foot screws (5), shielding bucket of the electronics (6)

a frequency of 12 kHz. Because the output voltage of the bridge is directly proportional to the amplitude of the supply voltage of the bridge, the amplitude of the output voltage of the oscillator is stabilized by a PID regulating circuit and the oscillator is temperature compensated. In this way an amplitude stability better than 10^{-4} was achieved. The output voltage of the capacitive bridge is amplified by a carrier-frequency amplifier. It consists of two stages. The first one has a high impedance and a high gain. The second stage is the phase-sensitive rectifier which ensures that the sign of the output voltage be in accordance with the direction of the movement of the moving plate of the differential condenser. The carrier-frequency amplifier ensures also a very high noise rejection, because it amplifies only the signal with the same frequency as its reference signal has. The whole electronics is very carefully shielded and placed near to the differential condenser to eliminate the effect of spurious capacitances.

To compensate the drift of the extensometers and to adjust the output voltage of the sensor amplifier to zero, a very fine mechanically adjustable support was developed for the precise movement of the standing plates of the differential condenser (Fig. 2). The mechanical unit is standing on three foot screws of fine thread (5) placed at the corners of a right-angled triangle. The standing plates of the differen-

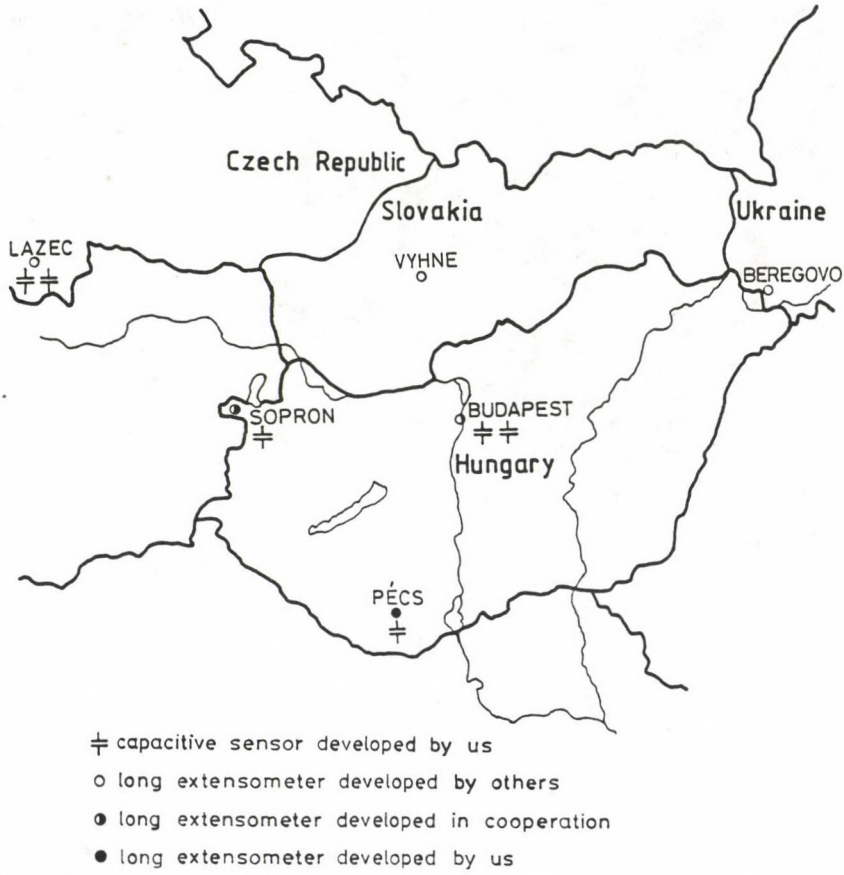

Fig. 3. Extensometer network for displacement measurements in the Pannonian Basin

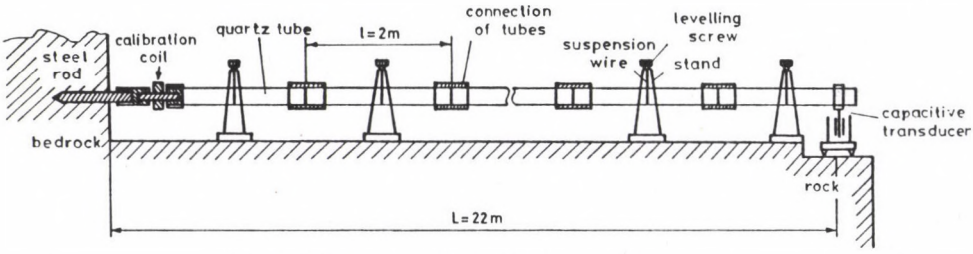

Fig. 4. The construction of the long extensometers

tial condenser (1) can be very finely moved by the fine adjusting screw (2) coupled via a lever transmission to their sliding clamp (3). The electronics (4) is shielded by the shielding bucket (6).

This type of sensor unit is used in two extensometers at the Budapest Geodynamical Observatory, in two extensometers in Lazec (Czech Republic), in one

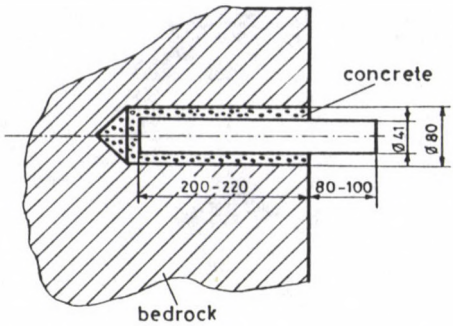

Fig. 5. Connection of the extensometer to the rock

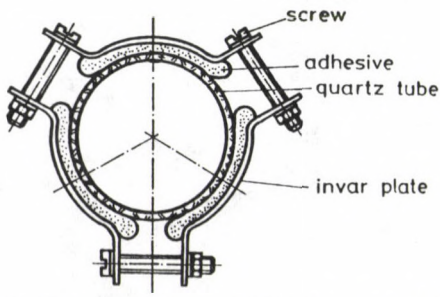

Fig. 6. Connection of quartz tubes by invar profiles

extensometer at the Sopron Geodynamical Observatory and at the Mecsek Uran Mining Company in Pécs (Fig. 3). On the one hand this figure shows the observatories taking part in the measurements of the movements of the Pannonian Basin, on the other hand it shows the observatories where extensometers and sensor units developed by us are operating.

3. Construction of long extensometers

We have constructed two extensometers which differ from each other in the calibration units and in the connection of the quartz tubes. The construction of the long extensometers developed by us is shown in Fig. 4. The extensometer is attached to the rock by means of a stainless steel bolt and concrete as shown in Fig. 5. The stability of this connection is very important. The calibration device is connected between the bolt and the quartz tube. The length of both extensometers is about 22 m and therefore they consist of several quartz tubes of 2–2.5 m length attached to each other. The setting strength of the attachment determines the stability of an extensometer. The extensometer in Sopron was made in cooperation with the O Yu Smidt Institute of Earth Physics, Moscow. At this instrument the attachment of tubes was made by means of invar profiles and adhesive consisting of cement, quartz sand and a two-component resin to accelerate the bonding of the adhesive (Fig. 6). The quartz tubes of the other extensometer installed at

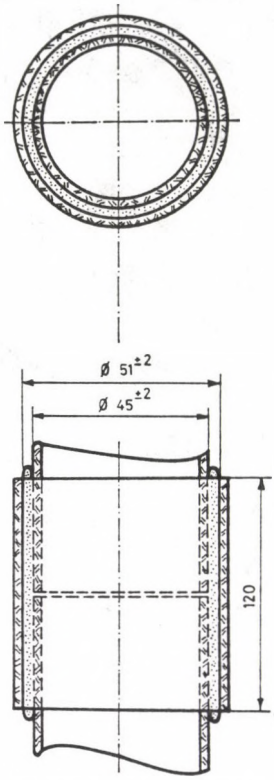

Fig. 7. Connection of quartz tubes by quartz muffles

the Mecsek Uran Mining Company in Pécs are attached by means of muffles made of quartz tubes which have a greater diameter (46–49 mm) than the long tubes (43–45 mm) to be attached (Fig. 7). The space between the tube and the muffle is luted by a two-component resin. The measurements showed that in both cases the junctions are very strong, they have about the same thermal expansion as the quartz tube and do not cause observable drift of the instruments. The attached tubes are suspended on very fine wires connected to supports (Fig. 8) making possible the levelling of the quartz tube (Fig. 9). The longer are the suspension wires, the smaller is the force obstructing the movements of the suspended quartz tubes. According to our measurements this force is negligible.

The mechanical unit of the capacitive sensor is placed at the free end of the extensometer and it is attached to the rock by concrete. The moving plate of the differential condenser is fixed to the end of the extensometer thus the movements of the rock at the fixed end of the extensometer are transferred by means of the rigid quartz tube to the free end of the extensometer and measured with respect to the standing plates of the capacitive transducer (Fig. 10).

For study of local tectonic effects extensometric data with a relative precision of $10^{-9} - 10^{-11}$ have to be used. To assure this accuracy very stable thermal conditions

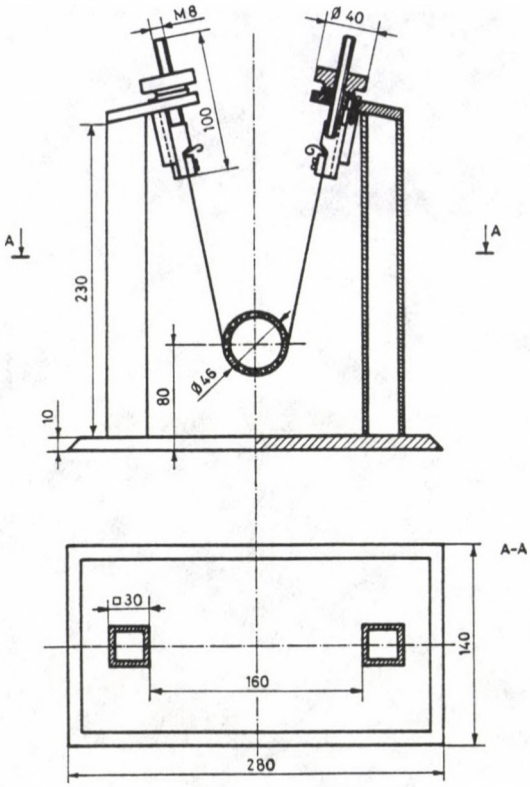

Fig. 8. Support of the quartz tubes

are needed at the recording places. The temperature variations should be lower than 0.1 – 0.01°C. To achieve this very low temperature variation the recording rooms of the extensometers are separated from the other parts of the observatories by air lock doors and in addition both extensometers are directly covered from side and above by means of poliethylen insulating blocks. The extensometer in the uranium mine is operating at a temperature of 48°C and the one at the Sopron Observatory at 9°C.

4. Calibration of the long extensometers

4.1 Magnetostrictive calibration device

The extensometer installed at the Sopron Geodynamical Observatory is calibrated by means of a magnetostrictive calibration device developed at the O Yu Smidt Institute of Earth Physics in Moscow. This device consists of a permendur-core and a coil (Fig. 11). The coil is supplied by direct-current and the core changes his length in function of the magnitude of the current. The free end of the extensometer moves to the same measure as the length of the magnetostrictive core changes. The extensometer and therefore the magnetostrictive calibration device was calibrated

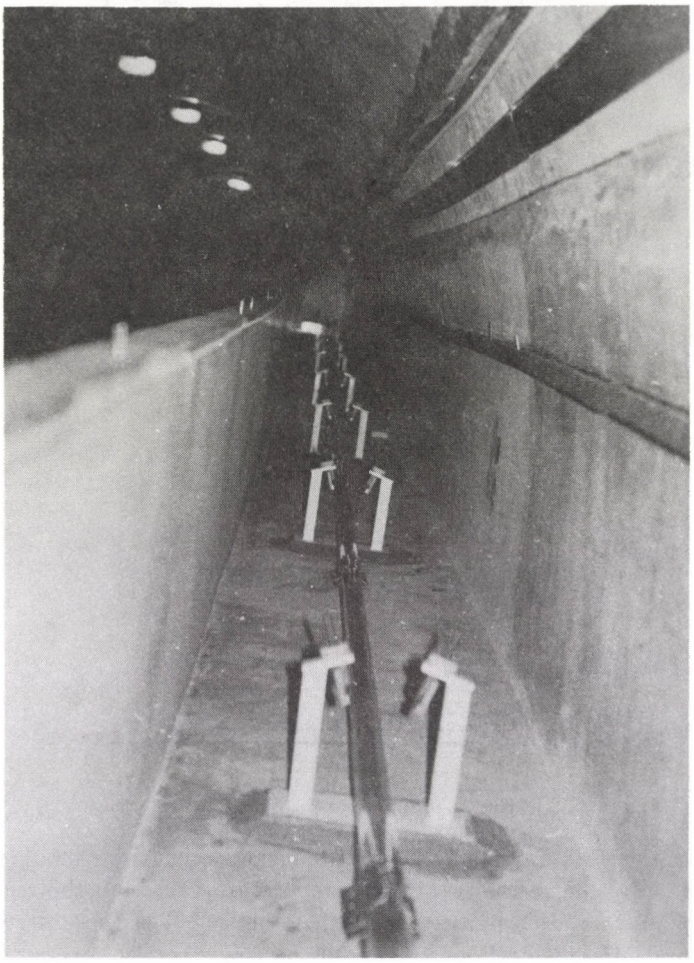

Fig. 9. The extensometer installed at the Sopron Geodynamical Observatory

at the installation by means of the Russian optical calibration device connected to the free end of the extensometer. From the measurements the following main parameters of the extensometer were obtained (Mentes 1991):

- calibration coefficients:
  optical coefficient: 2.467 nm/mA
  electrical coefficient: 1.447 mV/mA

- electrical sensitivity: 0.5865 mV/nm.

The accuracy of the above given parameters is about 2 %. The optical calibration unit can be used for yearly controls of these parameters but it cannot be done without disturbing the work of the extensometer. The magnetostrictive calibration

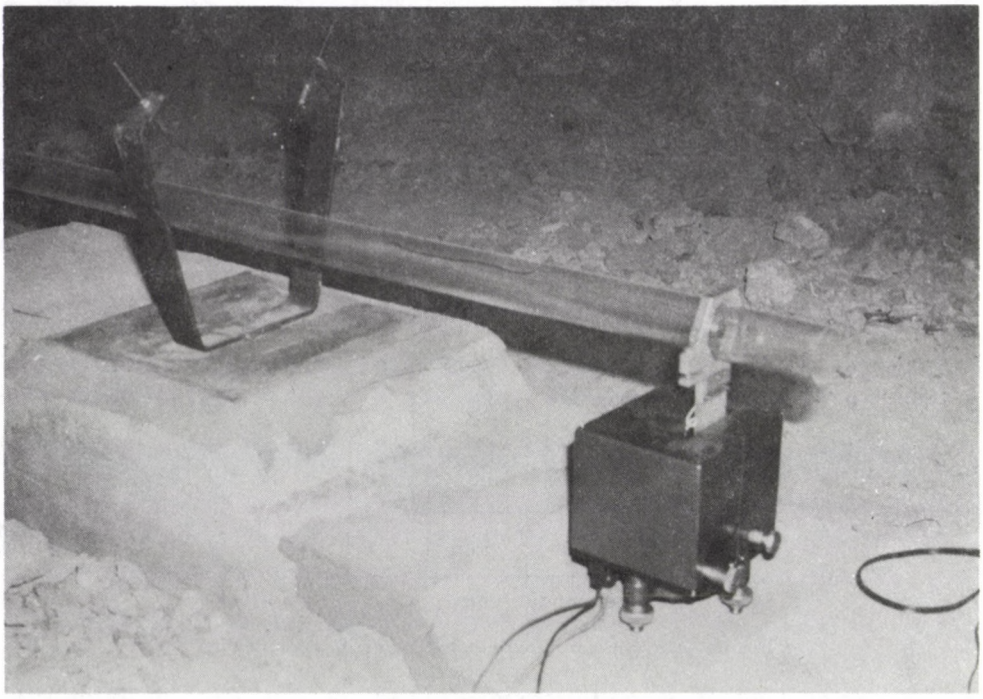

Fig. 10. The capacitive sensor unit connected to the end of the extensometer installed at the Mecsek Uran Mining Company in Pécs

device is used for the daily calibration of the instrument. An automatic controlling device gives a constant current of 150 mA to the coil and causes a displacement of 370 nm at the free end of the extensometer. This value has been strictly constant during the three years long recording since the installation.

4.2 The crapoudine

For the extensometer installed at the Mecsek Uran Minig Company in Pécs we did not have the possibility to make a magnetostrictive calibration device due to the lack of magnetostrictive material (permendur). Therefore we have developed a crapoudine for the calibration of extensometers. Earlier we used these devices for the calibration of horizontal pendulums. The crapoudine is a thick-walled (about 5 mm) closed membrane made of stainless steel. The interior of the membrane is jointed via a pipe junction and a flexible pressure-tight tube to a mercury vessel filled with mercury. By lifting the mercury vessel the pressure increases inside the membrane and its ends protrude with some nm/Hgcm. The construction of the crapoudine is shown in Fig. 12. The joint ends of the crapoudine are identical with the ones of the magnetostrictive device, so the same clamping frame (Fig. 13) can be used to both devices to connect them to the quartz tube. The calibration devices have the same measurements, hence they are interchangeable without any modification of the

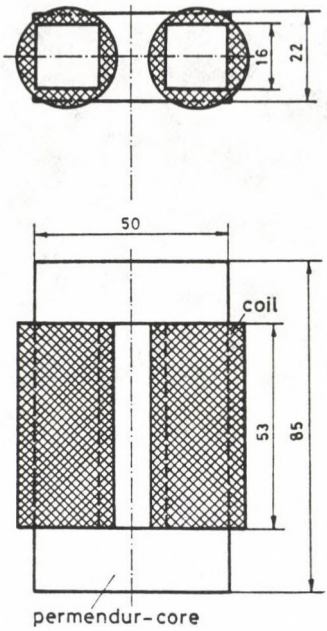

Fig. 11. The Russian magnetostrictive calibration device

extensometers. Figure 14 shows the crapoudine built into the extensometer directly at the fixed end of the instrument. The one side of the clamping frame is attached to the bolt concreted into the rock and the other end of the frame to the quartz tube. This instrument is calibrated once a week. The automatic lifting mechanism (Fig. 15) lifts the mercury vessel by 30 cm very slowly (about 20 minutes) then holds the vessel in this elevated position for 30 minutes and lowers then the vessel back to its original position. The dilatation of the crapoudine due to the pressure change of 30 Hgcm is exactly 400 nm. The crapoudine was calibrated by means of the high precision calibration device developed in our Institute (Mentes 1993). This instrument is used for the measurement of very small displacements with an accuracy of 1 nm.

According to our measurements the linearity of crapoudines is better than the one of the magnetostrictive coils. Generally the linearity errors of the crapoudines are about 1 % and the ones for the magnetostrictive coils are about 2.5 %.

5. Short extensometers

Our short extensometer is made of a 1 m long single quartz tube having a diameter of 10 mm and a wall thickness of 1 mm. The advantage of this solution is that the instabilities of the instrument rised from the connections of the tubes are omitted. The disadvantage is that the instrument has a far lower sensitivity as the long ones which must be counterbalanced by a higher gain of the electronic amplifiers. To achieve a higher gain with an acceptable electronic noise level a new

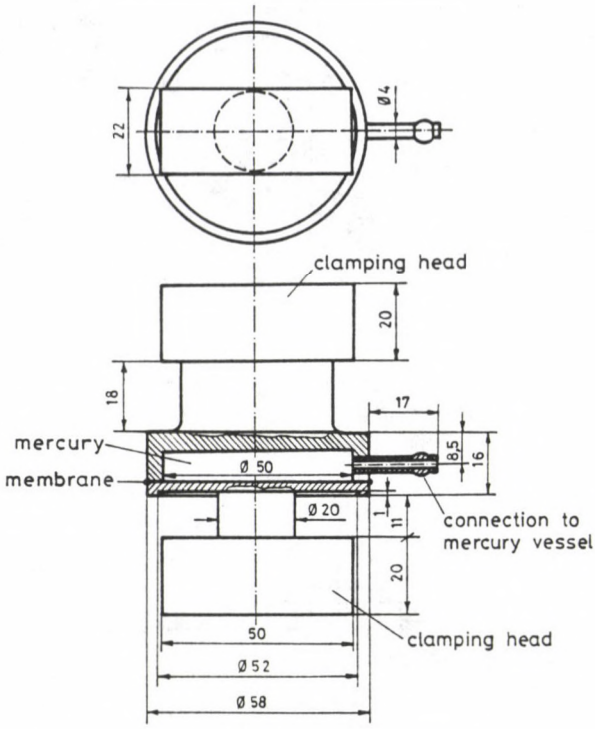

Fig. 12. The construction of the crapoudine

sensor unit was developed. The principle of the electronic solution is the same as at the long extensometers, but the amplifier is built up from new electronic parts according to the latest development stage of electronics.

Figure 16 shows the construction of our short extensometer. The one end of the quartz tube (1) is strengthened by means of a very robust holding fixture (2) to the bolt concreted into the bedrock. On the other end of the tube is a fastening clamp holding the moving plate (3) of the differential condenser. This plate hangs down into the interspace between the standing plates (4) of the capacitive transducer placed on the mechanical unit (5) which is fastened to an another bolt (6) concreted into the bedrock (7).

Figure 17 shows the complete sensor unit. The fastening clamp (8) holds the rigid base plate (7) which holds the whole electronics (9) and the precise moving mechanism of the standing plates (4) of the differential condenser. The moving mechanism consists of the slide (6) holding the standing plates (4). A very fine motion of the slide can be produced by the micrometer screw (10) via a single-armed lever (5).

The calibration of the long extensometer can be done by magnetostrictive devices. A proper solution is being developed in our Institute. We should like to make magnetostrictive calibration devices of smaller dimensions which have about the same conversion coefficient as the devices made for long extensometers.

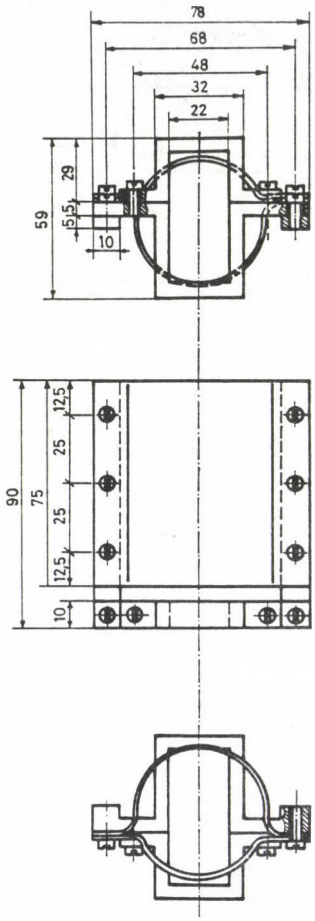

Fig. 13. The jointing part to connect the magnetostrictive device or the crapoudine to the quartz tube

The short extensometers make three-dimensional measurements possible and with more instruments simultaneously parallel measurements in small recording rooms. Of course three-dimensional parallel measurements are possible, too. In small recording rooms the steadiness of the environmental parameters (temperature, humidity, *etc.*) can be easier assured. Especially the constant temperature is very important due to the lower sensitivity of the short extensometers. The temperature must be kept within 0.1°C and the annual variation must be within 0.5°C. The only disadvantage of the small recording rooms is the higher cavity effect.

6. The microbarograph

The effect of the atmospheric pressure changes influences the extensometric measurements and in general all those which have a connection to the gravity variations

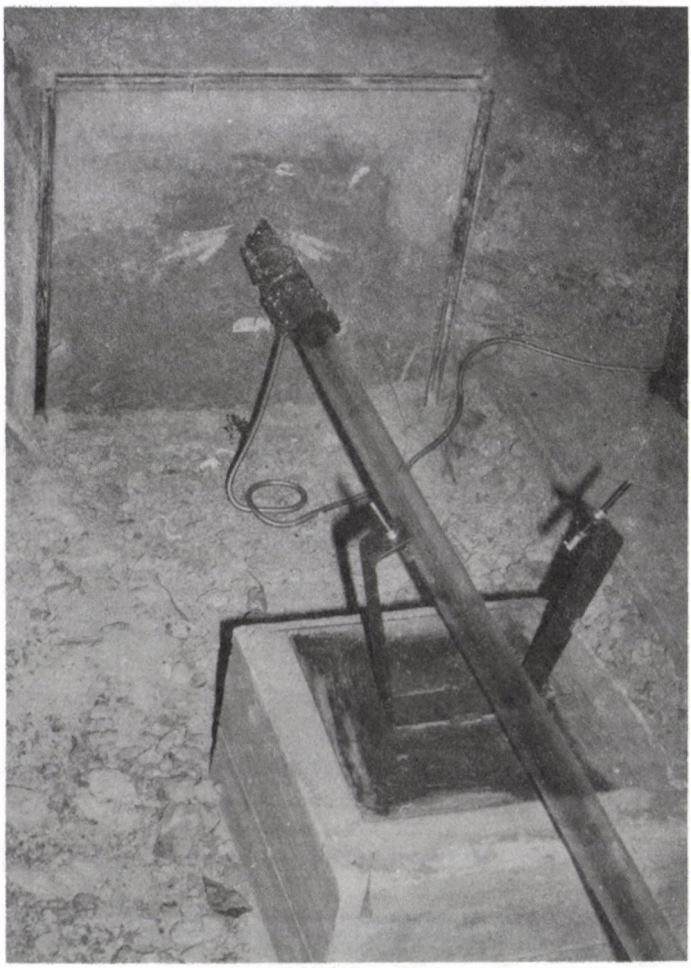

Fig. 14. The crapoudine built into the extensometer at the Uran Mining Company in Pécs

of the Earth. The mass of the atmosphere produces an effect on the gravity in two ways:

- direct effect: air mass attraction

- indirect effect: an elastic deformation of the solid Earth due to an air mass loading.

The sum of these effects results a magnitude of the atmospheric gravity effects of about 20 μGal and therefore — similarly to the gravity field variations — the atmospheric pressure influences the extensometric measurements. Consequently the

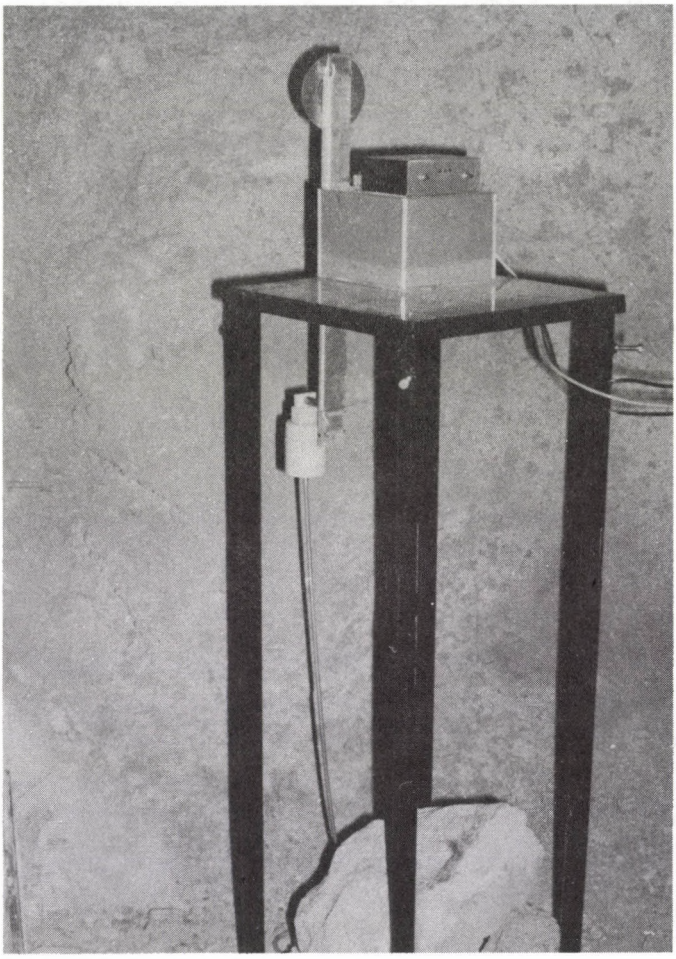

Fig. 15. The lifting mechanism of the mercury vessel

gravity changes can be recorded by a very sensitive barograph or the recorded atmospheric pressure can be used to correct the gravity measurements and so the extensometric measurements, too. For this purpose we have developed a microbarograph.

The principle of the microbarograph is shown in Fig. 18. The pressure sensor is a very sensitive closed diaphragm which produces large displacements between its two end faces for small air-pressure variations. This displacement is measured as follows: The standing plates of the differential condenser are fastened to the one end face of the closed diaphragm and the moving plate is joined to its other end face. The length changes of the membrane due to air-pressure variations cause

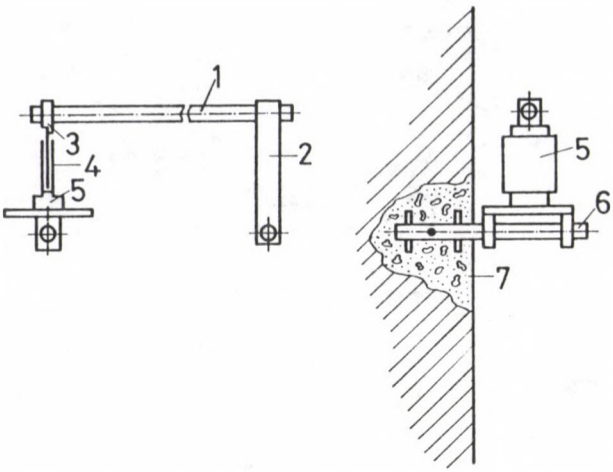

Fig. 16. The construction of the short extensometer: quartz tube (1), holding fixture of the quartz tube (2), fastening clamp of the moving plate (3), standing plates (4), mechanical support of the capacitive sensor (5), bolt holding the sensor unit (6), concrete (7)

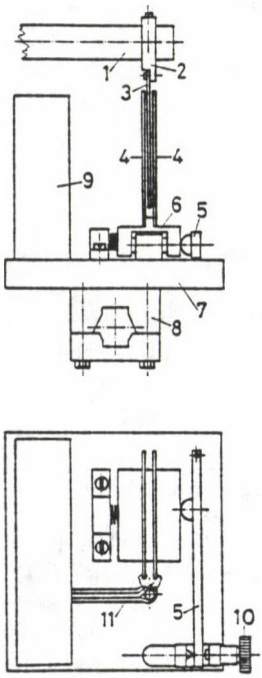

Fig. 17. The sensor unit of the short extensometer: quartz tube (1), fastening clamp of the moving plate (2), moving plate of the differential condenser (3), standing plates of the differential condenser (4), fine moving single-armed lever of the standing plates (5), moving slide of the standing plates (6), rigid base plate (7), fastening clamp of the sensor unit to the bolt (8), electronic unit (9), fine adjusting screw of the standing plates (10), lead-out wires of the differential condenser (11)

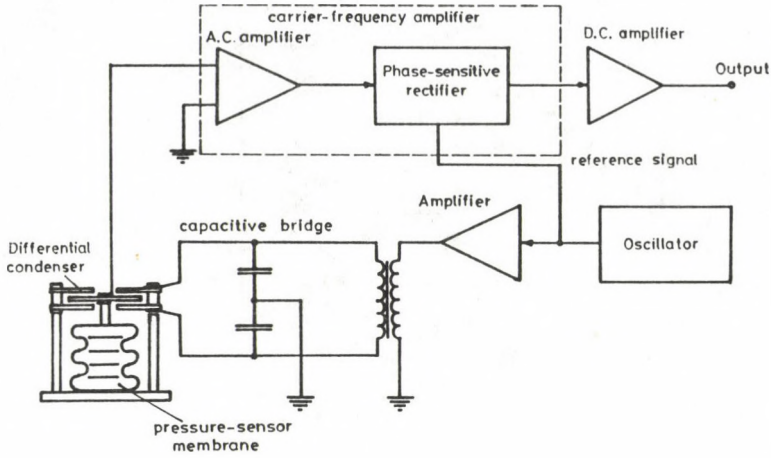

Fig. 18. The principle of the microbarograph

capacitance variations being measured in a bridge circuit in the known manner as done in the extensometer electronics.

The calibration of the instrument was made during the air-pressure recording. The output signal of the microbarograph was regularly compared by a barometer with an accuracy of 0.1 mbar. The sensitivity of the instrument according to the measurements is 341.8 μV/μbar and the linearity error of the microbarograph is less than 1 % in the whole measuring range.

7. Conclusion

The developed long extensometer and the sensor electronics have been working since many years satisfactorily. The measuring results prove that these instruments are suitable for the long term recording of very small displacements (Varga et al. 1993). In spite of the good results we continue the development of such instruments, since the increase of the accuracy and the stability is a steady requirement for a better interpretation of the measuring results at long term measurements.

Acknowledgements

The development of the microbarograph and partly the one of the short extensometer was supported by the Hungarian Science Fundation OTKA (Project No. 2131). Thanks are due to Tibor Molnár electrical engineer, Ferenc Schlaffer technician for their help in completion, maintenance of the instruments and to Gimesiné Németh Ágnes for her help in drawing the documentations of the instruments and for preparation of the measured data.

References

Mentes Gy 1981: *Acta Geod. Geoph. Mont. Hung.*, 16, 269–280.

Mentes Gy 1983: *Acta Geod. Geoph. Mont. Hung.*, 18, 359–368.
Mentes Gy 1991: *Bull. d'Information*, No.110, 7936–7939.
Mentes Gy 1993: *Bull. d'Information*, No.115, 8467–8471.
Varga T, Latynina L A, Brimich L, Mentes Gy, Katona Gy, Varga P 1993: *Bull. d'Information*, No.116, 8537–8545.

Acta Geod. Geoph. Hung., Vol. 29(1–2), pp. 179–195 (1994)

INTERPLANETARY MAGNETIC FIELD-DEPENDENT AND GEOMAGNETIC LATITUDE DEPENDENT PERIODS OF PULSATIONS — A RE-EVALUATION

J Cz MILETITS[1], L HOLLÓ[1], J VERŐ[1], B ZIEGER[1]

[Manuscript received March 13, 1994]

Some twenty years ago a detailed comparison was made based on data of the Nagycenk Observatory ($L \cong 2.0$) between periods of geomagnetic pulsations and interplanetary conditions. According to it, periods were clearly dependent on the scalar magnitude of the interplanetary magnetic field, (IMF), $|B|$, and even the form of the connection, $T = 160/|B|$ could be confirmed. Since then, the same problem has been several times attacked. The (geomagnetic) latitude dependence of the periods may veil this connection. Nevertheless, field line resonances are also influenced by interplanetary conditions, as e.g. the occurrence frequency of the resonance depends on IMF. Anyway, following types of the pulsation periods can be separated:

1. IMF dependent periods (upstream waves with slightly transformed spectra) everywhere in the zone studied (L 1.7 to 3 or more)
2. Latitude dependent periods, *i.e.* field line resonances, often simultaneously with upstream waves
3. Long period events with similar periods everywhere (from the tail?) mainly at higher latitudes
4. Short periods during geomagnetically active intervals (storms). It is not excluded that the IMF dependence of periods may be different at different L-values, even within the mid-latitude region, due to interference of the field line resonances.

Keywords: field line resonance; geomagnetic pulsations; Nagycenk geomagnetic observatory; upstream waves

1. Introduction

Based on data of the Nagycenk (NCK, $L \cong 2.0$) observatory, the connections between (Pc2-4) pulsation activity and several interplanetary-magnetospheric parameters have been studied since the International Geophysical Year, IGY, in 1957. Thus data and results for more than three complete solar cycles are available. Quite recently questions have been raised about the validity of the previously found connections. That is why we should like to summarize experience and results in this field, together with results concerning meridional changes of the pulsation activity (periods) based on results of several pulsation arrays in Central and Northern Europe.

[1]Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences, H-9401 Sopron, POB 5, Hungary

1217-8977/94/$ 4.00

In the following, we discuss mostly results confirmed by at least two, but in many cases by three or more independent sets of data. Data sets mean here when not otherwise stated all available data from one year, *i.e.* several thousand hourly values. As most results have been published previously (even if sometimes not in international journals), we do not want to give details about instruments, data collection and processing. Hourly occurrences and average amplitudes are from the pulsation catalogue of the observatory that contains these values from quick-run records for 12 period bands, namely

P1	1– 5 s	P7	30– 40 s
P2	5–10 s	P8	40– 60 s
P3	10–15 s	P9	60– 90 s
P4	15–20 s	P10	90–120 s
P5	20–25 s	P11	120–300 s
P6	25–30 s	P12	300–600 s

It is to be added that since 1987, the bands P1 and P2 are too disturbed for an estimation of the corresponding parameters. Amplitude and period values are hand-scaled, but due to the great experience of the colleague who makes the processing, they are quite stable, moreover, they are made in a routine basis, *i.e.* they cannot be influenced by wishful thinking.

In a few cases daily pulsation indices are also used. They indicate the activity

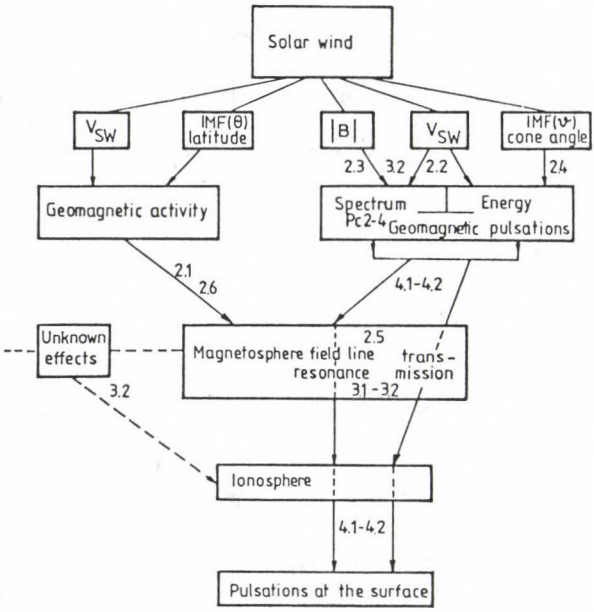

Fig. 1. Scheme of the connections surveyed in this paper. Numbers indicate sections where the corresponding connections are treated

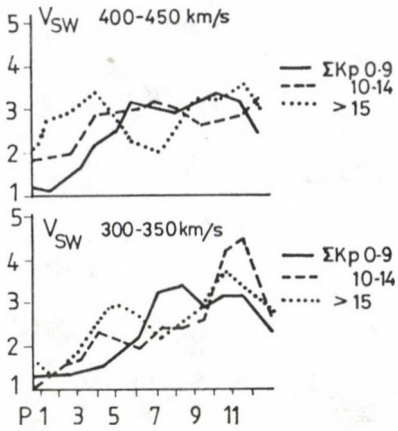

Fig. 2. Effect of the geomagnetic activity, ΣK_p, on the pulsation activity in two ranges of the solar wind velocity (for P1–12, see text) (Verő 1975)

of the period range 0–2 min by a quasi-logarithmic index between 0 and 9 (K1), as determined from normal records.

Observatories and temporal stations included in the arrays discussed in Section 3 are variable, but they are mostly restricted to a zone North/South of NCK. In recent arrays, mostly high-resolution dynamic spectra are used, details can be found in the original publications, too.

Data amounts of the arrays are naturally much less than in case of the NCK data, nevertheless, we only included into this paper generally valid conclusions, not exceptional cases.

Figure 1 gives a scheme on the connections to be surveyed in this paper, but it is not meant as a summary of physical processes.

2. Effects of interplanetary-magnetospheric parameters on the Pc2–4 activity

2.1 Effect of the geomagnetic activity

In the chronological order, the first (magnetospheric) parameter being known to influence pulsation activity was geomagnetic activity. Geomagnetic activity (as expressed e.g. by K_p) is naturally influenced by interplanetary parameters, mainly by solar wind velocity (V_{SW}) and by the latitude (Θ) of the direction of the interplanetary magnetic field (IMF). As far as possible, we tried to separate effects due to the geomagnetic activity and due to V_{SW}, as the latter may directly influence pulsation activity, too. From this point of view, Θ is less problematic, as it does not seem to influence significantly pulsations.

Increasing geomagnetic activity shifts pulsation periods towards shorter ones as early results have shown (Verő 1958). It is, however, possible that this shift is due to increasing V_{SW} or even to increasing scalar magnitude, $|B|$, of IMF.

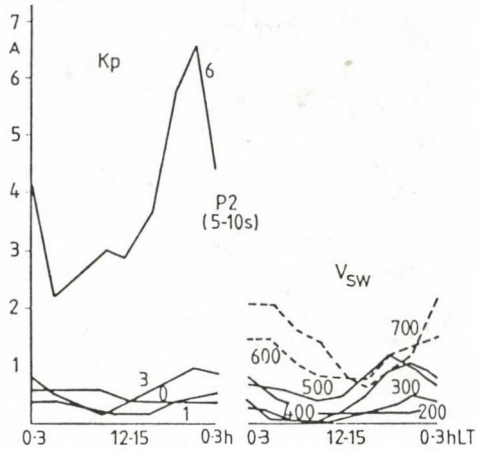

Fig. 3. Activity in the period range 5–10 s (P2) *vs.* K_p and V_{SW}, during different local times. (V_{SW} values are lower limits in 100 km/s wide bins) (Verő 1980)

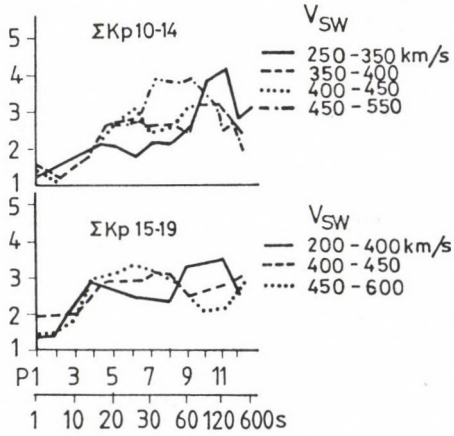

Fig. 4. Effect of the solar wind velocity, V_{SW}, on the pulsation activity at two levels of the geomagnetic activity (Verő 1975)

Evidences show that local field line resonances (e.g at NCK, 20 s periods) are less active at higher K_p-values (Fig. 2) if V_{SW}-effects are excluded. Field line resonances (at least in the longer period part of the corresponding range) are strongest at K_p-values of 1 or 2, somewhat less strong at K_p 0, and decrease continuously at K_p-values higher than 2.

The most significant effect of high geomagnetic activity is in the Pc2-range, namely the average amplitude of P2 (5–10 s) increases about 8 times at $K_p \geq 6$, and only about 3 times at high V_{SW} (Fig.3). Anyway, the increasing short-period activity connected with high geomagnetic activity is to be considered when analyzing $|B|$-effects.

2.2 *Effect of the solar wind velocity*

In contrast to geomagnetic activity, high solar wind velocity is favourable for field line resonances. Figure 4 shows that P5–P6 (20–30 s) activity increases by more than a factor of 2 between solar wind velocities 300 and 500 km/s. The activity in this period range increases smoothly with V_{SW} up to the highest speeds observed (Fig. 5).

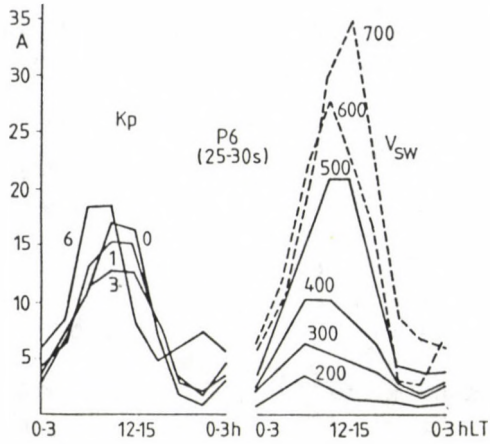

Fig. 5. Activity in the period range 25-30 s (P6) *vs.* K_p and V_{SW} (as Fig. 3 for P2) (Verő 1980)

Supposing a connection of the type $A(Pn) = c_n \cdot V_{SW}^{x_n}$ (here $A(Pn)$ means average amplitude in the range Pn, c_n a constant, and x_n the exponent of the solar wind velocity), the x_n-values are the following in some data sets:

	Year 1972	Year 1974	
	all data	all data	cone angle 20 – 40°
P2	2.9		
P3	3.5	3.0	2.7
P4	4.5	2.6	2.7
P5	2.2	2.3	1.8
P6	2.3	1.5	1.4
P7	1.0	1.0	1.0
P8	0.8	1.0	0.4
P9	0.5	−0.2	−0.4
P10	−0.3	−0.8	−0.8
P11	−0.2	0.8	0.7
P12	1.8	1.8	2.3

These values indicate the similarity between different data sets, too, with a slight shortening of the most affected period range in 1974 as the only significant effect. It is, however, to be kept in mind that actual values are to be taken with caution, as data used in the computation are hand-scaled values.

The fact that amplitudes in the shorter period part of the Pc2–4 range increase more quickly with V_{SW} means that spectra are shifted toward shorter periods at high solar wind velocities. Therefore the dependence of the (peak) period, T on $|B|$ can only be clearly identified at lower V_{SW}-values.

The increasing activity of the field line resonances results among others in an increasing share of regular — very regular pulsations at solar wind velocities higher than average (at 300 km/s, about 10 percent, at 500 km/s, about 20 percent, and at 700 km/s, about 15 percent of the pulsations is of a regular sinusoidal form).

The total pulsation activity (as a sum of amplitudes in the different period ranges, or K1) depends approximately linearly on V_{SW} (Verő and Holló 1983). At different local times, the connection between V_{SW} and pulsation activity does not change significantly, in contrast to pulsation activity at satellite altitudes (ATS 6).

2.3 Effect of the IMF scalar magnitude, $|B|$

Before going into details of the connection between $|B|$ and pulsation activity, the distribution of $|B|$ values should be given. In 1972 (from 56 percent of all possible values), 33.5 percent of $|B|$ values were less than 5 nT, 58 percent between 5 and 10 nT, and 8.5 percent more than 10 nT. Taking into account that the field line resonance has a characteristic period of 20 to 25 s at $L \geq 2$ (valid for NCK), most frequent values of $|B|$ (at least in the time interval studied) correspond to the period of the local field line resonance (from $T = 160/|B|$, at $|B|$ values between 6 and 8 nT), and therefore the periods of upstream waves and of field line resonance often coincide.

At any value of $|B|$, the most likely period to be found (peak of the spectrum) is $T = 160/|B|$. The only restriction to this statement is that if K_p and/or V_{SW} are high, the spectrum may be shifted towards shorter periods. Otherwise the statement remains valid even if field line resonances (e.g. very regular pulsations) are considered.

The amplification at the peak in the spectrum is, however, not very strong, with respect to the next minimum (or flat part) at longer periods, about 1.7-2.5 times (Fig.6); these minima are at periods about twice the period of the maximum.

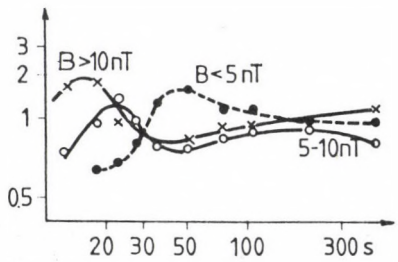

Fig. 6. Amplification of the pulsation amplitudes at different IMF scalar magnitudes, $|B|$, (amplitudes at given $|B|$ relative to the average amplitude) (Verő and Holló 1978)

The frequency of occurrence of regular pulsations increases from about 8 percent at 4 nT, to 14 at 6 nT, to 27 at 9 nT and to 38 at 11 nT, then above 12.5 nT it drops again to 9 percent.

There are some "secondary" effects together with the shift of the maximum: amplitudes of periods longer than 60 s decrease at high $|B|$-values (this may be some shielding effect of the short period activity in the processing); at very low $|B|$, there is a second peak at about 20 s, corresponding to the local field line resonance period, supposedly due to impulsive excitation of the field lines.

As mentioned in the previous section, high V_{SW}-values shift the spectrum towards shorter periods. Figure 7 shows data about this complex connection. At solar wind velocities 300–400 km/s, observed T-values correspond to those computed from $160/|B|$, at 500–600 km/s, periods are too short at low IMF, and at $V_{SW} > 700$ km/s, there is practically no connection between $|B|$ and T. This situation may result in low correlation between T and $|B|$ at high V_{SW} (or K_p) values.

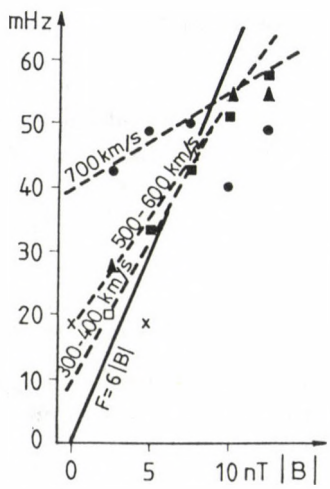

Fig. 7. Position of the peaks in the spectra of the pulsations at different solar wind velocities *vs.* $|B|$, with corresponding linear approximations (Verő 1986)

2.4 Effect of the IMF direction — cone angle (ϑ)

In the first survey carried out with NCK pulsation data and IMF direction, most advantageous latitude (Θ) of the IMF for the excitation was found at 0°, with decrease of the average amplitudes towards both higher negative and positive latitudes, even more quickly towards negative ones. Similarly, longitudes (Φ) around 0° and 180° were advantageous, perpendicular directions unadvantageous.

In a more detailed analysis, the maximum of the amplitudes in the range P4–P5 (15–25 s) was found at a cone angle (ϑ) of 30°, and in the range P6, there was a

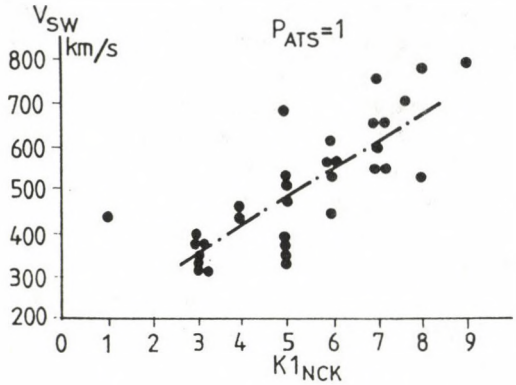

Fig. 8. Pulsation activity at NCK (K1) *vs.* V_{SW} on days with fixed, low pulsation activity ($P_{ATS} = 1$) on the satellite ATS 6 (Holló and Verő 1987)

flat maximum up to this angular value (Verő and Holló 1978). Similar surveys have been repeated three times, always with the same result, *i.e.* maximum at a cone angle of 30°, but only around the local field line resonance.

Not only amplitudes, but also the occurrence of regular pulsation is most probable at y-values around 30°. Occurrence frequencies of about 17 percent were found at lower and higher cone angles, with a maximum of 21 percent at 30°. Thus it seems that field line resonances are most likely if the cone angle is around 30°.

The effect of the cone angle was found to be very similar at different solar wind velocities, up to 700 km/s and above.

2.5 Magnetospheric effects

In this section, several effects are summarized which are supposed to be due to different magnetospheric conditions, but not immediately connected to field line resonance.

a) Average daily pulsation activities are influenced by the previous activity for a longer time than geomagnetic activity. Geomagnetic activity is in high approximation a Markovian process, whereas pulsation activity is non-Markovian. The difference can be seen on spectra on days with similar geomagnetic activity on the actual day, but with different activities previously (Verő 1974). The "memory" for past activity is most likely to be found in the position of the plasmapause.

b) The position of the magnetopause is closely connected with e.g. geomagnetic activity. It was found that the amplitude inversely depends on the dimension of the magnetosphere in the range P3 (10–15 s) if the magnetosphere is small, and in the range P4–P5, if it is large. As the amount of data was not sufficiently large in this case, it is not sure whether the period of the field line resonance could be really influenced by the dimension of the magnetosphere.

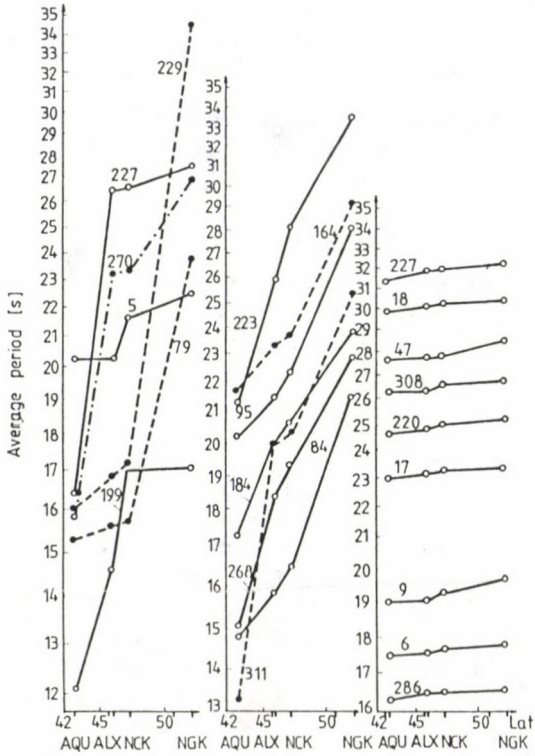

Fig. 9. Pulsation periods of some selected events (H-component) *vs.* geomagnetic latitude (Cz Miletits 1980)

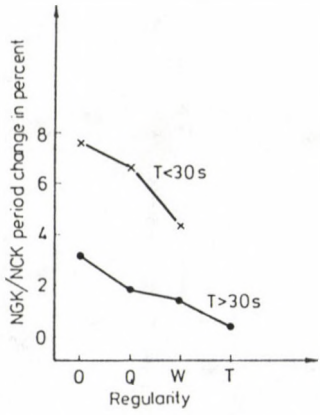

Fig. 10. Connection between average change of the pulsation period for one degree of geomagnetic latitude and the "regularity" of the pulsation event (regularity decreases in the order O, Q, W, T) (Cz Miletits 1980)

c) There are some seasonal effects in the pulsation activity. For example, in summer the morning start of the activity occurs earlier than in winter, moreover, the occurrence of regular pulsations is more frequent in summer. (This may be connected to the winter attenuation of field line rezonances.)

d) A comparison with data of the satellite ATS 6 has shown that the pulsation activity at the altitude of ATS 6 sets a lower limit to the pulsation activity at NCK. The NCK activity above a limit set by the ATS 6 activity depends on the solar wind velocity. With other words, if NCK activites are plotted *vs.* V_{SW} for a given level of the pulsation activity at ATS 6, a rather close correlation exists between both (Fig. 12). That means that incoming upstream waves are modified within the magnetosphere, and the amplification there depends on solar wind velocity, too. This is an indirect proof of the solar wind influence on magnetospheric parameters.

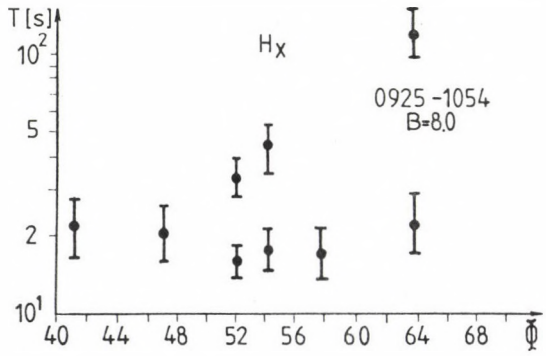

Fig. 11. Position of the spectral peaks in an interval with both field line resonances and slightly transformed upstream waves (Cz Miletits et al. 1990)

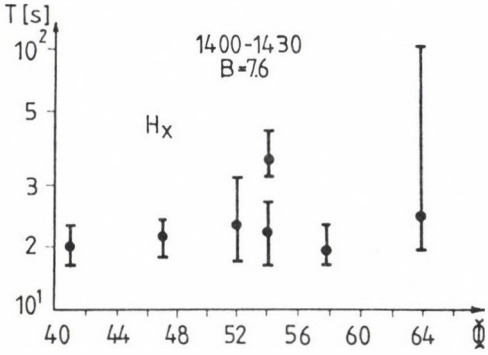

Fig. 12. Spectral peaks for an interval with only slightly transformed upstream waves; field line resonances are, nevertheless, apparent at 54 (WRH) and perhaps at NGK, too (broad peak) (Cz Miletits et al. 1990)

2.6 Effect of geomagnetic impulses (SI-s)

Since Veldkamp has found in 1958 a connection between SI-s and geomagnetic pulsations, this connection has been studied several times. The effect can be described so that longer period (30–60 s) pulsations present before the impulse get amplified during, and disappear after, the impulse. Sometimes after the impulse, shorter period (15–25 s) pulsations appear (Verő and Tátrallyay 1973). The duration of the effect is about 1–2.5 hours.

Subsequent results have shown that rather small SI-s (not qualified as those from geomagnetic observatory records) have more significant effects than greater ones. Statistical data proved a duration of these effects up to 12 hours (Holló and Verő 1985).

It is likely that such small SI-s are connected with switch-on and switch-off events in the upstream waves, in the pre-magnetospheric solar wind. Switch-ons are more likely to coincide with these small SI-s which can be then identified on the records in a rather great part of the events. The effect of switch-ons is similar to that of SI-s, as periods of 20–40 s are most influenced, and the duration of the effect is 2 to 3 hours. Switch-offs, on their part, are accompanied by period decrease, as mentioned for SI-s previously.

3. *L*-dependent pulsation periods and field line resonances

3.1 Short duration events

The disjunction of short-duration and long-duration events is certainly artificial in the context of pulsation arrays, there are notwithstanding certain other differences in the data sets to be discussed in this and in the following section that justify the separation.

Short-duration events consist of a dozen individual cycles (certainly quite often single "beats" due to interference of two periods differring by a fraction of the common period). The events of the 1977 array were selected to have these characteristics, and from one month of data, about 400 such events were processed. In addition, the 1977 array included a station pair with a distance of about 150 km between them (somewhat more than one degree of latitude), where periods could be compared at such a small distance, too.

From the present point of view, significant results of this array (and of some other small-scale campaigns) are the following (Cz Miletits 1980):

1. Pulsation periods do change in a significant part of the events with latitude (or with the L-value), in other events the period remained the same along the whole array (at least from L 1.7 to 3.0). Figure 9 shows some examples.

2. If pulsations are of a regular (*i.e.* sinusoidal) form, it is more likely that the period changes with the latitude. (Fig. 10).

3. In events when the period changes with latitude, there may be station pairs the records of which have the same period, *i.e.* the period does not change

smoothly with latitude, but there are steps in the function period *vs.* latitude. It is to be remarked that such cases were most often found in case of the station pair having the 150 km distance between them. As later arrays did not contain such neighbouring stations, this observation was not confirmed by a later study. Figure 9 shows a lot of examples for these steps, too.

4. The average rate of the period change is about 8 percent for one degree of latitude in case of stations located at around L 2 to 3.

These observations can be interpreted as field line resonances, (in the case of the steps more exactly, resonances of shells of field lines) in the case of L-dependent

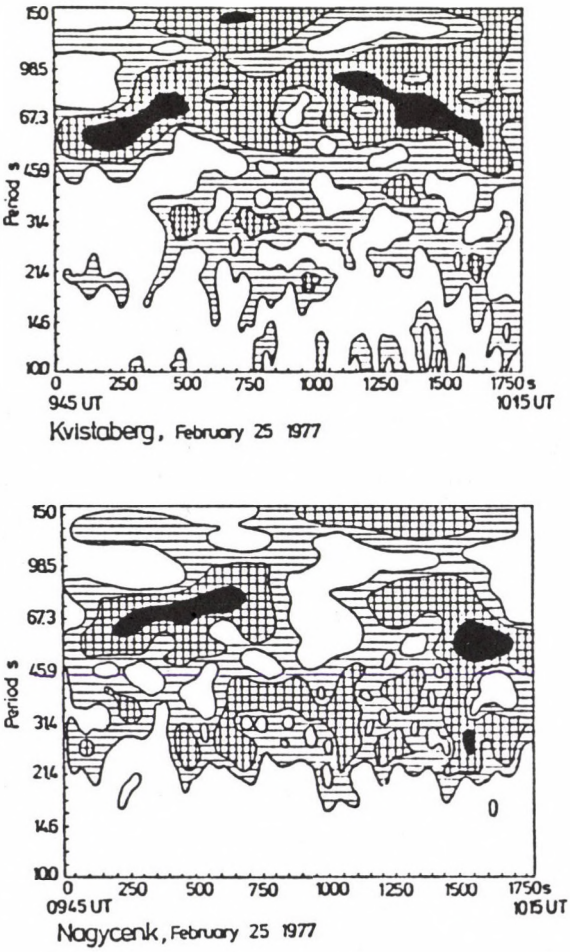

Fig. 13. Dynamic spectra from Kvistaberg and NCK, February 25, 1977, 0945–1015 UT. Upwards and downwards trends are visible in the period of the peak around 67 s (Cz Miletits et al. 1988)

periods, and waves conserving their original spectrum in the case of constant periods along the array. From the distance between the stations, the width of the shells was supposed to be one or a few degrees of latitude.

3.2 Long-duration intervals

Long-duration intervals are no "events" in the strict sense of the word, they have durations of 30 min or more. Such intervals were selected from the records of a campaign in 1984, including partly the same stations as the 1977 array, with a latitudinal extension being similar to the earlier one. Nevertheless, the later array did not include neighbouring stations, but some stations were at similar latitudes at slightly different longitudes (e.g. NCK and Fürstenfeldbruck). In the following, data not only from this array are considered, but similar "intervals" from other times, too.

The main results are the following (Cz Miletits et al. 1990):

1. Similarly to short duration events of the previous array, both latitude dependent and constant periods were found in this material, too. The occurrence frequency of the two types, however, was quite different in the two sets.

2. Several intervals were found when both types occurred simultaneously. In some cases, field line resonances only covered a part of the array, in other cases, they were simultaneously present along the whole array. Most frequent occurrence was observed at L values 2.5 to 3 (NGK, WRH). Periods of a typical event are presented in Fig. 11.

3. Field line resonances occur at latitudes where the local field line resonant period is longer than that due to upstream waves, *i.e.* $T = 160/|B|$. At stations lying near to the intersection of the field line resonance and of upstream waves, a strong maximum of amplitudes is found, but at even lower latitudes, *i.e.* at significantly shorter periods no field line resonances were observed (Fig. 12, and the effect is apparent in Fig. 11, too).

4. An analysis of beating structures (Verő and Cz Miletits 1994) led to the idea that they are due to field line resonances from neighbouring shells. These periods differ in average by 10 percent, *i.e.* with the previously found rate of the period change *vs.* latitude, the width of the shells is around 1°. According to this supposition, beating structures characterize field line resonances.

5. Frequency shifts occur sometimes in the pulsation records. They are mostly found at longer periods (at 40 s or more). Slow changes of the frequency are typical for upstream waves when $|B|$ changes, too. Such a case is shown in Fig. 13. In a few cases the change in the IMF could be identified with frequency shifts of the pulsations. (Such "slow" shifts are to be separated from sudden changes of the period, as the latter can be due to switch-on and switch-off events of the field line resonance, too) (Cz Miletits et al. 1988).

6. Impulses in the geomagnetic field/pulsation range were nearly without exception of the constant period type, as if impulses would not excite field line resonance. This fact is conflicting with the spectra for very low $|B|$-values at NCK, where a secondary peak at the local field line resonance was attributed to impulsive excitation of the field line resonance.

7. Figure 8 in Verő and Cz Miletits (1994) shows that the hyperbola $T = 160/|B|$ *vs.* $|B|$ crosses mostly the bottom part of the period range of all stations during a given time interval. Points corresponding to the data pairs $(T, |B|)$ of all stations belong to three zones: Zone C contains "upstream waves", Zone B field line resonances, Zone A perhaps pulsations from the tail (there is a misprint in the legend of this figure in the original paper).

The occurrence frequency of field line resonances changed significantly in different data sets collected at arrays. As based on data of a single station we cannot unambiguously separate the two types, field line resonances and upstream waves, from each other, it is impossible to carry out a more exhaustive survey of this problem. It would be moreover important to have data on the occurrence of field line resonances, as without them, the factors governing the excitation of field line resonances cannot be found. An immediate comparison of upstream waves observed in the near-Earth space and of surface pulsations confirmed the close connection between the two events (Verő et al. 1993), but did not yield conclusive evidence for a direct influence of upstream waves on field line resonances.

4. Discussion and conclusions

4.1 Expected and observed pulsation spectra at three L-values

The following facts follow from the previous observations and considerations:

1. Both slightly modified upstream waves and pulsations due to field line resonance can be observed at the surface, at mid-latitudes. The characteristics (e.g. period) of the two kinds are rather similar, nevertheless, there are some differences between the two types:

   a) Pulsations from field line resonances are more often regular (sinusoidal), or even very regular (spectra with narrow peaks), than upstream waves are.

   b) Field line resonances are more active than average at high solar wind velocity and are less active at high geomagnetic activity (or they are veiled by storm-time short period pulsations).

   c) The maximum occurrence of field line resonances can be supposed at cone angles around $\vartheta = 30°$, that of upstream waves around $\vartheta = 0°$.

   d) There are specific signatures of both types in dynamic spectra, as e.g. frequency/period shifts in the case of upstream waves (due to changes in the IMF) and beating structures in the case of field line resonances

(due to interference of neighbouring shells with slightly different resonant periods).

2. Field line resonances may occur with different frequency of occurrence at different L-values (geomagnetic latitudes) (Vellante et al. 1993). The standard plot period *vs.* latitude is valid between latitudes of about 40° and auroral latitudes with a maximum frequency of occurrence around $L \simeq 2.5$, 50° latitude (Fig. 14). At lower latitudes, mostly upstream waves and periods longer than expected are present, at auroral latitudes, the standard periods are sometimes found, but due to higher geomagnetic activity, they are often veiled.

3. The incoming spectrum of the upstream waves (being the energy source of field line resonances, too) has a sharp upper frequency limit (at short periods) and a rather long flank towards low frequencies (long periods) (Varga 1980). Thus there is no energy to excite field line resonances above the frequency $f = 6.|B|$, but long period resonances are often excited.

4. Both high geomagnetic activity and high solar wind velocity shift the spectrum towards short periods. Therefore field line resonances (and upstream waves) are to be studied at low geomagnetic activity and solar wind velocity.

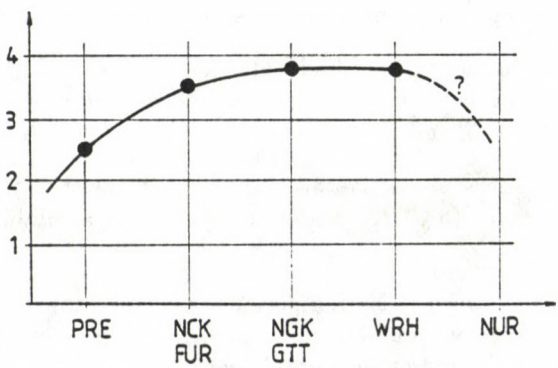

Fig. 14. Occurrence of field line resonances along the array, at L values around L 1.7 (PRE), 2.0 (NCK, FUR), 2.5 (NGK, GTT), 3.0 (WRH) and at even higher latitudes (NUR)

It is to be remarked here again that NCK is situated in the central part of the zone where IMF scalar magnitude most significantly influences pulsation periods. At latitudes below 40°, field line resonances are hardly excited at low $|B|$-values, *i.e.* the long period part of the spectrum may be lacking. At higher latitudes, the short period part of the spectrum can be veiled by strong long-period field line resonance.

4.2 Conclusions

The main conclusions from the present re-examination of earlier results are the following:

1. There are two types of Pc2-4 pulsations which can be hardly distinguished based on records of a single station.

   Slightly modified upstream waves conserve their spectra from the interplanetary medium, i.e the peak is found at $T_{SW} = 160/|B|$, with a shift towards higher frequencies at high solar wind velocity.

   Field line resonances are also governed by the spectrum of the upstream waves, parts of the spectrum corresponding to the local field line resonant period are amplified in a zone where the resonant periods are included in the spectrum of the upstream waves (see also Vellante et al. 1993).

2. Some features of the two pulsations types may be different, e.g. different cone angles are advantageous for their excitation, the widths of the spectral peaks are different, field line resonances are promoted by high solar wind velocity and hampered by high geomagnetic activity *etc.*

3. Field line resonance and upstream waves occur often simultaneously with a variable ratio of the two types. Field line resonances are most likely at L-values around 2.5 to 3, below $L = 2$, field line resonances are connected to higher $|B|$-values.

4. At very high geomagnetic activity ($K_p \geq 6$) a sudden increase of the pulsation amplitudes indicates the appearance of storm-time pulsations.

5. High geomagnetic activity and high solar wind velocity shift spectra towards shorter periods, both for upstream waves and field line resonances, thus the connection between T and $|B|$ can be only found for lower K_p and V_{SW} values.

6. Field line resonances are active at frequencies below the peak frequency of the upstream waves (6. $|B|$), above it there is no available energy for the excitation.

7. The connection between pulsation periods and IMF scalar magnitude is best reflected in pulsations around $L \simeq 2$, as the most frequent $|B|$-values correspond to the field line resonant period there.

8. The occurrence frequency of field line resonances changes significantly in time. Thus their efficiency is controlled by some (magnetospheric) agent. This else unknown agent is influenced by the solar wind velocity, and it is correlated with electron concentrations both in the F2 region of the ionosphere (Verö and Menk 1986) and in the equatorial magnetosphere at distances of a few R_E, where whistlers propagate (Verö 1965).

Field line resonances are capable to increase the surface pulsation activity at mid-latitudes irrespective of other interplanetary conditions.

These conclusions may change somewhat in different time intervals e.g. due to changes in the average parameters of the interplanetary medium as such changes may influence field line resonances, too. Very little is known on the changes of the

periods of field line resonances with time. A perhaps related time variable effect is the mentioned winter minimum of the pulsation activity at high solar activity, *i.e.* at high ionospheric-magnetospheric particle densities.

Acknowledgements

The authors thank A Best, I Best and M Vellante for records of the observatories Niemegk, Warnkenhagen and l'Aquila, to them and to H Lühr and J Streštik for fruitful discussions. The research has been supported by a Hungarian state grant, OTKA No. 1171.

References

Cz Miletits J 1980: *J. Atm. Terr. Phys.*, 42, 563–567.

Cz Miletits J, Verő J, Stuart W 1988: *J. Atm. Terr. Phys.*, 50, 649–656.

Cz Miletits J, Verő J, Szendrői J, Ivanova P, Best A, Kivinen M 1990: *Planet. Space Sci.*, 38, 85–95.

Holló L, Verő J 1985: *J. Atm. Terr. Phys.*, 47, 523–528.

Holló L, Verő J 1987: *J. Atm. Terr. Phys.*, 49, 1147–1153.

Varga M 1980: *J. Atm. Terr. Phys.*, 42, 365–369.

Vellante M, Villante U, Core R, Best A, Lenners D, Pilipenko V A 1993: *Ann. Geoph.* (in press)

Verő J 1958: *Zschr. f. Geoph.*, 24, 307–317.

Verö J 1965: *J. Geoph. Res.*, 70, 2254–2257.

Verő J 1974: *Acta Geod. Geoph. Mont. Hung.*, 9, 291–297.

Verő J 1975: *J. Atm. Terr. Phys.*, 37, 561–564.

Verő J 1980: *J. Atm. Terr. Phys.*, 42, 371–380.

Verő J 1986: *J. Geoph., 60*, 106–120.

Verő J, Cz Miletits J 1994: *J. Atm. Terr. Phys.*, 56, 433–445.

Verő J, Holló L 1978: *J. Atm. Terr. Phys.*, 40, 857–867.

Verő J, Holló L 1983: *Acta Geod. Geoph. Mont. Hung.*, 18, 403–419.

Verö J, Menk F W 1986: *J. Atm. Terr. Phys.*, 48, 231–245.

Verő J, Tátrallyay M 1973: *J. Atm. Terr. Phys.*, 35, 1507–1515.

Verő J, Zieger B, Lühr H 1994: In: Solar Wind Sources Magnetospheric Ultra-Low-Frequency Waves. Geophysical Monograph Series 81. American Geophysical Union, Washington, 365–373.

Acta Geod. Geoph. Hung., Vol. 29(1–2), pp. 197–208 (1994)

A SHORT HISTORY OF EARTHQUAKE RESEARCH IN HUNGARY

GY SZEIDOVITZ[1]

[Manuscript received April 14, 1994]

The history of seismological research in Hungary is summarized. The seismograms, available in Hungary, are listed.

Keywords: earthquake research; history of seismology; seismograms

The first major scientifically investigated earthquake in Hungary burst out in the vicinity of Komárom in 1763, in consequence buildings were seriously damaged and a lot of people died. The different effects of the earthquake exerted on its surroundings were discussed in details in the contemporary press, and the earthquake became well-known in the whole Austro-Hungarian Monarchy. It is likely that the subsequent detailed study of the shocks in Hungary was inspired by this devastating shock.

The first comprehensive Hungarian catalogue on the earthquakes was edited and published by Grossingen in 1783. It was followed by studies by Sternberg (1786), Kitaibel and Tomtsányi (1814), Holéczy (1824), Hunfalvy (1859), Saly (1860), Jeittels (1860) and Koch (1880). In addition, the earthquake activity of particular minor areas was investigated by a lot of authors. It may be worth listing a few authors among our predecessors without striving at completeness: Inkey (1877, 1881), Hantken (1882), Schafarzik (1880, 1889, 1901), Réthly (1906, 1907, 1908, 1909, 1912, 1914, 1915, 1952), Szilber (1914), Csengeri (1916), Schréter (1925), Moravetz (1925), Simon (1926, 1931, 1937).

Proposed by Schafarzik, a Permanent Earthquake Committee was established within the Hungarian Geological Society in 1881, as the second in Europe, following the Swiss one. Its first president was József Szabó, an eminent geologist. The following people were elected to the committee: Lajos Lóczy, as vice-president, Ferenc Schafarzik, Miksa Hantken, Tamás Szontágh, Miklós Vály a, as members. At the first meeting of the committee, they decided to send letters in Hungarian and German to each editorial office in Hungary as well as to the towns in the country in which the description of the earthquake observations and its brief explanation were summarized. They enclosed questionnaires to the letters, in preparing of which they used the similar material of A Heim, a professor in Zürich, as an aid.

Besides the acquisition and evaluation of macroseismic data, they regarded the building up of a domestic seismological network as their important task.

[1]Seismological Observatory, Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences, H-1112 Budapest, Meredek u. 18, Hungary

1217-8977/94/$ 4.00

It would exceed the size of this brief historic summary to recite where and what kind of instruments were set and from what time till what time they operated, therefore we only refer to Bisztricsány and Csomor's study (1981), in which more detailed answer can be found to the questions in connection with the seismological stations.

In 1900, the Earthquake Committee sent Radó Kövesligethy to the Strassburg Institute to study the building up and instrumentation of the seismological station operating there. It turned out from his travel report that he visited other earthquake observatories, too (Kövesligethy 1900).

In the 1902 report of the Ógyalla Observatory, Réthly (1909) had already reported the establishment of some seismological stations. The first station began to operate in Kalocsa. In 1901, a pair of horizontal pendulums were installed in Ógyalla, and somewhat later in Budapest, in the cellar of the Geological Institute with the assistance of Dr Ignác Darányi and Dr Andor Semsey. 5000 koronas were made available by the Ministry of Agriculture, from which they wanted to buy 5 Vicentini pendulums. The instruments were planned to be set up in Budapest, Ógyalla, Segesvár and Fiume. At the first seismological meeting in Strassburg, 1901, Schafarzik gave an account on the activity of the Hungarian Earthquake Committee. Kövesligethy reviewed the earthquake catalogue compiled by Ferenc László. Kövesligethy was elected as the seventh member of the international committee by the participants of the meeting. The Earthquake Committee received regular support from the Hungarian Academy of Sciences from the year 1902 on.

From 1903 on, the seismological research activities were continued in the framework of the Meteorological Institute. The macroseismic reports on the earthquakes occurred in Hungary were edited by Réthly (1909), which were first published as the publications of the Institute for Meteorology and Geomagnetism, and then under the auspices of the University Earthquake Observatory directed by Kövesligethy. The annual microseismic reports on the activity of the earthquake observatories in Hungary were edited by Kövesligethy.

The report on the earthquakes recorded at the Hungarian stations between 1913 and 1919 was published after World War I 1920. In its introduction Kövesligethy (1920) gave a summary of the damage caused in the seismological network. Hungary lost the stations Fiume, Ógyalla, Kolozsvár, Ungvár and Temesvár, and the observatories of Budapest, Kalocsa, Kecskemét and Szeged only remained.

The next microseismic report edited by Simon was published in 1926, and it only contained the shocks recorded at the station Budapest.

As the macroseismic observations, the situation was even more unfavourable: no reports were made on the shocks from 1913 till 1929. In fact, this gap remained unfilled up to now. Though Réthly (1952) collected the earthquake observations from the years before 1918 in his book Earthquakes in the Carpathian Basin, a ten years interval remained unprocessed yet. Csomor and Kiss (1962) published in a catalogue like form a list of shocks occurred in this particular period in their study on the seismicity of Hungary, but they did not deal with the detailed description of the individual shocks. The same applies to the catalogue edited by Zsíros et al. (1988).

After World War I, the Hungarian earthquake research gradually recovered. The macroseismic and microseismic reports generally edited by Simon and Mrs Szilber and controlled by Kövesligethy were published more or less regularly. After Kövesligethy's death (1934), Béla Simon became first the temporary and then the appointed director of the Earthquake Observatory of Budapest. In the preface of the 1938 microseismic report, he announced that the name of the seismological Institute of Budapest had been changed for Péter Pázmány University National Earthquake Observation Institute according to the decision of the Minister of Religion and Public Education (24759–1937/IV.). The shorter name of National Earthquake Observatory was used in the reports.

In World War II during the siege of Budapest, 12 destructive and two incendiary bombs fell upon the headquarters of the Institute (Deák tér 21, Budapest) (Simon 1946). As a consequence, the building was consumed by fire and collapsed down to the bombshelter, thus the manuscript of the macroseismic material for the years 1943–44 was destroyed, too. The list of the available seismograms has been compiled in Table I. From 1943 till 1949, the macroseismic and microseismic reports were published by the National Earthquake Observation Institute of the Péter Pázmány University of Budapest. The macroseismic and microseismic reports after 1951 were published under the name National Earthquake Observation Institute. The Institute kept its name unchanged for a relatively long time, only the name of the then mother institute was added to it depending on which major institute it was transformed. After 1951, the Institute belonged to the Ministry of Mining and Energy for one year and then to the Loránd Eötvös Geophysical Institute of Hungary (ELGI) between 1952 and 1963. Directed by Professor László Egyed, the building of the seismological observatory of Sashegy started in Budapest in 1955. The seismological investigations were launched at the Geophysical Department of the Loránd Eötvös University. Egyed supplemented the Hungarian seismological network with stations established at Piszkéstető and Sopron. In 1963, the Seismological Division of ELGI was transferred to the Geophysics Department of ELTE as an academic research team, but it continued to present itself as the National Earthquake Observation Institute in international relations. After Professor Egyed's death (1970) the research team of the department joined the Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences on the proposal of Ede Bisztricsány, the head of the research team, and it has been an independent division of this institute since 1971.

Bisztricsány retired in 1987 and Győző Szeidovitz was appointed as head of the Seismological Department.

During the last years a few seismological stations have been closed, and two new stations have been installed.

At the present time five seismological stations (Budapest, Sopron, Piszkéstető, Paks and Uzd) run in Hungary.

Recently at the station Piszkéstető, a new Adebahr-Quanterra broad-band instrument of high sensitivity and dynamics has been put up. Data are freely available for everybody through the X.25 network. This station has been integrated into the German seismological network.

Table I.

Name and coordinates of the observatory and the period in which the instruments were used	Type of instruments	Head of the Observatory	Seismograms begin stored in	References and remarks
BUDAPEST	19°03′55″ EGr 47°29′29″ N			
1902–1905	Strassbourg horizontal further Bosch	S. Kalecsinszky K. Emszt	–	1/a, 3, 4, Worked in the cellar of Hungarian Geological Institute
1906–1912	Bosch Vicentini-Konkoly Wiechert (1000 kg)	R. Kövesligethy A. Pécsi	–	5, 6, 7, Worked in the cellar of the Hungarian National Museum
1913–1920	Wiechert (1000 kg)	R. Kövesligethy	–	8
1920–1926	"	"	from 1926 Kecskemét	9
1926–1934	"	"	"	11, 12, Kövesligethy died
1935–1945	"	B. Simon	"	12, 13, 14, 15, The observatory was out or order
1946–1949	Wiechert (1000 kg) Krumbach (100 kg)	"	"	16, 17
1950	Wiechert (1000 kg)	"	"	17
1951–1962	Wiechert (1000 kg) N-S, E-W	"	"	20, 21, 22, 23, 24, 25
1962–1972	VEGIK Z	L. Egyed	"	25, 26, 27, 28, 29, 30, 31, 32, 33, 34
	Kirnos, N-S, E-W, Z	E. Bisztricsány	"	Modified Kirnos and Kirnos with magnification of 200
1973–1979	Kirnos, N-S, E-W, Z M. Kirnos	E. Bisztricsány	"	34, 35, 36, 37, 38 Modified Kirnos with filter galvanometer
	Ullmann, Teupser (UT)			

Table I. (contd.)

Name and coordinates of the observatory and the period in which the instruments were used	Type of instruments	Head of the Observatory	Seismograms begin stored in	References and remarks
1980	Ullmann, Teupser (Z) Kirnos, N-S, E-W, Z M. Kirnos Z	"	Budapest	39
1981–1990	Kirnos, N-S, E-W, Z M. Kirnos Z	E. Bisztricsány from 1987 Gy. Szeidovitz	"	39
PISZKÉSTETŐ	19°53′40″ E 47°55′10″ N			
1964	M. Kirnos (Z) K(Z)	L. Egyed E. Bisztricsány	Kecskemét?	VEGIK short period seismograph worked 1963–1964 27, 28, 29
1965–1966	Kirnos N-S, E-W, Z M. Kirnos (Z)			
1967–1976	M. Kirnos (Z)	L. Egyed E. Bisztricsány E. Bisztricsány		30, 31, 32, 33, 34, 35 Observatory of Piszkéstető was not included in Bulletin of 1977. From 1977 the seismological data has been transmitted from Piszkéstető to Budapest by radio 36, 37, 38, 39
1977–1990	short period seismograph	Gy. Szeidovitz		
PÉCS	18°15′ EGr 47°06′ N			
1906	Konkoly avisatore	E. Czirer	–	5
KECSKEMÉT	19°41′54″ EGr 46°54′44″ N			
1937–1942	Wiechert vertical (80 kg)	K. Murányi	Kecskemét	13

Table I. (contd.)

Name and coordinates of the observatory and the period in which the instruments were used	Type of instruments	Head of the Observatory	Seismograms begin stored in	References and remarks
1951–1966	Krumbach (100 kg) N-S, S-W		"	equipment Krumbach worked till 1972 20, 21, 22, 23, 24, 25, 26, 27, 28, 29
1967–1971	Kirnos N-S, E-W, Z	K. Murányi	" 1968–1971	Data vailable from time interval between 1967–1971 in bulletin 30, 31, 32, 33
	Quervain-Piccard (25 kg)			
SZEGED	20°8′29″ EGr 46°14′54″ N			
1909–1911	Mainka		–	6, it was not included in Microseismic bulletin
1939	Mainka (135 kg)		Kecskemét	13
1940–1944	Mainka (210 kg)		"	15
1952–1963	" N-S, E-W	B. Simon	"	There are seismograms from 1964 as well 20, 21, 22, 23, 24, 25, 26
KALOCSA	18°48′35″ EGr 46°31′43″ N			
1900?	Brassart vertical	?	–	2/a
1903?–1906	Rossi Avisatore	Gy. Fényi	–	2
1907–1908				It is not involved in bulletin of 1909
1910–1912	Wiechert (200 kg)		–	7
1940–1950	"	B. Simon	Kecskemét	14, 15, 17
1951–1963	Wiechert (200 kg) N-S, E-W	"		20, 21, 22, 23, 24, 25

Table I. (contd.)

Name and coordinates of the observatory and the period in which the instruments were used	Type of instruments	Head of the Observatory	Seismograms begin stored in	References and remarks
JÓSVAFŐ	20°32′21.7″ EGr 48°29′48.8″ N			
1970–1982	M. Kirnos (Z)	E. Bisztricsány	Kecskemét	33, 34, 35, 36, 37, 38, 39
UNGVÁR	(Uzhgorod) 22°14′ EGr 48°37′ N			
1910?	Bosch (10 kg)	Gulovics	–	7
1911–1912	"	Gulovics?	–	7
1940–1942	Wiechert (80 kg)	?	out of order	13, Equipment was put in another place. The new co-ordinates: 22°17′57″ EGr, 48°37′33″ N
KOLOZSVÁR	(Cluj Napoca) 23°22′ EGr 46°46′ N			
1911?	Mainka (210 kg)	M. Cholnoky	–	7
1912	"	"	–	7
1940–1941	?	?	–	14, Equipment was out of order
ÓGYALLA	(Hurbanovo) 18°11′32″ EGr 47°52′24″ N			
1903	Strassburg, further Bosch horizontal		–	2, 3
1904	"		–	3
1905	Bosch, Vicentiny-Konkoly		–	4
1906–1908	"	B. Szabó	–	6
1909–1910	Bosch		–	7

Table I. (contd.)

Name and coordinates of the observatory and the period in which the instruments were used	Type of instruments	Head of the Observatory	Seismograms begin stored in	References and remarks
1911–1912	Mainka (210 kg)		–	7
1938–1943	"		Kecskemét	13, 14, 15, Out of order in 1939
TEMESVÁR	(Timisoara) 21°15′55″ EGr	45°45′32″ N		
1901–1902	Cacciatore Agamennone avisatore	?	–	1/a
1903–1904	Rossi-forel avisatore	E. Berecz	–	3
1905	Vicentini-Konkoly Rosi-Forel avisatore	"	–	4
1906–1909	Vicentini-Konkoly	"	–	5, 6
1909–1912	"	M. Ottilie	–	7
FIUME	(Rijeka) 14°25′40″ EGr	45°19′56″ N		
1903–1904	Vicentini horizontal	P. Salcher	–	1a, 2, 3
1905	Vicentini vertical	"	–	4
1906–1909	Vicentini horizontal	"	–	5, 6, 7
1910–1912	"	M. Schieht	–	7
ZAGRAB	15°58′48″ EGr	45°48′54″ N		
1906–1908	Vicentini-Konkoly	A. Mohorovičić	–	5, 6, Separate bulletin in Croatian and German languages

Station Sopron will be modernized during the year 1994 and a new station will be installed in the SE part of Hungary. Both stations will be equipped with KINEMETRICS SSR-1 seismographs.

The recorded earthquake data are published in Microseismic Bulletins. There is some lag in the bulletin publication (they are not available from 1980 till 1990), but it is tried to make up some arrears of the publications of Microseismic Bulletins. Recently Microseismic Bulletins are available in disk (dBase format) for the years 1969, 1970, 1971, 1977, 1978, 1979, 1991 and 1992.

One of the main research works of the department is the seismic hazard assessment of Hungary. In this framework the following themes have been studied:

1. Revaluation of the parameters of historical earthquakes.
2. Completing of the earthquake catalogue.
3. Development of a new method for the location of near earthquakes by a genetic algorithm.
4. Investigation of geological structures and geoanomalies in the epicentral area of large earthquakes.
5. Calculation of amplification of loose layers.
6. Calculation of synthetic seismograms.

The seismological observation of underground blastings has appeared in the program of the department since several years.

References to Table I

1a. Kövesligethy R 1900: Über einige seismische Institute. Földtani Közlöny, Vol. 30.
1b. Réthly A 1909: Az 1900., 1901. és 1902. évi magyarországi földrengések (Earthquakes in Hungary in the years 1900, 1901 and 1902). Budapest
2a. Réthly A 1906a: Az 1903. évi magyarországi földrengések (Earthquakes in Hungary in the year 1903). Budapest
2b. Réthly A 1903: III. Jelentés a M. Kir. Országos Meteorológiai és Földmágnességi Intézet és az Ógyallai Obszervatórium 1902. évi működéséről (Report on the 1902 activity of the Hungarian Meteorological and Geomagnetic Institute and of the Ógyalla Observatory). Pesti Könyvnyomda RT., Budapest
3. Réthly A 1906b: Az 1904. évi magyarországi földrengések (Earthquakes in Hungary in the year 1904). Budapest
4. Réthly A 1906c: Az 1905. évi magyarországi földrengések (Earthquakes in Hungary in the year 1905). Budapest
5. Réthly A 1907: Az 1906. évi magyarországi földrengések (Earthquakes in Hungary in the year 1906). Budapest
6. Kövesligethy R 1909: Jelentés a Magyar Szent Korona országai területén lévő obszervatóriumokban 1907. és 1908 években végzett megfigyelésekről (Report on the observations made in the years 1907 and 1908 in the observatories of the lands of the Hungarian Holy Crown in years). Budapest

7. Kövesligethy R 1913: Rapport sur les observations faites pendant les années 1909–1912 aux observatories sismologiques de Hongrie. Budapest
8. Kövesligethy R 1920: Rapport sur les observations sismologiques faites pendant les années 1913–1919. Budapest
9. Simon B 1937: Rapport sur les observations sismologiques faites pendant les années 1920–1925 a L'observatoire de Budapest. Budapest
10. Simon B 1926: Rapport sur les observations sismologiques faites a L'observatoire de Budapest. Budapest
11. Kövesligethy R 1927: Rapport sur les observations sismologiques faites a L'observatoire de Budapest. Budapest
12. Szilber M 1928, 1930, 1932, 1934, 1937, 1938, 1939: Rapport sur les observations sismologiques faites a L'observatoire de Budapest. Budapest, 1928, 1929, 1930, 1931, 1932, 1933, 1934, 1935, 1936
13. Szilber M, Petrich G 1937: Rapport microseismique de L'observatoire seismologique de Budapest. Budapest
14. Szilber M 1940, 1941: Rapport microseismique de L'observatoire central seismologique de Hongrie. Budapest
15. Szilber M 1942, 1943, 1945: Rapport microseismique de L'institut national seismologique de Hongrie
16. Szilber M, Szeredai E 1946: Rapport microseismique de L'institut national seismologique de Hongrie
17. Szilber M, Turi St 1947, 1948, 1949, 1950: Rapport microseismique de L'institut national seismologique de Hongrie, Budapest
18. Simon B 1948: A földrengéskutatás története (History of Earthquake Research). Pázmány Péter Tudományegyetem. Az Országos Földrengésvizsgáló Intézet kiadványai
19. Bisztricsány E, Csomor D 1981: 75 Years of seismological research in Hungary. *Acta Geod. Geoph. Mont. Hung.*, 16, 423–431.
20. Szilber M, Tiru St, Csomor D 1952, 1953, 1954: Rapport microseismique de L'institut national seismologique de Hongrie, 1951, 1952, 1953. Nehézipari Könyv- és Folyóiratkiadó V., Budapest
21. Szilber M, Csomor D 1955: Rapport microseismique de L'institut national seismologique de Hongrie, 1954. Műszaki Könyvkiadó, Budapest
22. Szilber M, Csomor D 1957, 1958, 1960, 1961: Rapport microseismique de L'institut national seismologique de Hongrie, 1955, 1956, 1957, 1958. Műszaki Könyvkiadó, Budapest
23. Szilber M, Csomor D, Kiss Z 1962: Rapport microseismique de L'institut national seismologique de Hongrie, 1959. Műszaki Könyvkiadó, Budapest
24. Csomor D 1967: Rapport microseismique de L'institut national seismologique de Hongrie, 1960, 1961. Országos Földrengésvizsgáló Intézet, Budapest
25. Csomor D et al. 1964: Rapport microseismique de L'institut national seismologique de Hongrie, 1962. Országos Földrengésvizsgáló Intézet, Budapest
26. Csomor D 1969: Rapport microseismique de L'institut national seismologique de Hongrie, 1963. Országos Földrengésvizsgáló Intézet, Budapest
27. Csomor D 1968: Rapport microseismique de L'institut national seismologique de Hongrie, 1964. Országos Földrengésvizsgáló Intézet, Budapest
28. Csomor D 1971: Rapport microseismique de L'institut national seismologique de Hongrie, 1965. Országos Földrengésvizsgáló Intézet, Budapest
29. Csomor D 1966: Rapport Microseismique de Hongrie. Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences

30. Kiss Z 1967: Rapport Microseismique de Hongrie. Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences
31. Csomor D, Kiss Z 1968: Rapport Microseismique de Hongrie. Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences
32. Gellén P 1970, 1971: Microseismic bulletin of Hungary, 1969. (Manuscript)
33. Gellén P, Kovács L 1970, 1971: Microseismic bulletin of Hungary. Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences
34. Csomor D 1972, 2973, 1974, 1975: Rapport Microseismique de Hongrie. Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences
35. Kiss Z 1976: Rapport Microseismique de Hongrie. Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences
36. Gellén P 1977: Microseismic bulletin of Hungary. Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences
37. Gellén P, Kovács L 1978: Microseismic bulletin of Hungary. Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences
38. Gellén P, Kovács L 1979: Microseismic bulletin of Hungary. Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences
39. Microseismic bulletin of Hungary 1980–1990. (Manuscript)

References

Csengeri M 1916: Az 1913. március 14-iki sangiri földrengésről. Doktori értekezés (The Sangiri earthquake, March 14, 1913)

Csomor D, Kiss Z 1962: Magyarország szeizmicitása (Seismicity of Hungary). *Geofizikai Közlemények*, 2, No. 1–4, 51–75.

Grossinger J B 1783: Dissertatio de Terrae Motibus Regni Hungariae. Jaurini, Győr

Hantken M 1882: Az 1880. évi zágrábi földrengés 2 térképpel és 6 táblával (The Zágráb earthquake, 1880). A Földtani Intézet Évkönyvei, Budapest, Vol. 6, K. 1.

Holéczy M 1824: 'A Komáromi Földindulások'a XVIII-dik században, és az XIX-diknek huszonkét elsőbb esztendejében (The Komárom earthquakes in the 18th century and in the first 22 years of the 19th century). Tudományos Gyűjtemény, Pest, Vol. 5.

Hunfalvi J 1859: Az 1858. január 15-én történt földrengés (The earthquake on January 15, 1858). Természettudományi Társulat Évkönyvei, Pest, Vol. 6, 1857–1859.

Inkey B 1877: Adatok az 1876. évi Somogy megyében észlelt földrengésekről (Data on the county Somogy earthquakes in 1876). *Földtani Közlöny*, Vol. 7 and *Természettudományi Közlöny*, Vol. 9.

Inkey B 1881: A zágrábi 1880. évi földrengés forgatási tüneményeiről (The rotational effects of the Zágráb earthquake in 1880). *Földtani Közlöny*, Vol. 11, 24–31.

Jeitteles H 1860a: Das Erdbeben am 15. Januar 1858 in den Karpathen und Sudeten in seinen Beziehungen zur Atmosphaere. K. K. Geographische Gesellschaft, Band III.

Jeitteles H 1860b: A földrengések legnevezetesebb kiindulási vagyis középpontjai Magyar és Erdélyországban (térképpel) (Most important epicentres of earthquakes in Hungary and in Transylvania). Természettudományi Közlöny, Pest

Kitaibel P, Tomtsányi Á 1814: Dissertatio de terrae motu in genere ac in specie Mórensi anno 1810. die 14. januário orto

Koch A 1880: A földrengésről (On the earthquake). Kolozsvári Orvos Természettudományi Társulat Értesítője, Kolozsvár, 279–290.

Moravetz K 1925: A folyó évi január 31-i egri földrengés (The Eger earthquake on January 31 in this year). Természettudományi Közlöny, Vol. 57.

Réthly A 1908: Jelentés az 1908. évben Magyarországon észlelt földrengésekről (Report on the earthquakes observed in Hungary in the year 1908). Budapest

Réthly A 1912: Az 1911. évben észlelt földrengések hazánkban (Earthquakes in our country observed in the year 1911). Földtani Közlöny, Vol. 42.

Réthly A 1914: Az 1896–1899 években Magyarországon észlelt földrengések (Earthquakes observed in Hungary in the years 1896–1899). Budapest

Réthly A 1915: Az 1894–1895 években Magyarországon észlelt földrengések (Earthquakes observed in Hungary in the years 1894–1895). Budapest

Réthly A 1952: A Kárpátmedencék földrengései (455–1918) (Earthquakes in the Carpathian Basins (455–1918). Akadémiai Kiadó, Budapest

Saly A 1860: Földrengések Magyar-hazánk határain, különösen városunkban; történeti adatok és kéziratok nyomán. (A Pannonhalmi Szent-Benedekrendiek rév-komáromi algymnáziumának tizedik programja az 1859/60. évben) (Earthquakes within the limits of our country Hungary, mainly in our town Komárom; after historical data and manuscripts), Komárom

Schafarzik F 1880: A földrengések Dél-Magyarországon és a szomszédos területeken, 1879. okt. 10.–1880. márc. 1. (Earthquakes in Southern Hungary and in adjacent areas, between October 10, 1879 and March 1, 1880). *Földtani Közlöny*, 10, 53–75.

Schafarzik F 1889: Az 1885–1886. évi magyarországi földrengésekről (Earthquakes in Hungary in the years 1885 and 1886). *Földtani Közlöny*, 19, 32–44.

Schafarzik F 1901: Az 1901. febr. 16-i észak-bakonyi földrengésről (The Nothern Bakony earthquake, February 16, 1901). *Földtani Közlöny*, 31.

Schréter Z 1925: Az 1925. január egri földrengése (The Eger earthquake, January 1925). *Földtani Közlöny*, 55, Budapest

Simon B 1931: Várpalota és környékének földrengései (Earthquakes in an around Várpalota). *Bányászati és Kohászati Lapok*, 30.

Simon B 1937: A Várpalotán és környékén észlelt földrengések 1038–1930 (Earthquakes observed in and around Várpalota 1038–1930). Budapest

Simon B 1946: Az Országos Földrengésvizsgáló Intézet mikroszeizmikus jelentése 1943–1945 (Microseismic report of the National Seismological Institute, 1943–1945). Budapest

Sternberg J 1786: Versuch einer Geschichte der Ungarischen Erdbeben. Abhandlungen der Böhmischen Gesellschaft der Wissenschaften auf das Jahr 1786. Prag-Dresden

Szilber J 1914: A május 13-i pestmegyei földrengés (The county Pest earthquake, May 13). *Természettudományi Közlöny*, 46.

Zsíros T, Mónus P, Tóth L 1988: Hungarian Earthquake-Catalog. Budapest

Acta Geod. Geoph. Hung., Vol. 29(1–2), pp. 209–220 (1994)

THREE-DIMENSIONAL FORMS OF THERMAL CONVECTION IN POROUS LAYERS

Z BUS[1] and L CSEREPES[1]

[Manuscript received June 10, 1994]

Three-dimensional planforms of thermal convection in porous media are studied by solving the equations of hydrodynamics numerically. Rectangular cells of a homogeneous plane layer constitute the modelling domain where a combined spectral-finite-difference algorithm is used to calculate the temporal evolution of the temperature and velocity fields. The Rayleigh number is in the range 40–240; further variables are the cell size and the initial and boundary conditions. When the upper and lower boundaries are impermeable and isothermal, the so-called [111] mode symmetric solution (Beck 1972, Steen 1983) is reproduced. With permeable top or iso-flux bottom, new asymmetric forms are found. The highest degree of asymmetry is obtained with permeable top which can be a model of an unconfined phreatic aquifer. Heat transfer efficacy is described in terms of the Nusselt number for all the observed planforms.

Keywords: asymmetry; cell structure; porous media; thermal convection

1. Introduction

Hydrothermal convection driven by buoyancy and gravitational instability can occur in high heat flow terrains (e.g. volcanic areas, for a review see Stefánsson and Björnsson 1982) or in highly permeable thick sediments (e.g. in the Hungarian Plain, see Lenkey 1993). The characteristics of hydrothermal flow have been investigated by numerical methods for almost twenty years now, solving the equations of motion typically in two-dimensional sections taken as models of the porous medium (e.g. Ribando and Torrance 1976, Ribando et al. 1976, Straus and Schubert 1977, McKibbin and Tyvand 1982, Schubert and Straus 1982, Rosenberg and Spera 1990).

Three-dimensional (3D) studies revealing the true spatial structure of porous convection have been much less numerous. In the linear approximation of the theory, Beck (1972) has shown that convection in a porous cube with isothermal top and bottom and insulated vertical sides begins in the form of two-dimensional rolls when the Rayleigh number R exceeds $R_{\text{crit}} = 4\pi^2 \approx 39.5$, and real three-dimensional modes exist when $R > 4.5\pi^2$. Carrying out an eigenfunction-expansion stability analysis, Steen (1983) has proved that the so-called [111] mode becomes stable for $R > 4.87\pi^2 = 48.06$. Previous and subsequent three-dimensional numerical studies concentrated almost exclusively on this particular 3D flow. Holst and Aziz (1972), probably in a first attempt to calculate 3D circulations, obtained results for $R = 60$

[1] Geophysical Department of Eötvös University, H-1083 Budapest, Ludovika tér 2, Hungary

1217-8977/94/$ 4.00

and 120. Straus and Schubert (1979) observed that the two-dimensional rolls carry more heat than the 3D flow until $R < 97$, while the opposite is true for $R > 97$. The same authors found (Schubert and Straus 1979) that steady 3D convection in a cube is replaced by time-dependent oscillatory flow somewhere in the range $R = 300 - 320$. Extending the calculations to higher values of R and improving the numerical accuracy, Kimura et al. (1989), Stamps et al. (1990) and Graham and Steen (1991) located the transition of the [111] mode from the steady to oscillatory regime at $R \approx 580$.

Stamps et al. (1990) gave examples for a porous cube with heat-conducting lateral sides where the flow pattern may be distinctly different from the well-studied [111] mode.

The present paper focuses on the convective patterns occurring in more general circumstances. Numerical experiments have been carried out for unit cells of convection appearing in infinite horizontal layers. The applied boundary conditions on the top and bottom covered more general cases than in any previous study. The porous medium have been assumed to be homogeneous and isotropic, and the Rayleigh number has been varied between 40 and 240 typically. Before discussing the results, a short description of the mathematical methods will be given in the next two sections.

2. The equations to be solved

The behaviour of a convective fluid saturating a porous rock is described by the equation of continuity, Darcy's law and the heat transport equation:

$$\operatorname{div} \mathbf{u} = 0 \,, \tag{1}$$

$$\varrho_0(1 - \alpha(T - T_0))g\mathbf{e} - \operatorname{grad} p - \frac{\eta}{k}\mathbf{u} = 0 \,, \tag{2}$$

$$\varrho_m c_m \frac{\partial T}{\partial t} + \varrho_0 c_0 \mathbf{u} \operatorname{grad} T = \lambda_m \nabla^2 T \,, \tag{3}$$

where $\mathbf{u}$ is the Darcy velocity (filtration flux), ϱ_0 and ϱ_m are the densities of water and the saturated medium, respectively, α is the thermal expansivity of water, T is the temperature and T_0 its reference value, g the gravitational acceleration, $\mathbf{e}$ is the unit vector directed downwards, p is the pressure, η is the viscosity of water, k is the permeability of the rock matrix, c_0 and c_m are the specific heat coefficients for water and the medium, respectively, λ_m is the heat conductivity of the medium. When setting up the equations, it is assumed that the fluid and the porous matrix are in a common thermal equilibrium and that the Boussinesq approximation is valid. This latter states that the fluid density can be kept constant everywhere except in the buoyancy term (first term in Eq. (2)) where its temperature dependence — source of the convective instability — cannot of course be neglected. Since, consequently, the divergence of the flow field $\mathbf{u}$ is zero (see Eq. (1)), $\mathbf{u}$ can be expressed as the curl of a vector potential. On the other hand, the curl of $\mathbf{u}$ ("vorticity") has a vanishing vertical component which follows directly from Eq. (2). This means that the vector

potential is purely poloidal, i.e. it is deduced form a single scalar V. Translating this fact to the Darcy velocity, we have

$$\mathbf{u} = \nabla \times \nabla \times V\mathbf{e}$$

or for the components $(u,\ v,\ w)$ of $\mathbf{u}$:

$$u = \frac{\partial^2 V}{\partial x\, \partial z}, \quad v = \frac{\partial^2 V}{\partial y\, \partial z}, \quad w = -\left(\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2}\right), \tag{4}$$

and Darcy's law reduces to

$$\nabla^2 V = \frac{k\alpha\varrho_0 g}{\eta} T\,. \tag{5}$$

Turning now to the non-dimensional form of the equations, let us define the following scale units: L for length (L = depth of the convective layer), L^2/χ for time ($\chi = \lambda_m/\varrho_0 c_0$ = heat diffusivity), ∇T for temperature (∇T = temperature drop across the layer, to be precised later). Then Eq. (5) and the heat transport equation become

$$\nabla^2 V = RT\,, \tag{6}$$

$$\gamma\frac{\partial T}{\partial t} + \mathbf{u}\,\mathrm{grad}\,T = \nabla^2 T \tag{7}$$

with

$$R = \frac{k\alpha\varrho_0 g\Delta T L}{\chi\eta}, \quad \gamma = \frac{\varrho_m c_m}{\varrho_0 c_0}\,.$$

The non-dimensional form of the components of $\mathbf{u}$ is the same as in Eq. (4). The qualitative and quantitative features of convection is governed by the Rayleigh number R. It has to exceed a first critical value R_{crit} in order to give non-vanishing thermal convection. This R_{crit} depends on the boundary conditions and can be determined by linearizing the equations.

In the present study, the top (cool) boundary of the convective layer is kept at a constant temperature. The bottom is either isothermal (warm) or it is the incoming heat flux q which is prescribed (and taken as constant). In the first case having two isothermal boundaries, the scale unit ΔT is the temperature difference assumed between bottom and top. In the second case ΔT is defined with the heat flux q as $\Delta T = qL/\lambda_m$.

The kinematic boundary conditions on the horizontal boundaries may be the following: on an impermeable horizontal surface $w = 0$ and $V = 0$ (Dirichlet condition for V); on a permeable boundary $\partial p/\partial x = \partial p/\partial y = 0$ is assumed, which translates to $\partial V/\partial z = 0$ (Neumann condition for V).

3. The numerical method

The 3D models to be presented have been obtained by a "hybrid" spectral+finite-difference solution of Eqs (4), (6) and (7). The method is basically a finite-difference algorithm but spectral decomposition is used in x and y to solve Eq. (6) (Cserepes

et al. 1988). The solution is calculated for a rectangular model box. Its non-dimensional depth is 1, the lengths of the horizontal sides are D_x and D_y. This box can be regarded a closed cell of the flow, its sides being symmetry planes. The full solution for the infinite layer is obtained by repeating the cell periodically in each horizontal direction. In terms of the temperature field, symmetry on the sides is equivalent to insulation — the usual boundary condition of previous studies about the [111] mode circulation.

The numerical solution is represented at the nodes (x_j, y_k, z_ℓ) of a uniform rectangular mesh. The horizontal Fourier-decomposition of V and T is written as

$$\left.\begin{aligned} V(x_j, y_k, z) &= \sum_{n,m} \tilde{V}_{nm}(z) e^{i\alpha_n x_j} e^{i\beta_m y_k} , \\ T(x_j, y_k, z) &= \sum_{n,m} \tilde{T}_{nm}(z) e^{i\alpha_n x_j} e^{i\beta_m y_k} , \end{aligned}\right\} \tag{8}$$

where α_n and β_m are discrete sets of wavenumbers:

$$\alpha_n = \frac{\pi n}{D_x}, \qquad \beta_m = \frac{\pi m}{D_y} .$$

Substitution of Eq. (8) into Eq. (6) yields ordinary differential equations for $\tilde{V}_{nm}$:

$$\left[\frac{d^2}{dz^2} - (\alpha_n^2 + \beta_m^2)\right] \tilde{V}_{nm}(z) = R\tilde{T}_{nm}(z) .$$

This is solved for every pair of n and m, then V is obtained as Fourier transform according to Eq. (8), and the velocities are calculated as numerical derivatives according to Eq. (4). At this stage, the T field is advanced in time by Eq. (7), using a finite time step Δt. This is done by the alternating direction implicite (ADI) method (see e.g. Peaceman 1977):

$$\left.\begin{aligned} \left(\delta_x - \tfrac{2\gamma}{\Delta t_n}\right) T^{(n+1/3)} &= -\left(\delta_x + 2\delta_y + 2\delta_z + \tfrac{2\gamma}{\Delta t_n}\right) T^{(n)} - 2H \\ \left(\delta_y - \tfrac{2\gamma}{\Delta t_n}\right) T^{(n+2/3)} &= \delta_y T^{(n)} - \tfrac{2\gamma}{\Delta t_n} T^{(n+1/3)} \\ \left(\delta_z - \tfrac{2\gamma}{\Delta t_n}\right) T^{(n+1)} &= \delta_z T^{(n)} - \tfrac{2\gamma}{\Delta t_n} T^{(n+2/3)} \end{aligned}\right\}$$

where

$$\begin{aligned} \delta_x &= \frac{\partial^2}{\partial x^2} - u\frac{\partial}{\partial x} , \\ \delta_y &= \frac{\partial^2}{\partial y^2} - v\frac{\partial}{\partial y} , \\ \delta_z &= \frac{\partial^2}{\partial z^2} - w\frac{\partial}{\partial z} , \end{aligned}$$

and $\Delta t = \Delta t_n$ is the n-th time step (it is variable), $T^{(n)}$ and $T^{(n+1)}$ are two subsequent stages of the evolution of the temperature field, $T^{(n+1/3)}$ and $T^{(n+2/3)}$ are auxiliary functions needed by the ADI method. The finite-difference equivalents of δ_x, δ_y and δ_z are defined with centred differencing (Cserepes 1985). After advancing T in time, we return to updating V and $\mathbf{u}$, and this cycle is repeated until a steady or quasi-stationary state is reached.

4. Results

4.1 Short summary of the numerical experiments

Several series of model computations have been carried out for the range $R = 40 - 240$ by varying the boundary and initial conditions as well as the box size $(D_x,\ D_y)$.

The boundary conditions for the top and bottom surface have been prescribed in the following four combinations: *i*mpermeable top with iso*t*hermal bottom (case IT), *p*ermeable top with iso*t*hermal bottom (case PT), *i*mpermeable top with iso-*f*lux bottom (case IF) and *p*ermeable top with iso-*f*lux bottom (case PF). Iso-flux means given constant heat flux at the bottom. Otherwise the top is isothermal and the bottom is impermeable. All these conditions are usual idealizations of the complex geophysical reality. For example impermeable top (together with the impermeable bottom) means that the porous layer is a confined reservoir; permeable top means that it is bordered by a free water table from above.

Typical initial conditions were

$$T(x, y, z) = z + A \sin \pi z \cos \frac{2\pi x}{\lambda} \cos \frac{2\pi y}{\lambda} \tag{9a}$$

or

$$T(x, y, z) = z + A \sin \pi z \left(\cos \frac{2\pi x}{\lambda} + \cos \frac{2\pi y}{\lambda} \right) , \tag{9b}$$

where $T(x, y, z) = z$ corresponds to the non-convective equilibrium state, A and λ are the amplitude and wavelength of the initial perturbation, respectively.

The basic structure of the resulting 3D patterns have been found not to depend essentially on the form of initial conditions and box size. The solutions to be presented have been obtained in a box with $D_x = D_y = 3$ and a mesh of $48 \times 48 \times 16$ points. There is a slight dependence on the wavelength λ of the starting perturbation (see later in Section 4.3), but the really important factor in determining the cell structure is the choice of boundary conditions. What is found is that the solutions fall into two different categories according to the symmetry or asymmetry of the boundary conditions. Case IT is symmetrical with respect to the $z = 0.5$ plane since the type of boundary conditions is the same on the top and the bottom. This leads to solutions with remarkable symmetries. All the other cases (PT, IF, PF) represent asymmetrical conditions and result in the asymmetry of the flow patterns. What follows is the description of these two classes of solutions.

4.2 Symmetric solutions

Case IT (Fig. 1) produces the well-studied [111] mode symmetric solution (Beck 1972, Steen 1983, Kimura et al. 1989, Graham and Steen 1991).

Figure 1a shows horizontal sections of the model box for $R = 60$: this is approximately 1.5 times R_{crit} ($R_{\text{crit}} = 4\pi^2$ in case IT). The contours are isolines of the vertical velocity w. The symmetry is as follows: the pattern at a distance ℓ upwards from the midplane (i.e. at $z = 0.5 - \ell$) is the same as the pattern at the

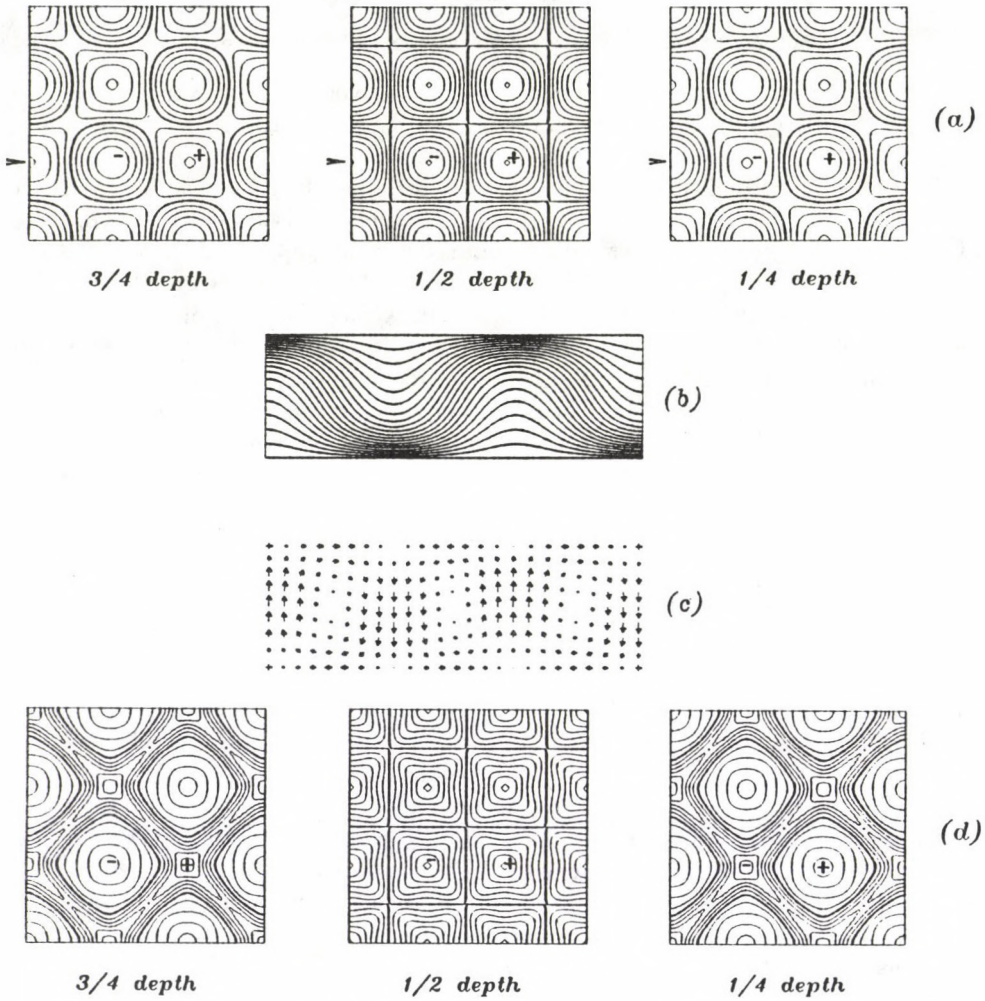

Fig. 1. Convection models for case IT. In (a), (b), (c): $R = 60$. (a) Horizontal sections at depths $z = 0.75$, 0.5 and 0.25 with contours of w. (b) Vertical cross section for the temperature in the plane indicated by arrow heads in (a). (c) Velocity distribution in the same vertical plane. (d) Model for $R = 220$: horizontal sections for w. Contours are equi-spaced in every diagram. + and − indicate warm upwellings and cold downwellings, respectively, while the thick contours mark $w = 0$ in the diagrams (a) and (d)

same distance downwards (at $z = 0.5 + \ell$) but it is shifted by a half cell width in both horizontal directions. The cell width or the basic wavelength of the solution is $\lambda = 2$ in the present case. This symmetry is valid not only for w but for any other physical quantity. The thick contour ($w = 0$) separates downwelling and upwelling domains in each section. In the lower half of the layer the cold descending flow is located in isolated, roughly circular areas. The opposite is true for the upper half where the ascending currents are isolated. In the midplane $z = 0.5$ the oppositely

directed vertical flows form identical squares interconnected at the square vertices (chequerboard pattern). Typical vertical cross-sections for the temperature and velocity are shown in Figs 1b and 1c.

The basic symmetry of the circulation is the same at higher Rayleigh numbers ($R = 220$ in Fig. 1d), but the pattern shows sharp geometrical features away from the midplane. In the lower half of the layer upwelling is concentrated in "ridges" along lines of a uniform square grid inclined at 45° to the box walls: these ridges separate the isolated downwelling spots. The structure of the upper half is similar but opposite in sense.

4.3 Asymmetric solutions

A typical case producing asymmetric solutions is PT, i.e. a layer with permeable top and impermeable bottom (both isothermal). The consequences of the change in the mechanical character of the top boundary with respect to the previous IT case can be studied in Fig. 2. The critical Rayleigh number for PT is $R_{\text{crit}} = 27.1$ (Ribando and Torrance 1976). At low R ($R = 60$ in Figs 2a-c) the solution is not much different from case IT, although the symmetry is lost, e.g. the downwelling and upwelling currents do not occupy equal squares in the midplane any more. At higher Rayleigh numbers ($R = 220$ in Fig. 2d) the asymmetry is obvious: downwelling and upwelling are different structures throughout the layer. The same square grid of warm (upwelling) ridges is observed as in Fig. 1d, but this feature is topologically the same in the lower and upper part of the fluid (i.e. the change in sign at $z = 0.5$, obvious in Fig. 1d, is absent now). With increasing R, the cold center of the oblique squares of Fig. 2d grows gradually at the expense of the warm ascending flow which, in turn, becomes more and more narrow.

The structure of the flow is much the same in cases IF and PF as in the above described case PT. All these 3 cases are inherently asymmetric. However, the change of the upper boundary condition from impermeable to permeable introduces a higher degree of asymmetry than the change of the bottom conditions from isothermal to iso-flux.

In the asymmetric models it is found that the temporal behaviour of the solutions may depend on the wavelength λ of the initial perturbation. When $\lambda = 2$ either in Eq. (9a) or in Eq. (9b), the flow converges quickly to the final steady state shown e.g. in Fig. 2. This steady state is cellular (strictly periodic in space). When we applied $\lambda = 1$ or $\lambda = 3$, some cases, especially at higher Rayleigh numbers ($R > 100$), did not end in cellular steady states but showed non-damping irregular oscillations with continuously changing planforms. An example is given in Fig. 3 (case PT). The same effect of the choice of initial wavelengths is known from studies of Newtonian convection (Cserepes 1993). This shows that the flow pattern cannot easily find the stable steady state if the starting conditions are far from it. This is all the more difficult that the finite size of the box of the numerical experiments imposes artificial symmetries and artificial pattern selection on the circulation — which would be absent in an infinite layer. The qualitative character of the irregular patterns of Figs 3a-d is, however, the same as that found in Fig. 2d: large cold

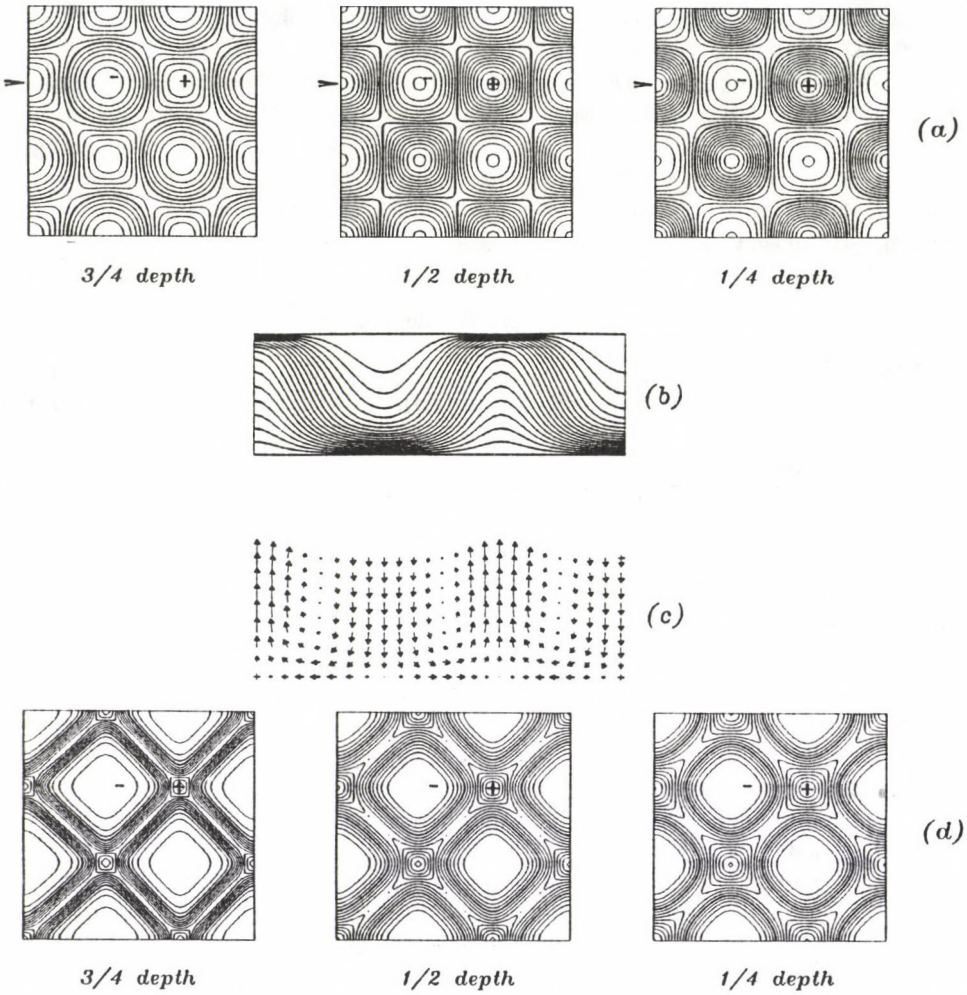

Fig. 2. Convection models for case PT. In a, b, c: $R = 60$. a) Horizontal sections with contours of w. b) Vertical cross section for the temperature in the plane indicated by arrow heads in (a). c) Velocity distribution in the same vertical plane. d) Model for $R = 220$: horizontal sections for w. For the notation see Fig. 1

descending parts are surrounded by "ridges" of the ascending current. Figure 3e) shows that the time-dependence of the example case is chaotic.

4.4 Heat transfer

Efficacy of the heat transfer due to convection is measured conventionally in terms of the Nusselt number Nu. For the cases with isothermal bottom boundary,

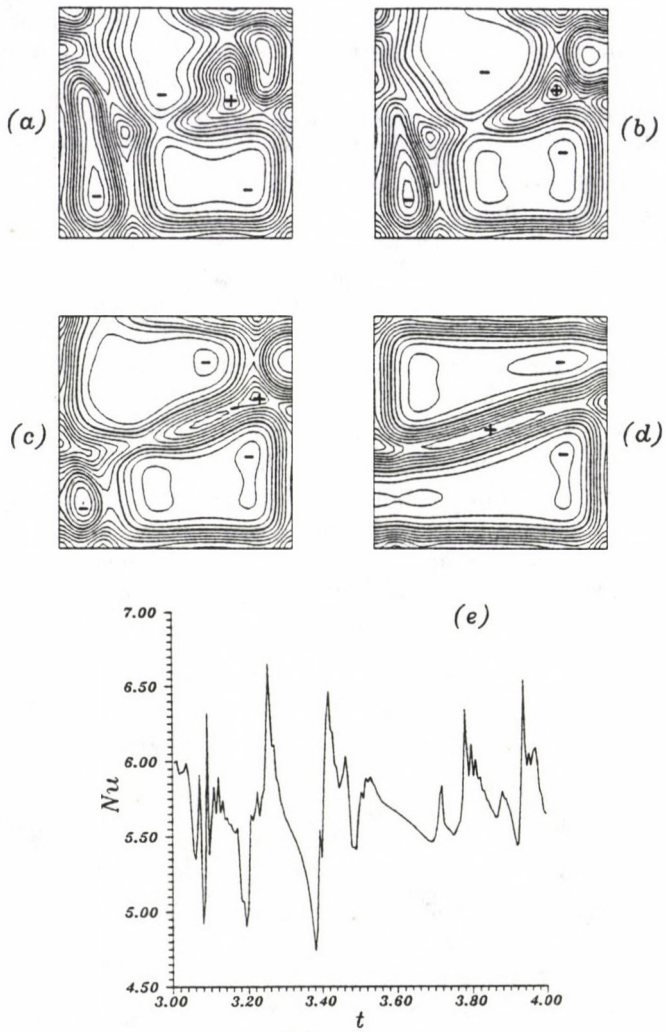

Fig. 3. Temporal evolution of a model at $R = 180$, case PT, started with $\lambda = 3$. Four stages: a) $t = 3.412$, b) $t = 3.440$, c) $t = 3.470$, d) $t = 3.559$. Diagram (e) shows the variation of the Nusselt number between $t = 3$ and 4

Nu is defined as the average non-dimensional heat flux at the surface:

$$Nu = < \partial T / \partial z >_{\text{top}} \ .$$

When, instead of a constant temperature, it is the input heat flux that is prescribed at the bottom, the above Nu would always give 1. In this case, the scaling is different and the Nusselt number can be defined with the average bottom temperature as

$$Nu = \frac{1}{< T >_{\text{bottom}}} \ .$$

Our results for Nu are displayed in Fig. 4 in a Nu *versus* R diagram. On the horizontal axis, R is normalized with R_{crit}. The curves should start with $Nu = 1$ at $R = R_{\text{crit}}$ when the conductive equilibrium ceases to be stable. Beyond that point, different curves exist for different cases. If stable cellular patterns can be found with different λ basic wavelengths, then Nu depends slightly on λ. A more important point is that the increase of Nu with R depends on the boundary conditions. The curves of Fig. 4 form two groups: the upper group corresponds to the cases IT and PT having an isothermal bottom, while the lower curves give Nu for the cases IF and PF when the input flux is given. These two groups cannot be compared because different scaling is applied to them: the definitions of both Nu and R are different. However, there is no scaling difference between IT and PT, neither between IF and PF. It is interesting to see that — at least in the range of the studied Rayleigh numbers — the permeable or impermeable nature of the top surface does not influence too much the heat transfer. Only in the case PT can we observe that the increase of Nu with R is faster than in case IT. This can be explained by the fact that a mechanically "open" surface, i.e. the permeable top allows easier heat transport across than an impermeable boundary. What is a bit surprising is that this is not observed in the relation of the other two cases IF and PF. This shows that the equalizing effect of the uniform bottom heat flux compensates the differences arising from the two opposing permeability conditions applied on the top.

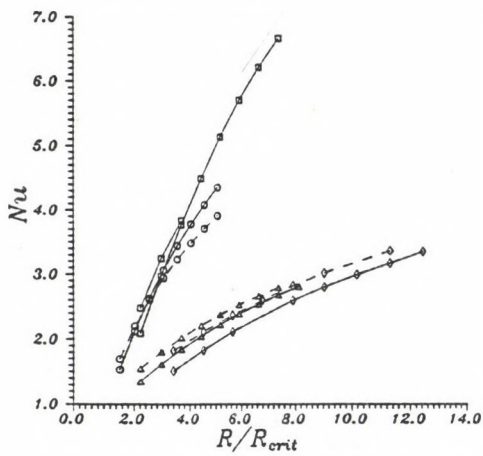

Fig. 4. Variation of the Nusselt number with R. Circles: case IT; squares: case PT; triangles: case IF; diamonds: case PF. Solid lines: models with $\lambda = 2$; dashed lines: models with $\lambda = 3$

5. Conclusions

Pattern selection of three-dimensional convection in idealized circumstances, e.g. in infinite layers, is an important model problem in studying the behaviour of dynamical systems. Numerical investigation of the convective flow in Newtonian fluids (Busse and Frick 1985, Cserepes 1993) have already revealed that there exist two

main classes of flow patterns depending on the vertical symmetry or asymmetry of the fluid layer with respect to its midplane. Symmetric conditions result in a mirror symmetry of the ascending and descending currents, while the lack of symmetry produces differently shaped upwellings and downwellings.

The problem of convection in a porous layer with isothermal and impermeable boundaries — example of symmetric external conditions — have been thoroughly analyzed (Beck 1972, Steen 1983, Kimura et al. 1989, Stamps et al. 1990, Graham and Steen 1991): the resulting [111] mode circulation exhibits symmetric structures in the ascending and descending parts of the flow. The present study shows that the asymmetry introduced by a permeable surface or an iso-flux bottom breaks down the geometrical similarity of upwellings and downwellings, shaping them differently. This is exactly what could be expected by analogy from models of the Newtonian convection mentioned above.

A natural hydrothermal circulation occurring in porous layers will probably never be symmetrical in the exact mathematical sense. Beyond this evidence, the most important hydrogeological situation where free thermal convection can be detected under suitable conditions is the case of an unconfined phreatic aquifer bordered from above by the water table open to the atmosphere and by an impermeable boundary somewhere in the depth. This is equivalent to our model case PT or PF — both of which are highly asymmetric. The symmetric case IT could be realized in a confined aquifer enclosed by impermeable layers. This means necessarily that the confined aquifer is buried at some depth, and if convection occurs in it, its manifestation will be concealed or attenuated by the covering strata (Lenkey 1993). Consequently, free hydrothermal convection found and observed in nature will most probably occur in an unconfined phreatic aquifer — i.e. in asymmetric physical circumstances.

In either case, the detection of natural porous convection relies on near-surface geothermal measurements. It is interesting to note the difference in the near-surface temperature distribution when the symmetric IT and the asymmetric PT, IF, PF cases are compared. This difference can be deduced from the contours of vertical velocity in a horizontal plane in the upper half of the convective layer (Figs. 1d and 2d): the isotherms give practically the same pattern. The near-surface manifestation of the symmetric flow (Fig. 1d) is a wide, more or less circular warm spot surrounded by narrow cold stripes along the sides of a square. The asymmetric models (e.g. Fig. 2d) shows just the opposite: the warm spot is smaller and it is connected by four "spokes" to neighbouring warm spots, while the "spokes" enclose circular cold spots. This difference of the two model classes is not significant at low Rayleigh numbers (near R_{crit}) but becomes more and more pronounced as R increases.

Acknowledgement

The present study has been supported by the OTKA grant No. 2134.

References

Beck J L 1972: *Phys. Fluids*, 15, 1377–1383.
Busse F H, Frick H 1985: *J. Fluid Mech.*, 150, 451–465.
Cserepes L 1985: *Annales Univ. Sci. Budapest., Sectio Geophys. Meteor.*, I–II, 52–67.
Cserepes L, 1993: *Geophys. Res. Lett.*, 19, 2091–2094.
Cserepes L, Rabinowicz M, Rosemberg-Borot C 1988: *J. Geophys. Res.*, 93, 12009–12025.
Graham M D, Steen P H 1991: *J. Fluid Mech.*, 232, 591–609.
Holst P H, Aziz K 1972: *J. Heat Mass Transfer*, 15, 73–90.
Kimura S, Schubert G, Straus J M 1989: *J. Fluid Mech.*, 207, 153–189.
Lenkey L 1993: *Magyar Geofizika*, 34, 30–45.
McKibbin R, Tyvand P A 1982: *J. Fluid Mech.*, 118, 315–339.
Peaceman D W 1977: Fundamentals of numerical reservoir simulation. Elsevier, Amsterdam
Ribando R J, Torrance K E 1976: *J. Heat Transfer*, 95, 42–48.
Ribando R J, Torrance K E, Turcotte D L 1976: *J. Geophys. Res.*, 81, 3007–3012.
Rosenberg N D, Spera F J 1990: *Geophys. Res. Lett.*, 17, 235–238.
Schubert G, Straus J M 1979: *J. Fluid Mech.*, 94, 25–38.
Schubert G, Straus J M 1982: *J. Fluid Mech.*, 121, 301–313.
Stamps D W, Arpaci V S, Clark J A 1990: *J. Fluid Mech.*, 213, 377–396.
Steen P H 1983: *J. Fluid Mech.*, 136, 219–241.
Stefánsson V, Björnsson S 1982: In: Continental and oceanic rifts, G Pálmason (ed.), Geodynamics Series, 8, 123–145.
Straus J M, Schubert G 1977: *J. Geophys. Res.*, 82, 325–333.
Straus J M, Schubert G 1979: *Fluid Mech.*, 91, 155–165.

Acta Geod. Geoph. Hung., Vol. 29(1–2), pp. 221–234 (1994)

VERTICAL MAGNETIC DIPOLE OVER VERTICALLY INHOMOGENEOUS EARTH WITH EXPONENTIAL VARIATION OF CONDUCTIVITY

N P SINGH[1], T LAL[1], A HAKIM[2]

[Manuscript received June 3, 1994]

The electromagnetic (EM) field response of vertical magnetic dipole placed over the surface of a vertically inhomogeneous earth model is analysed. The problem is formulated for the 1-D multilayered inhomogeneous earth model with one of the layers having exponential variation of conductivity with depth. Analytical solutions and numerical computations have been performed for the three layered transitional earth model with the intermediate inhomogeneous layer possessing above conductivity variation. Computations are performed for the absolute-amplitude ratio of various field components for the case of quasi-static approximation and as such results, showing the effects of variation of transition layer thickness and conductivity contrast between the top and bottom layers are presented in the form of absolute-amplitude ratio of various field components expressed as function of numerical distance. The analysis of the results reveal the characteristics dependence of the EM field response on conductivity inhomogeneity.

Keywords: absolute-amplitude ratio; electromagnetic response; numerical distance; quasistatic zone; time dependence of source; vertically inhomogeneous earth model

Introduction

Reliability and accuracy of the quantitative interpretation of electromagnetic data depends upon the electrical properties of the subsurface materials and conversely the electrical properties (conductivity and dielectric constants) of the subsurface materials govern the electromagnetic response. The EM response is differential in extended frequency range. The high frequency EM response is controlled by dielectric constant, whereas the low frequency EM response, pertaining to geophysical exploration work is controlled by mainly conductivity parameter.

The general trend in electromagnetic interpretation is to consider the earth models to be made of a number of stratified homogeneous layers. Though this assumption of homogeneity of layers is common and expedient, yet it does not represent the real subsurface structures. There are a number of cases, where it has been observed that in certain zones in the earth, conductivity does not changes abruptly from one set of values to the others in accordance with the assumption of homogeneity of layers, rather it varies continuously on local as well as global scale

[1]Department of Geophysics, Banaras Hindu University, Varanasi-221005, India
[2]Geological Survey of India, Nagpur, India

1217-8977/94/$ 4.00

(Mallick and Roy 1968, Keller 1971). Thus, it is felt logical to study the effect of conductivity inhomogeneity on EM response of a vertical magnetic dipole, which is the best EM source suitable for detecting relatively good conductors.

Mallick and Roy (1971) have derived the analytical solutions for the EM field components of vertical magnetic dipole over an inhomogeneous earth model, considering linear variation of conductivity with depth, while Abramovici and Chlamtac (1978) have presented the computational results for the above case for the various electric and magnetic field components. Further, Chlamtac and Abramovici (1981) presented the EM response of horizontal electric dipole over inhomogeneous earth model possessing linear variation of conductivity with depth. Singh and Lal (1994a, b) have presented the EM field response of horizontal magnetic dipole over vertically inhomogeneous earth models, considering linear and exponential variation of conductivity with depth.

In this paper, the problem of calculating EM field response of the vertical magnetic dipole placed over the surface the inhomogeneous earth model is presented for the multi-layered earth model with one of the layer having exponential variation of conductivity with depth. Analytical solutions and computational results are presented for the three layer transitional earth models with intermediate transition layer possessing exponential variation of conductivity with depth.

1. Specification of the problem

The geometry of 1D multilayered earth model under consideration is shown in Fig. 1. It comprises of a sequence of N-layers, in which $N-1$ are homogeneous layers and the remaining one j^{th} is inhomogeneous. A cylindrical co-ordinate system (ϱ, ϕ, z) with z-axis directed vertically upward is used. A vertical magnetic dipole or its equivalent, a current carrying horizontal circular loop with its axis in z-direction is placed at a height 'h' above the model.

Let σ_j, μ_j, h_j $(j = 1, 2, \ldots \mathrm{N})$ be the conductivities, permeabilities and depth to the lower boundaries the layers. The conductivity and permeability of the free space are assumed to be σ_o and μ_o respectively. Further, as the magnetic permeability values for the rock formations in general equals the free space value, it is assumed that for all the layers μ_j $(j = 1, 2, \ldots \mathrm{N}) = \mu_o$. The conductivity σ_j in the jth inhomogeneous layer, which acts as a transition zone, is assumed to follow the exponential variation with depth in accordance with the relation;

$$\sigma_j(z) = \sigma_{j-1} \exp[\alpha_2(z + h_{j-1})] \tag{1}$$

where the conductivity of the $(j-1)^{\text{th}}$ layer gradually merges with that of the $(j+1)^{\text{th}}$ layer through the jth transition layer so that

$$\sigma_j(z) = \sigma_{j-1} \quad \text{at} \quad z = -h_{j-1} \qquad \text{and} \qquad \sigma_j(z) = \sigma_{j+1} \quad \text{at} \quad z = -h_j \ ,$$

where α_2 is the constant dependent upon the model parameters.

The radiation constants γ_o in the air and $\gamma_1, \gamma_2, \ldots \gamma_{j-1}, \gamma_{j+1} \ldots \gamma_N$ in the homogeneous layers assume constant values while in the transition layer $\gamma_j(z)$ varies

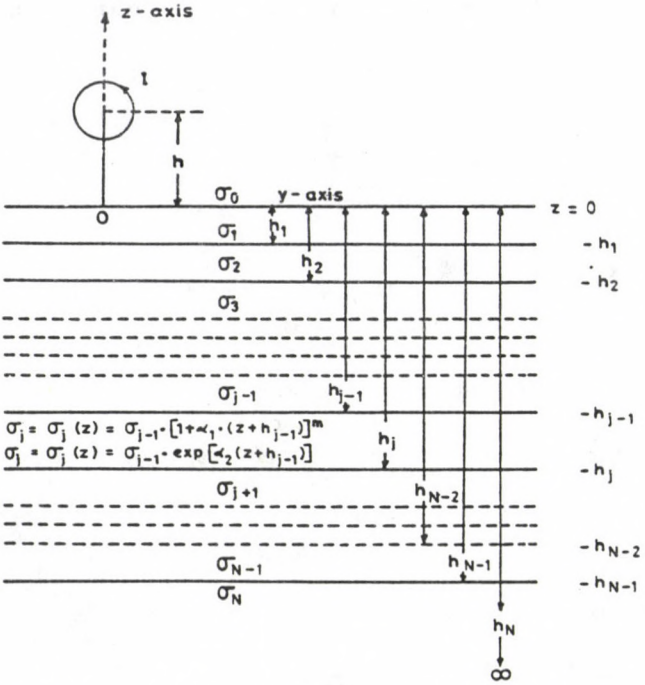

Fig. 1. Geometry of 1-D multi-layered earth model under investigation

following the exponential variation with depth in correspondence with conductivity variation. Therefore, in terms of radiation constants, above relation can be expressed as

$$\begin{aligned} \gamma_j^2(z) &= \gamma_{j-1}^2 \exp[\alpha_2 \cdot (z + h_{j-1})] \\ &= \gamma_{j-1}^2 \qquad \text{at } z = -h_{j-1} \\ &= \gamma_{j+1}^2 \qquad \text{at } z = -h_j \,, \end{aligned} \tag{2}$$

where $\alpha_2 = \dfrac{1}{(h_{j-1} - h_j)} \cdot \ln\left(\dfrac{\gamma_{j+1}^2}{\gamma_{j-1}^2}\right)$ and $\gamma_j^2 = i\omega\mu_j\sigma_j$.

The time dependence of the source is assumed to be of the form $e^{i\omega t}$ which is implied. The conduction currents are assumed to dominate over the displacement currents in all the conductive media. Further, we have restricted ourselves to the discussion of fields at distances within the quasi-static zone, therefore, the displacement current in air and all the conductive media have been neglected.

Let dA be the area of the loop representing the vertical magnetic dipole and I be the low frequency alternating current circulating the loop. The moment M of the dipole is given by

$$M = \frac{I \cdot dA}{4\pi} \,. \tag{3}$$

The primary Hertz-vector, which has only z-component in this case, is written as

$$\pi_p = M \cdot \frac{e^{-\gamma_o r}}{r} , \tag{4}$$

where $\gamma_o = \sqrt{i\omega\mu_o\sigma_o}$ and $r = \sqrt{\varrho^2 + |z-h|^2}$.

Using Sommerfeld's (1949) relations, one can write

$$\pi_p = M \cdot \frac{e^{-\gamma_o r}}{r} = M \int_0^\infty \frac{\lambda}{n_o} \cdot e^{-n_o|z-h|} \cdot J_o(\lambda\varrho) d\lambda , \tag{5}$$

where λ is separation constant and $n_o = \sqrt{\lambda^2 + \gamma_0^2}$.

The harmonic electric and magnetic vector fields E and H may be derived from the Hertz vector $\boldsymbol{\pi}$, which possess only z-component, using the relations

$$E = -i\omega\mu \text{ curl } \pi \tag{6}$$

$$H = -i\omega\mu\sigma\pi + \text{grad}(\text{div } \pi) . \tag{7}$$

The components of the fields can be written directly as

$$\begin{aligned} E_\varrho &= 0; & H_\varrho &= \tfrac{\partial^2\pi}{\partial\varrho\cdot\partial z} \\ E\phi &= i\omega\mu\tfrac{\partial\pi}{\partial\varrho}; & H\phi &= 0 \\ Ez &= 0; & Hz &= -\tfrac{1}{\varrho}\cdot\tfrac{\partial}{\partial\varrho}\left(\varrho\cdot\tfrac{\partial\pi}{\partial\varrho}\right) = \tfrac{\partial^2\pi}{\partial z^2} - \gamma^2\cdot\pi^2 . \end{aligned} \tag{8}$$

The Hertz-vector $\boldsymbol{\pi}$ is a solution of wave equations

$$\nabla^2\pi_j = \gamma_j^2 \cdot \pi_j \tag{9}$$

which in cylindrical coordinate system with azimuthal symmetry reduces to

$$\frac{\partial^2\pi_j}{\partial\varrho^2} + \frac{1}{\varrho}\cdot\frac{\partial\pi_j}{\partial\varrho} + \frac{\partial^2\pi_j}{\partial z^2} - \gamma_j^2\cdot\pi_j = 0 . \tag{10}$$

The solution of the above equations for air (π_o), homogeneous layers (π_j) and lower half-space(π_N) of the model with constant radiation constants are well known and can be written as

$$\pi_o = M \cdot \int_0^\infty \frac{\lambda}{n_o} \cdot e^{-n_o|z-h|} \cdot J_o(\lambda\varrho) d\lambda + \int_0^\infty A(\lambda) \cdot e^{-n_o z} \cdot J_o(\lambda\varrho) d\lambda \tag{11}$$

$$\begin{gathered} \pi_j = \int_0^\infty \{B_{j1}(\lambda)\cdot e^{n_j z} + B_{j2}(\lambda)\cdot e^{-n_j z}\}\cdot J_o(\lambda\varrho) d\lambda \\ \text{for} \quad j = 1, 2, \ldots, j-1, j+1, \ldots, N-1 \end{gathered} \tag{12}$$

$$\pi_N = \int_0^\infty D(\lambda) \cdot e^{n_N z} \cdot J_o(\lambda\varrho) d\lambda \tag{13}$$

with $n_j = \sqrt{\lambda^2 + \gamma_j^2}$, for all j.

In the following, the solution of the equation for the transition layer, where the radiation constant varies with depth has been obtained. After obtaining the solution, the appropriate boundary conditions requiring the continuity of tangential electric and magnetic fields at different interfaces have been utilised to derive the field components over the surface of the proposed model. In general, these boundary conditions at the interface of jth and $(j+1)^{\text{th}}$ layers can be written as

$$\pi_j = \pi_{j+1} \tag{14}$$

$$\frac{\partial \pi_j}{\partial z} = \frac{\partial \pi_{j+1}}{\partial z} \,. \tag{15}$$

2. Analytical solution

For the exponential variation of $\gamma_j(z)$ in the transition layer, the wave Eq. (10) assumes the form

$$\frac{\partial^2 \pi}{\partial \varrho^2} + \frac{1}{\varrho} \cdot \frac{\partial \pi}{\partial \varrho} + \frac{\partial^2 \pi}{\partial z^2} - \left\{\gamma_{j-1}^2 \exp\left[\alpha_2 \cdot (z + h_{j-1})\right]\right\} \cdot \pi = 0 \,. \tag{16}$$

Substituting $\pi = R(\varrho) \cdot Z(z)$ in the above equation and separating the variables, we get following equations

$$\frac{d^2 R}{d\varrho^2} + \frac{1}{\varrho} \cdot \frac{dR}{d\varrho} + \lambda^2 R = 0 \tag{17}$$

and

$$\frac{d^2 Z}{dz^2} - \left[\gamma_{j-1}^2 \exp\left\{\alpha_2 (z + h_{j-1})\right\} + \lambda^2\right] Z = 0 \,, \tag{18}$$

where λ is the separation constant.

A non-diverging solution of Eq. (17) is $R(\varrho) = J_o(\lambda\varrho)$.

Making a substitution of the form $\xi = e^{\frac{\alpha_2(z + h_{j-1})}{2}}$, modifies the Eq. (18) to

$$\frac{d^2 Z}{d\xi^2} + \frac{1}{\xi} \cdot \frac{dZ}{d\xi} - \left[\frac{4\gamma_{j-1}^2}{\alpha_2^2} + \frac{4\lambda^2}{\alpha_2^2} \cdot \frac{1}{\xi^2}\right] Z = 0 \,. \tag{19}$$

On comparison of this equation with the standard Bessel equation (Mclachlan 1955), we get the solution of Eq. (19) as,

$$Z = \left[C_1(\lambda) \cdot I_{2\lambda/\alpha_2}\left(\frac{2\gamma_{j-1}}{\alpha_2} \cdot \xi\right) + C_2(\lambda) \cdot K_{2\lambda/\alpha_2} \cdot \left(\frac{2\gamma_{j-1}}{\alpha_2} \cdot \xi\right)\right] \,. \tag{20}$$

Hence, the Hertz-vector potential in transition layer is given by

$$\pi(Z) = \int_0^\infty \left[C_1(\lambda) \cdot I_{2\lambda/\alpha_2} \left(\frac{2\gamma_{j-1}}{\alpha_2} \cdot e^{\frac{\alpha_2(z+h_{j-1})}{2}} \right) + \right. \\ \left. + C_2(\lambda) \cdot K_{2\lambda/\alpha_2} \cdot \left(\frac{2\gamma_{j-1})}{\alpha_2} \cdot e^{\frac{\alpha_2(z+h_{j-1})}{2}} \right) \right] \cdot J_o(\lambda\varrho) d\lambda \,. \tag{21}$$

The constants $A(\lambda)$, $B_{j1}(\lambda)$, $B_{j2}(\lambda)$ for all j, except $j = j; C_1(\lambda), C_2(\lambda)$ and $D(\lambda)$, hereafter denoted as $A, B_{j1}, B_{j2}, C_1, C_2$ and D are evaluated using the boundary conditions given by Eqs (14) and (15), at different interfaces. The application of these $2N$ boundary conditions at N interfaces give rise to $2N$ linear equations, with the help of which one has to find out the value of $2N$ unknowns. These $2N$ linear equations resulting from the boundary conditions can be written in appropriate matrix form. The solution of this matrix equation would give the value of constants for a general N-layered model having presumed type of conductivities in different layers. Whereafter the field components over the surface of the model with desired number of layers can be obtained.

In the following, the solution for the three layer earth model with the intermediate inhomogeneous layer having exponentially varying conductivity, acting as a transition zone between the top and bottom homogeneous layers is presented. The matrix equation for this model is Eq. (22) (see next page),

$$\text{where} \xi_1 = \frac{2\gamma_1}{\alpha_2}; \xi_2 = \frac{2\gamma_1}{\alpha_2} \cdot e^{\frac{\alpha_2(h_1 - h_2)}{2}} \text{and} \gamma_2^2(z) = \gamma_1^2 \cdot e^{\frac{\alpha_2(z+h_1)}{2}} \,.$$

On solving the matrix Eq. (22), we get the required constant A as

$$A = M \cdot \frac{\lambda}{n_o} \cdot \left[\frac{\left(\frac{n_o - n_1}{n_o + n_1}\right) + \left\{ \frac{(1+\lambda/n_1)+(\gamma_1/n_1)\cdot v}{(1-\lambda/n_1)-(\gamma_1/n_1)\cdot v} \right\} \cdot e^{-2n_1 h_1}}{1 + \left(\frac{n_o - n_1}{n_o + n_1}\right) \left\{ \frac{(1+\lambda/n_1)+(\gamma_1/n_1)\cdot v}{(1-\lambda/n_1)-(\gamma_1/n_1)\cdot v} \right\} \cdot e^{-2n_1 h_1}} \right] \cdot e^{-n_o h} \,, \tag{23}$$

where

$$v = \frac{K_{(2\lambda/\alpha_2)-1}(\xi_1) + u \cdot I_{(2\lambda/\alpha_2)-1}(\xi_1)}{K_{2\lambda/\alpha_2}(\xi_1) - u \cdot I_{2\lambda/\alpha_2}(\xi_1)}$$

$$u = \frac{(n3 + \lambda) \cdot K_{2\lambda/\alpha_2}(\xi_2) + \frac{\alpha_2 \xi_2}{2} \cdot K_{(2\lambda/\alpha_2)-1}(\xi_2)}{(n3 + \lambda) \cdot I_{2\lambda/\alpha_2}(\xi_2) - \frac{\alpha_2 \xi_2}{2} \cdot I_{(2\lambda/\alpha_2)-1}(\xi_2)} \,.$$

Now substituting the value of constant A from Eq. (23) in Eq. (11), we get the

$$
\begin{bmatrix}
M\lambda e^{-n_o h} & n_o & -n_o & -n_o & 0 & 0 & 0 & 0 \\
M\lambda e^{-n_o h} & -n_o & -n_1 & n_1 & 0 & 0 & 0 & 0 \\
0 & 0 & e^{-n_1 h_1} & e^{n_1 h_1} & -I_{2\lambda/\alpha_2}(\xi_1) & -K_{2\lambda/\alpha_2}(\xi_1) & 0 & 0 \\
0 & 0 & n_1 e^{-n_1 h_1} & -n_1 e^{n_1 h_1} & -\frac{d}{dz}[I_{2\lambda/\alpha_2}(\xi_1)] & -\frac{d}{dz}[K_{2\lambda/\alpha_2}(\xi_1)] & 0 & 0 \\
0 & 0 & 0 & 0 & I_{2\lambda/\alpha_2}(\xi_2) & K_{2\lambda/\alpha_2}(\xi_2) & -e^{-n_3 h_2} & 0 \\
0 & 0 & 0 & 0 & \frac{d}{dz}[I_{2\lambda/\alpha_2}(\xi_2)] & \frac{d}{dz}[K_{2\lambda/\alpha_2}(\xi_2)] & -n_3 e^{n_3 h_2} & 0
\end{bmatrix}
\begin{bmatrix} 1 \\ A \\ B_{11} \\ B_{12} \\ C_1 \\ C_2 \\ D \\ 0 \end{bmatrix}
\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \qquad (22)
$$

expression of Hertz-vector in the air as

$$\pi_o = M \cdot \int_0^\infty \frac{\lambda}{n_o} \cdot \left[e^{-n_o|z-h|} + \left\{ \frac{\left(\frac{n_o - n_1}{n_o + n_1}\right) + \left\{\frac{(1+\lambda/n_1)+(\gamma_1/n_1)\cdot v}{(1-\lambda/n_1)-(\gamma_1/n_1)\cdot v}\right\} \cdot e^{-2n_1h_1}}{1 + \left(\frac{n_o - n_1}{n_o + n_1}\right)\left\{\frac{(1+\lambda/n_1)+(\gamma_1/n_1)\cdot v}{(1-\lambda/n_1)-(\gamma_1/n_1)\cdot v}\right\} \cdot e^{-2n_1h_1}} \right\} \cdot e^{n_o(z+h)} \right] \cdot J_o(\lambda\varrho) d\lambda \,. \tag{24}$$

Taking the conductivity of air $\sigma_o = 0$; and placing the source and observation point on the ground surface ($h = 0$ and $z = 0$), the Eq. (24) reduces to

$$\pi_o \begin{pmatrix} z = o \\ h = o \end{pmatrix} = 2M \int_0^\infty \frac{\lambda}{(\lambda + n_1)} \cdot \left[\frac{1 + \left\{\frac{(1+\lambda/n_1)+(\gamma_1/n_1)\cdot v}{(1-\lambda/n_1)-(\gamma_1/n_1)\cdot v}\right\} \cdot e^{-2n_1h_1}}{1 + \left(\frac{n_o - n_1}{n_o + n_1}\right)\left\{\frac{(1+\lambda/n_1)+(\gamma_1/n_1)\cdot v}{(1-\lambda/n_1)-(\gamma_1/n_1)\cdot v}\right\} \cdot e^{-2n_1h_1}} \right] \cdot J_o(\lambda\varrho) d\lambda \,. \tag{25}$$

Further, this Hertz-vector potential expression (25), for a homogeneous earth, $h_1 \to \infty$, reduces to

$$\pi_o \begin{pmatrix} z = o \\ h = o \end{pmatrix} = 2M \cdot \int_0^\infty \frac{\lambda}{(\lambda + n_1)} \cdot J_o(\lambda\varrho) d\lambda \,. \tag{26}$$

This expression (26) is the same as that derived by Wait (1951) for the homogeneous half space, which provides the confirmation for the result derived in the expression (24).

The magnetic and electric field components in the air can be derived from the Eq. (8). In air, on the earth surface these components are

$$E_\varrho = 0 \tag{27}$$

$$E_\phi = i\omega\mu \cdot \frac{\partial \pi_o}{\partial \varrho} = \\ = -2i\omega\mu_o \cdot M \cdot \int_0^\infty \frac{1}{(\lambda + n_1)} \left[\frac{1 + W \cdot e^{-2n_1h_1}}{1 + \left(\frac{\lambda - n_1}{\lambda + n_1}\right) W \cdot e^{-2n_1h_1}} \right] \cdot \lambda^2 \cdot J_1(\lambda\varrho) d\lambda \tag{28}$$

$$E_z = 0 \tag{29}$$

$$H_\varrho = \frac{\partial^2 \pi}{\partial \varrho \cdot \partial z} = -2M \int_0^\infty \frac{n_1}{(n_o + n_1)} \cdot$$
$$\cdot \left[\frac{1 - W \cdot e^{-2n_1 h_1}}{1 + \left(\frac{\lambda - n_1}{\lambda + n_1}\right) W \cdot e^{-2n_1 h_1}} \right] \cdot \lambda^2 \cdot J_1(\lambda\varrho) d\lambda \tag{30}$$

$$H_\phi = o \tag{31}$$

$$H_z = -\frac{1}{\varrho} \cdot \frac{\partial}{\partial \varrho} \left(\varrho \cdot \frac{\partial \pi_o}{\partial \varrho} \right) = \frac{\partial^2 \pi_o}{\partial z^2} - \gamma_o^2 \pi_o =$$
$$= 2M \cdot \int_0^\infty \frac{n_1}{(n_o + n_1)} \cdot \left[\frac{1 + W \cdot e^{-2n_1 h_1}}{1 + \left(\frac{\lambda - n_1}{\lambda + n_1}\right) W \cdot e^{-2n_1 h_1}} \right] \cdot \lambda^3 \cdot J_o(\lambda\varrho) d\lambda \ , \tag{32}$$

where $W = \left[\frac{(1 + \lambda/n_1) + (\gamma_1/n_1) \cdot v}{(1 - \lambda/n_1) - (\gamma_1/n_1) \cdot v} \right]$.

3. Computational scheme

The computation of the complex integral expressions occurring in the expressions of various electromagnetic field components have been performed by expressing the infinite integrals in terms of Hankel transforms of order 0 and 1 and thereafter applying the computer program ZHANKS (Anderson 1979) to calculate the resulting Hankel transforms of order 0 and 1. As per requirement of the scheme the convergence of the integrals have been tested and in case of slow convergence or divergent nature of the kernels associated with these integrals a known integral expression with an analytic equivalent has been added/subtracted inside the integral expression and subsequently has been adjusted outside the integral expression. As an illustration, the expression for the, E_ϕ is expressed as

$$E_\phi = -2i\omega\mu_o M \left[\int_0^\infty \left\{ \frac{1}{1 + \frac{R_N}{R_M} \cdot \frac{n_1}{\lambda}} - 0.5 \right\} \cdot \lambda \cdot J_1(\lambda\varrho) d\lambda + 0.5 \int_0^\infty \lambda \cdot j_1(\lambda\varrho) d\lambda \right] \tag{33}$$

with $R_M = 1 + We^{-2n_1 h_1}$ and $R_N = 1 - We^{-2n_1 h_1}$.

In this expression (33), the first part is evaluated using computer program ZHANKS of Anderson (1979), whereas the second part is directly evaluated using the following relation (Watson 1962)

$$\int_0^\infty \lambda^2 \cdot J_1(\lambda\varrho) d\lambda = 1/\varrho^2 \ . \tag{34}$$

Further, for the evaluation of Kernels, an approximate expression of v, for the resistive basement case has been used for the ease of computations. The evaluation of the modified Bessel functions, occurring in the expressions has been accomplished by using the standard expression for them (see Mclachlan 1955).

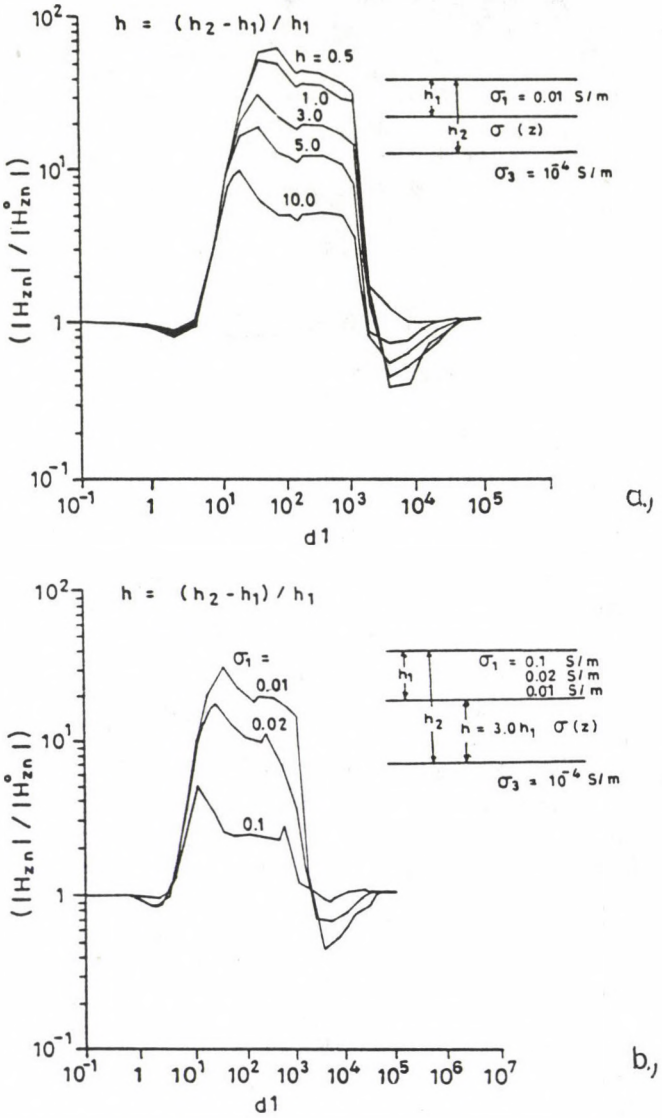

Fig. 2. Plot of the $|E_{\varrho n}|$ / $|E^{o}_{\varrho n}|$ vs d_1, showing the effects of variation of a) transition layer thickness; b) conductivity contrast between the top and bottom layers

4. Results and discussions

To study the effect of exponential variation of subsurface conductivity on EM response, various three layer models have been chosen with the presumption that the conductivity of the top layer gradually merges with that of the basement or substratum following the exponential variation in the transition layer. The computations for the absolute-amplitude ratio (the ratio of the absolute-amplitude of

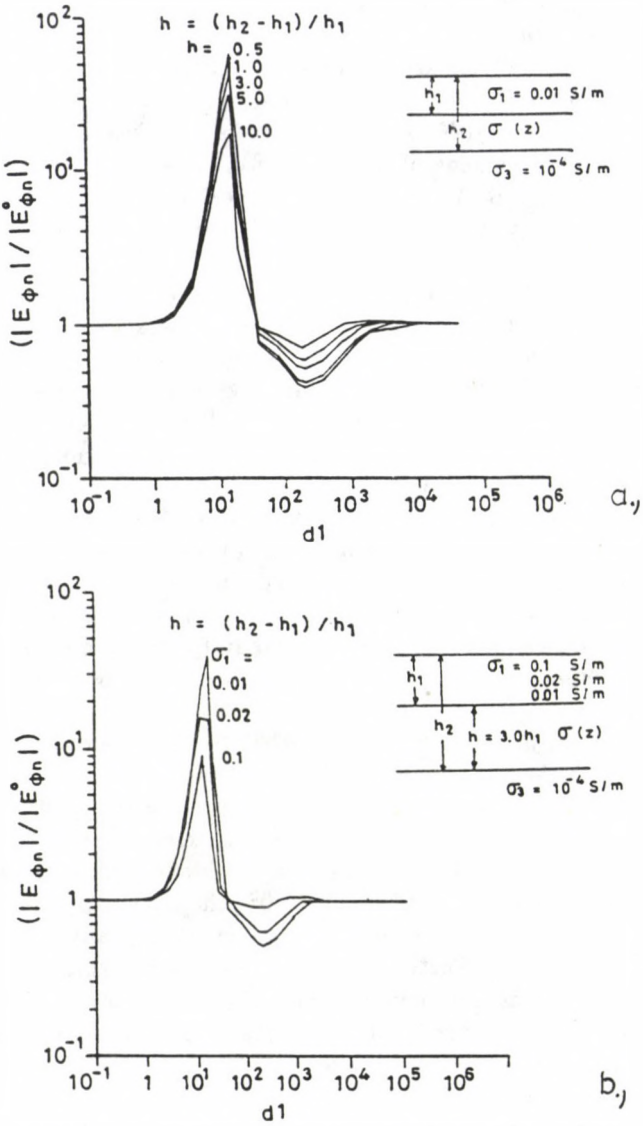

Fig. 3. Plot of the $|H_{\phi n}|$ / $|H^{o}_{\phi n}|$ vs d_1, showing the effects of variation of a) transition layer thickness; b) conductivity contrast between the top and bottom layers

the field component and the corresponding component for the homogeneous half space having the conductivity of the top layer of the proposed model) values of different field components have been performed, and the results showing the effect of transition layer thickness and the conductivity contrast between the top and bottom layers have been presented in the form of absolute-amplitude ratio of field components expressed as a function of numerical distance, $d_1 = \{(\omega\mu_o\sigma_1)^{1/2} \cdot \varrho\}$.

The variation of absolute-amplitude ratio of H_z, E_ϕ and H_ϱ components with numerical distance (d_1), for the relative transition layer thicknesses $h = (h_1 - h_2)/h_1$ = 0.5, 1.0, 3.0, 5.0 and 10.0 are presented in the Figs 2a, 3a and 4a, for the models shown in the inset of figures and for the relative conductivity contrast between the top and bottom layers with constant thickness of transition zone are presented in Figs 2b, 3b and 4b. A general observation of these curves reveal that the curves have values close to unity for very small and very large values of numerical distance and show their characteristics variations only for the intermediate value of d_1.

From the plot of the absolute-amplitude ratio of H_z-component with d_1, for the various relative thickness of transition layer as shown in Fig. 2a, it is clear that these curves show a fixed value unity for small values of d_1 ($d_1 \leq 3$), and thereafter increase to attain a maximum peak at about $d_1 \simeq 4$, and then decrease sharply to a moderate value of the amplitude which is maintained upto $d_1 \simeq 5 \cdot 10^2$. Again an increase begins at about $d_1 \simeq 10^3$ forming a secondary maximum of lesser amplitude than that of the primary maximum, whereafter, the curves descend to approach unity value at $d_1 \simeq 4 \cdot 10^4$, with minor undulations in between $d_1 \simeq 4 \cdot 10^3$ and $4 \cdot 10^4$. As the thickness of transition layer increases, amplitude lower down and peaks become sharper and well defined. From the Fig. 2b showing the effect of conductivity contrast between the top and bottom layers on the absolute-amplitude ratio curve of H_z-component, it is clear that the maximum and minimum peak become more pronounced with decreasing conductivity contrast.

Figure 3a showing the variation of absolute-amplitude ratio of E_ϕ-component for the various thickness of transition layer reveals that the curves have a value very close to one for small values of d_1 and then at $d_1 \simeq 0.8$, start increasing sharply to attain a maximum at about $d_1 \simeq 10$ and thereafter decrease sharply to regain unity value at $d_1 \simeq 10^3$, after forming a small lower peak in between $d_1 \simeq 40$ and $d_1 \simeq 10^3$. For larger value of $d_1 (d_1 \geq 10^3)$ the curves maintain the fixed value unity. As the thickness of transition layer increases, peaks of the curves decrease in amplitude and width. Further, from the Fig. 3b showing the effect of variation of conductivity contrast on absolute-amplitude ratio curve of E_ϕ component, it is clear that as the conductivity contrast decreases, maximum peak increases and the minimum peak decreases in amplitude.

Figure 4a showing the variation of absolute-amplitude ratio of H_ϱ-component with numerical distance (d_1), for various thickness of transition layer reflects that, the curves are smooth, well defined and have a clear pattern. These curves show a fixed value unity for small values of d_1, then decrease and form a lower peak in the intermediate value range of d_1 and thereafter again increase to attain the value unity for large value of d_1. These curve have a valley type shape. As the thickness of transition layer increases peak decreases in size and width. From the Fig. 4b, showing the effect of variation of conductivity contrast between the top and bottom layers, it is clear that with the decreasing conductivity contrast, minimum peak decreases to lower values and thus marks a slight structural difference.

5. Conclusions

The present work is devoted to the study of the EM response, of a vertical magnetic dipole, situated over the surface of vertically inhomogeneous earth model having exponential variation of conductivity with depth, in the inhomogeneous layer. Both analytical and computational results have been obtained for the three

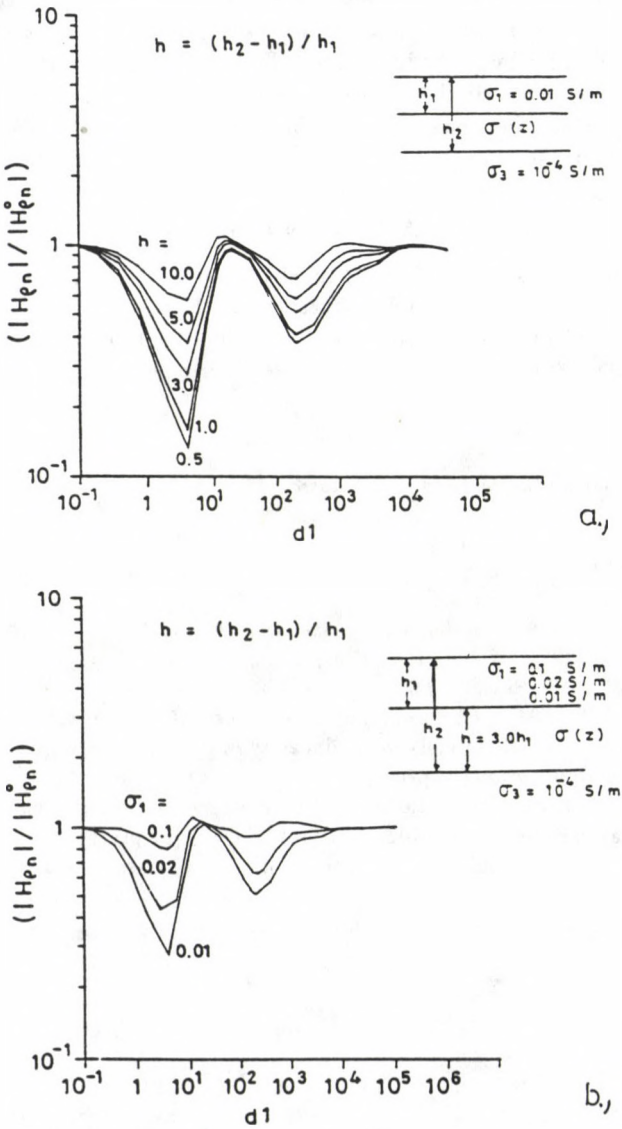

Fig. 4. Plot of the $|H_{zn}|$ / $|H^o_{zn}|$ vs d_1, showing the effects of variation of a) transition layer thickness; b) conductivity contrast between the top and bottom layers

layered earth model with the intermediate layer having above type of variation. The analytical solutions are obtained in terms of integral equations and the numerical computations are performed by using the computer program based on digital filter theory. The procedure is quite simple and comprehensive.

From the plots of the absolute amplitude ratio of the different field components H_z, E_ϕ and H_ϱ, against the numerical distance (d_1) showing the effect of variation of relative transition layer thicknesses and condcutivity contrast, it is clear that all the curves show a unity value for very small values of d_1 and then show their characteristics variation for the intermediate values of d_1. Thereafter, they again converge to the unity value for larger values of d_1. The peaks of the curves decrease in size, amplitude and width with the increasing thickness of transition layer. Further, it is also noticed that the peaks increase in amplitude and width, with decreasing conductivity contrast. This variation may be assigned to the model characteristic.

Acknowledgements

The authors are thankful to Shri B Singh, Scientist, National Geophysical Research Institute, Hyderabad, for his valuable suggestions in the computation of the problem. One of the authors (NPS) is thankful to CSIR, New Delhi, for providing the financial support in the form of fellowship.

References

Abramovici F, Chlamtac M 1978: *Geophysics*, 43, 954–966.

Anderson W L 1979: *Geophysics*, 44, 1287.

Chlamtac M, Abramovici F 1981: *Geophysics*, 46, 904–915.

Keller G V 1971: In: Electromagnetic probing in geophysics. J R Wait (ed.), Golden Press, Boulder

Mallick K, Roy A 1968: *Geophysical Prospecting*, 16, 436.

Mallick K, Roy A 1971: *Geophysical Prospecting*, 19, 388.

Mclachlan N W 1955: Bessel functions for engineers. Oxford

Singh N P, Lal T 1994a: Electromagnetic response of horizontal magnetic dipole over inhomogeneous earth models with linear variation of conductivity. *Journal of EM Waves and Application* (in press)

Singh N P, Lal T 1994b: Horizontal magnetic dipole over inhomogeneous earth model with exponential variation of conductivity. *PAGEOPH* (in press)

Sommerfeld A 1949: Partial differential equations in physics. Academic Press Inc., New York

Wait J R 1951: *Can. Journ. Phys.*, 29, 577–592.

Acta Geod. Geoph. Hung., Vol. 29(1–2), pp. 235–249 (1994)

A MAGNETOTELLURIC 2-D INVERSION TECHNIQUE

CHEN LE-SHOU[1], WANG GUANG-E[1], CHEN JIU-PING[1], JIAO JIAN-WU[1], ZHAN SHUANG-QING[2], DONG DA-KUI[2], CAI GANG[3]

[Manuscript received July 13, 1994]

2-D MT imaging was performed through successfully applying the Generalized Pulse Spectrum Technique (GPST) (Chen 1985, Li Jian hua et al. 1987, Xie Gang quan and Li Jian hua 1988, Yang Wen cai 1989) which has been widely used in the tomographic technique, to the problem of 2-D MT inversion. The inversion was speeded up because the Frechet derivatives were calculated while computing the 2-D MT forward problem using Finite Element Method (FEM). The theory of this imaging method is rigorous and it is numerically stable and fast in convergence. The method is easy to be adopted on microcomputer 386. Test with theoretical data and field data shows good imaging results in this paper.

Keywords: imaging technique; magnetotelluric static effect; magnetotellurics

1. Introduction

Recently magnetotellurics (MT) has been widely used in the world. It is necessary to develop an effective interpretational technique for 2-D model. In this paper, a new unique 2-D inverse technique for MT data is presented from the Generalized Pulse Spectrum Technique (GPST) which has been widely used to solve the inverse problem of partial differential equations in the tomographic technique. By delineating the geo-electrical structure by a continuous model whenever an initial model is given, a geoelectrical model fitted the observed data can be obtained by means of automatic iterative inversion. Using the Finite Element Method (FEM), the field is computed and the Frechet derivative function of the apparent resistivity has been computed also, thereby the computation of inversion is speeded up significantly. This method is superior to the fashionable methods of 2-D MT inversion, e.g. to the minimum model inversion method. Some achievements in technique-scientific research have been used in this method. It is theoretically rigorous and comparable with the effect of resistivity imaging in practice. It is numerically stable and fast in convergence. The program can be run on a PC.

[1]China University of Geosciences, 100083 Beijing, China

[2]Geophysical Exploration Party No. 1, Ministry of Geology and Mineral Resources, 233005 Bang Fu, An hui, China

[3]Academy of Geoexploration, Ministry of Geology and Mineral Resources, 100812 Beijing, China

2. Theory

The objective function of 2-D inversion for the magnetotelluric apparent resistivity data is defined as:

$$F[\lambda] = \sum_{i=1}^{M\times N} \{[\ln \varrho_{aci} - \ln \varrho_{ai}] + Q'_{\lambda i}\Delta\lambda\}^2 \tag{1}$$

where M is the number of MT stations, N the number of frequency points, λ the vector of model parameters, the parameter may be both resistivity or conductivity, $Q'_{\lambda i}$ is the Frechet derivative function of Q or $\ln \varrho_{aci}$ with respect to λ, $Q_i = \ln \varrho_{aci} - \ln \varrho_{ai}$, ϱ_{ai} is the observed apparent resistivity, ϱ_{aci} theoretical apparent resistivity.

In the 2-D MT problem, the apparent resistivities for the two kinds of polarization are as follows (Liu Guo dong and Chen Le shou 1984):

$$\varrho_a = \frac{i}{\omega\mu}\left(\frac{V}{J}\right)^2, \quad \text{for E polarization} \tag{2}$$

$$\varrho_a = \frac{i}{\omega\mu}\left(\frac{J}{V}\right)^2, \quad \text{for B polarization} \tag{3}$$

where V is the main field, E_x and H_x, respectively, J is the auxiliary field H_y or E_y, x is the strike of the structure.

Taking the natural logarithm of Eqs (2) and (3) leads to:

$$\ln \varrho_a = \ln \frac{i}{\omega\mu} + 2\ln V - 2\ln J, \quad \text{for E polarization} \tag{4}$$

$$\ln \varrho_a = \ln \frac{i}{\omega\mu} + 2\ln J - 2\ln V, \quad \text{for B polarization} \tag{5}$$

by differentiation it follows that:

$$Q_\lambda \cdot \Delta\lambda = 2\left[\frac{\Delta V}{V} - \frac{\Delta J}{J}\right], \quad \text{for E polarization} \tag{6}$$

$$Q_\lambda \cdot \Delta\lambda = 2\left[\frac{\Delta J}{J} - \frac{\Delta V}{V}\right], \quad \text{for B polarization.} \tag{7}$$

First of all, we deal with the problem of B polarization. In this time, H_x, for the sake of brevity, will be written as H, it fulfils the following equations:

$$\frac{\partial}{\partial y}\left(\varrho\frac{\partial H}{\partial y}\right) + \frac{\partial}{\partial z}\left(\varrho\frac{\partial H}{\partial z}\right) + i\omega\mu H = 0, \tag{8}$$

$$H = 1, \quad \text{for } z = z_{\min} \tag{9}$$

$$\frac{\partial H}{\partial y} = 0, \quad \text{for } y = y_{\max},\ y = y_{\min} \tag{10}$$

$$\varrho\frac{\partial H}{\partial z} + Z_{H_S}\cdot H = 0, \quad \text{for } z = z_{\max} \tag{11}$$

where $y_{\min}$ and $y_{\max}$ are the y-ordinates of the left and right boundaries of the rectangular zone, respectively; $z_{\min}$ and $z_{\max}$ are the z-ordinates of the upper and lower boundaries of the ractangular zone which is taken into account for the FEM computation; Z_{H_s} is the surface impedance at the lower boundary.

The variation of the parameter of the medium, and the variation of the field due to a parameter variation, are given by

$$\Delta H = H - H_0 \tag{12}$$

$$\Delta\varrho = \varrho - \varrho_0 \tag{13}$$

substituting Eqs (12), (13) into Eqs (8)–(11) and omitting the higher order infinitesimals of ΔH and $\Delta\varrho$, an elliptical boundary value problem is obtained for H_0 which is exactly similar to Eqs (8)–(11) formally, and an elliptical type hybrid boundary value problem for ΔH, the latter being:

$$\begin{aligned} &\tfrac{\partial}{\partial y}\left(\varrho_0\tfrac{\partial\Delta H}{\partial y}\right) + \tfrac{\partial}{\partial z}\left(\varrho_0\tfrac{\partial\Delta H}{\partial z}\right) + i\omega\mu\Delta H = \\ &= -\left[\tfrac{\partial}{\partial y}\left(\Delta\varrho\tfrac{\partial H_0}{\partial y}\right) + \tfrac{\partial}{\partial z}\left(\Delta\varrho\tfrac{\partial H_0}{\partial z}\right)\right] = -Ln(y,z,\Delta\varrho,H_0) \end{aligned} \tag{14}$$

$$\Delta H = 0, \quad \text{for } z = 0 \tag{15}$$

$$\frac{\partial\Delta H}{\partial y} = 0, \quad \text{for } y = y_{\min},\ y = y_{\max} \tag{16}$$

$$\varrho_0\frac{\partial\Delta H}{\partial z} + Z_{H_s}\Delta H = 0, \quad \text{for } z = z_{\max}\,. \tag{17}$$

The corresponding Green function $G(y,z,y',z',\omega)$ is the solution of the following definite-solution problem:

$$\frac{\partial}{\partial y}\left(\varrho_0\frac{\partial G}{\partial y}\right) + \frac{\partial}{\partial z}\left(\varrho_0\frac{\partial G}{\partial z}\right) + i\omega\mu G = -\delta(y-y')(z-z')\,, \tag{18}$$

$$G = 0, \quad \text{for } z = 0 \tag{19}$$

$$\frac{\partial G}{\partial y} = 0, \quad \text{for } y = y_{\min},\ y = y_{\max} \tag{20}$$

$$\varrho_0\left(\frac{\partial G}{\partial z}\right) + Z_{H_s}G = 0, \quad \text{for } z = z_{\max}\,. \tag{21}$$

This problem of definite solutions formally is exactly similar to Eqs (14)–(17), except the source term, thus the Green function can be evaluated by FEM.

Using the Green function method for solving the elliptical boundary value problem leads to an integral equation with respect to $\Delta\varrho$ and ΔH from Eqs (14)–(17):

$$\begin{aligned} &\iint G(y,z,y',z',\omega)\left[-\tfrac{\partial}{\partial y'}\left(\Delta\varrho\tfrac{\partial H_0}{\partial y'}\right) - \right. \\ &\left. -\tfrac{\partial}{\partial z'}\left(\Delta\varrho\tfrac{\partial H_0}{\partial z'}\right)\right] dy'dz' = \Delta H(y,z,\omega)\,. \end{aligned} \tag{22}$$

By taking ΔH as the difference between successive results, namely:

$$\Delta H^{(n)}(y,z,\omega) = H^{(n+1)}(y,z,\omega) - H^{(n)}(y,z,\omega) \tag{23}$$

and substituting Eq. (23) into Eq. (22) and partially integrating (22), as well as partially differentiating with respect to z from boundary conditions we obtain:

$$\begin{aligned} \iint_{\Omega} \left[\frac{\partial}{\partial y'} \left(\frac{\partial G^{(n)}}{\partial z} \right) \frac{\partial H^{(n)}}{\partial y'} + \frac{\partial}{\partial z'} \left(\frac{\partial G^{(n)}}{\partial z} \right) \frac{\partial H^{(n)}}{\partial z'} \right] \Delta\varrho^{(n)}(y',z')dy'dz' = \\ = \frac{\partial H^{(n+1)}(y,z,\omega)}{\partial z} - \frac{\partial H^{(n)}(y,z,\omega)}{\partial z} . \end{aligned} \tag{24}$$

Taking (y,z) as coordinates of the MT station at the surface $(y,0)$ and according to the complementary boundary condition at the surface, namely:

$$\left. \frac{\partial H(y,z,\omega)}{\partial z} \right|_{z=0} = -\frac{1}{\varrho(y,0)} Z H(y,0,\omega) , \tag{25}$$

where Z is the observed surface impedance, substituting the value into Eq. (25) for the term $\frac{\partial H^{(n+1)}(y,z,\omega)}{\partial z}$ in the Eq. (24), a first kind Fredholm integral equation for $\Delta\varrho^{(n)}(y,z)$ is obtained:

$$\begin{aligned} \iint_{\Omega} \left[\frac{\partial}{\partial y'} \left(\frac{\partial G^{(n)}}{\partial z} \right) \Big|_{z=0} \cdot \frac{\partial H^{(n)}}{\partial y'} + \right. \\ \left. + \frac{\partial}{\partial z'} \left(\frac{\partial G^{(n)}}{\partial z} \right) \Big|_{z=0} \cdot \frac{\partial H^{(n)}}{\partial z'} \right] \Delta\varrho^{(n)}(y',z')dy'dz' = \\ = -\frac{1}{\varrho(y,0)} Z \cdot H(y,0,\omega) - \frac{\partial}{\partial z} H^{(n)}(y,z,\omega)|_{z=0} , \end{aligned} \tag{26}$$

where Z and $H(y,0,\omega)$ are observed values, $\varrho(y,0)$ is the resistivity at surface at the MT station which should be obtained by some other way or from other data.

From the discrete form of Eq. (26), by using an iterative method to solve this equation for $\Delta\varrho^{(n)}$, the modification of the geoelectrical model is obtained. Some steps are iterated until a fit is obtained.

Up to now, according to Eqs (7), (15), (24) the Frechet derivative function $Q'^{(n)}_{\lambda i}$ or the corresponding ordinary differential at the surface of MT station can be derived as follows:

$$\begin{aligned} \tfrac{1}{2} Q'^{(n)}_{\varrho} \delta\varrho^{(n)} = \frac{\Delta J^{(n)}}{J^{(n)}} = \frac{\partial \Delta H^{(n)}/\partial z}{\partial H^{(n)}/\partial z} = \frac{1}{\partial H^{(n)}/\partial z|_{z=0}} \cdot \\ \cdot \iint_{Q} \left[\frac{\partial}{\partial y'} \left(\frac{\partial G^{(n)}}{\partial z} \right) \Big|_{z=0} \cdot \frac{\partial H^{(n)}}{\partial y'} + \frac{\partial}{\partial z'} \left(\frac{\partial G^{(n)}}{\partial z} \right) \Big|_{z=0} \cdot \frac{\partial H^{(n)}}{\partial z'} \right] \Delta\varrho^{(n)} \cdot \Delta y' \Delta z' . \end{aligned} \tag{27}$$

In summary, the initial model $\varrho_0(y,z) > 0$ should be given according to results of 1-D interpretation, then the field $H^{(n)}$ and the corresponding Green function $G^{(n)}$ are evaluated by FEM with the aid of an identical coefficient matrix. After numerical discretization of the Fredholm integral equation, the resulting linear equations are solved by normalized generalized inverse matrix inversion method for

the modification of electrical parameter $\Delta\varrho^{(n)}$ and the modified parameter $\varrho^{(n+1)}$ can be obtained.

For E polarization, from an analogous derivation, the following elliptic boundary value problem for ΔE can be obtained:

$$\frac{\partial}{\partial y}\left(\frac{1}{i\omega\mu}\frac{\partial\Delta E}{\partial y}\right)+\frac{\partial}{\partial z}\left(\frac{1}{i\omega\mu}\frac{\partial\Delta E}{\partial z}\right)+\sigma_0\Delta E=-\Delta\sigma\cdot E_0\ , \tag{28}$$

$$\Delta E = 0, \quad \text{for} \quad z=z_{\min} \tag{29}$$

$$\frac{\partial\Delta E}{\partial y} = 0, \quad \text{for} \quad y=y_{\min},\ y=y_{\max} \tag{30}$$

$$\frac{1}{i\omega\mu}\frac{\partial\Delta E}{\partial z}-Z_{H_x}\cdot\Delta E = 0, \quad \text{for} \quad z=z_{\max}\ . \tag{31}$$

Substituting G for ΔE in Eqs (28)–(31) and $-\delta(y-y')(y-z)$ for the right hand term of source in Eq. (28) leads to the corresponding boundary value problem which the Green function meets.

Using the Green function method to solve the elliptic boundary value problem, an integral relationship between ΔE and $\Delta\sigma$ is obtained

$$\Delta E=\iint\limits_Q (E_0G_0)\Delta\sigma dy'dz'\ . \tag{32}$$

By partially differentiating Eq. (32) with respect to z, and by using a complementary boundary condition for E polarization which corresponds to Eq. (25), it follows that:

$$\begin{aligned}&\iint\limits_Q\left[E^{(n)}(y',z')\left.\frac{\partial G^{(n)}(y,z,y',z',\omega)}{\partial z}\right|_{z=0}\right]\Delta\sigma^{(n)}dy'dz'=\\&=\frac{\partial}{\partial z}E^{(n+1)}(y,z)\Big|_{z=0}-\left.\frac{\partial E^{(n)}(y,z)}{\partial z}\right|_{z=0}=\\&=i\omega\mu Z^{-1}E(y,0,\omega)-\left.\frac{\partial E^{(n)}(y,z,\omega)}{\partial z}\right|_{z=0}\ .\end{aligned} \tag{33}$$

This is a Fredholm integral equation corresponding to Eq. (31), thereby the parameter modification $\Delta\sigma^{(n)}$ can be evaluated.

From Eqs (6) and (32), the Frechet derivative function of Q, corresponding to Eq. (27), for E polarization can be obtained.

3. The evaluation of the Green function and its derivative function

To obtain a stable and convergent modification $\Delta\varrho^{(n)}$ or $\Delta\sigma^{(n)}$ from Eqs. (26) and (33), the following two technical keys must be carefully handled:

1. satisfactory accuracy should be assured for the computation of the Green function and its second derivatives;

2. the ill-definition problem of the Fredholm integral equation should be solved. For the latter, the well known normalized generalized inverse method has been used (Tikhonov and Arsenin 1977).

As for the first problem, e.g. for B polarization, it can be seen from Eq. (26) that the partial derivative of the Green function $G^{(n)}$ is contained in the coefficients. To increase the accuracy of computation, the FEM has been used to compute the partial derivative of the Green function directly rather than to compute the Green function itself, thus the times of difference operation can be decreased.

Equations (18)–(21) constitute a definite-solution problem, which is met by the Green function G. $G^{(n)}$ and $\varrho^{(n)}$ are substituted for G and ϱ_0 in the equations, and partial derivatives with respect to z are evaluated on the two sides of each equation. Let

$$V_n(y,z,y',z',\omega) = \frac{\partial}{\partial z} G^{(n)}(y,z,y',z',\omega)\,, \tag{34}$$

then V_n meets the following definite-solution problem:

$$\begin{aligned} &\tfrac{\partial}{\partial y'}\left[\varrho^{(n)}(y',z')\cdot\tfrac{\partial V_n}{\partial y'}\right] + \tfrac{\partial}{\partial z'}\left[\varrho^{(n)}(y',z')\tfrac{\partial V_n}{\partial z'}\right] + \\ &i\omega\mu V_n(y,z,y',z',\omega) = -\tfrac{\partial}{\partial z}\delta(y'-y)(z'-z) \end{aligned} \tag{35}$$

$$V_n(y,z,y',z'\omega) = 0, \quad \text{for} \quad z = z_{\min} \tag{36}$$

$$\frac{\partial}{\partial y'} V_n(y,z,y',z',\omega) = 0, \quad \text{for} \quad y = y_{\min},\ y = y_{\max} \tag{37}$$

$$\begin{aligned} &\varrho^{(n)}(y',z')\cdot\frac{\partial}{\partial z'} V_n(y,z,y',z',\omega) + \\ &+ Z_{H_x} V_n(y,z,y',z',\omega) = 0, \quad \text{for} \quad z = z_{\max}\,. \end{aligned} \tag{38}$$

According to the variational principle, solving the boundary value problem, *i.e.* Eqs (35)–(38), is equivalent to determine the mininum of functional $J[V_n]$ in the following equation:

$$\begin{aligned} J[V_n] = \iint_Q &\left\{\tfrac{\varrho}{2}\left(\tfrac{\partial V_n}{\partial y'}\right)^2 + \tfrac{\varrho}{2}\left(\tfrac{\partial V_n}{\partial z'}\right)^2 - \tfrac{i\omega\mu}{2}V_n^2 - \right. \\ &\left. -\tfrac{\partial}{\partial z}\delta(y'-y)(z'-z)V_n\right\} dy'dz' + \int_{z_{\max}}\left(\tfrac{Z_{H_x}}{2}V_n^2\right) dS = \min\,. \end{aligned} \tag{39}$$

Evaluating the fourth term in the surface integral of Eq. (39) by partial integrating and from boundary condition as well, where (y,z) is considered as the coordinate of MT station on the surface, thus we have:

$$\begin{aligned} J[V_n] = \iint_Q &\tfrac{1}{2}\left\{\varrho(y',z')\left[\left(\tfrac{\partial V_n}{\partial y'}\right)^2 + \left(\tfrac{\partial V_n}{\partial z'}\right)^2\right] - i\omega\mu V_n^2\right\} dy'dz' + \\ &+ \int_{z=z_{\max}}\left(\tfrac{Z_{H_x}}{2}\cdot V_n^2\right) ds + \tfrac{\partial}{\partial z'}V_n(y,0,y',z',\omega)\Big|_{y'=y,\ z'=z}\,. \end{aligned} \tag{40}$$

The equations relative to V_n are almost similar to the equations relative to the main field $H^{(n)}$, except the source term. Thus, in each iteration, meanwhile making

forward computation, we can evaluate the partial derivative of the Green function by increasing a little effort of computation, then the second derivative of $G^{(n)}$ can be obtained by accomplishing difference operation only once.

Similarly, the corresponding formulas evaluating the derivative function of Green function for E polarization could be derived.

4. Test of synthetic data and of observed data for B polarization

Test computation for synthetic data

a) In the model shown in Fig. 1a, the initial value of resistivity is 300 Ωm. After 11 iterations, the variance has been decreased from 76.5 percent to 4.6 percent. The resulting section is shown in Fig. 1b.

b) A model consists of three horizontal layers including two shallower conductors as shown in Fig. 2a. There are 11 MT observation stations and 20 frequency points, the mesh for FEM is 23×27. The pseudosection for apparent resistivity and phase of impedance are shown in Fig. 2b and Fig. 2c, respectively. The figures show different characteristic distortions due to the "static effect". At MT stations near two anomalous bodies, their effect on ϱ_a still exists, if the frequency decreases to 0.015 Hz. On the ϱ_a pseudosection there are two parallel conductive zones which almost run through the whole section, but the distortion of φ is confined to the skin depth, namely near high frequencies on the φ pseudosection. 28 blocks have been divided for inversion. After 8 iterations, the variance decreased from 95 percent to 2 percent. On the PC 386 the computation takes 2 hours and 19 minutes. The resulted section is shown in Fig. 2d: two shallow anomalous bodies have been well illustrated.

c) The model shown in Fig. 3a consists of two outcropping resistive bodies and a resistive horst. The mesh for FEM is 27×35. There are 12 MT stations, 20 frequency points and 50 blocks for inversion. The initial values for higher and lower resistivities are 600 Ωm and 200 Ωm, respectively. After 8 iterations, the variance decreased from 89 percent to 1.4 percent. The computation takes 4 hours and 22 minutes. The corresponding imaging section is shown in Fig. 3b, two anomalous bodies and deep horst have been very well resolved. The results shows that the inversion is successful.

d) In Fig. 4a the model consists of a horst complex and a small shallow anomalous body. Mesh for FEM is 28×29. There are 11 MT stations and 18 frequency points. 5 percent random error is added to the synthetic data to simulate observed data. There are 54 blocks for inversion. After 16 iterations, variance decreased from 32 percent to 5.5 percent, it takes 6 hours and 6 minutes. The resulted imaging section is shown in Fig. 4b.

All three models shown in Figs 2a, 3a and 4a contain shallow heterogeneities which can cause "static effect". The imaging results show that although distortions caused by the "static effect" are complicated, those could be identified on the apparent resistivity and phase pseudosections and shallow heterogeneities could be recovered directly by the MT GPST imaging method.

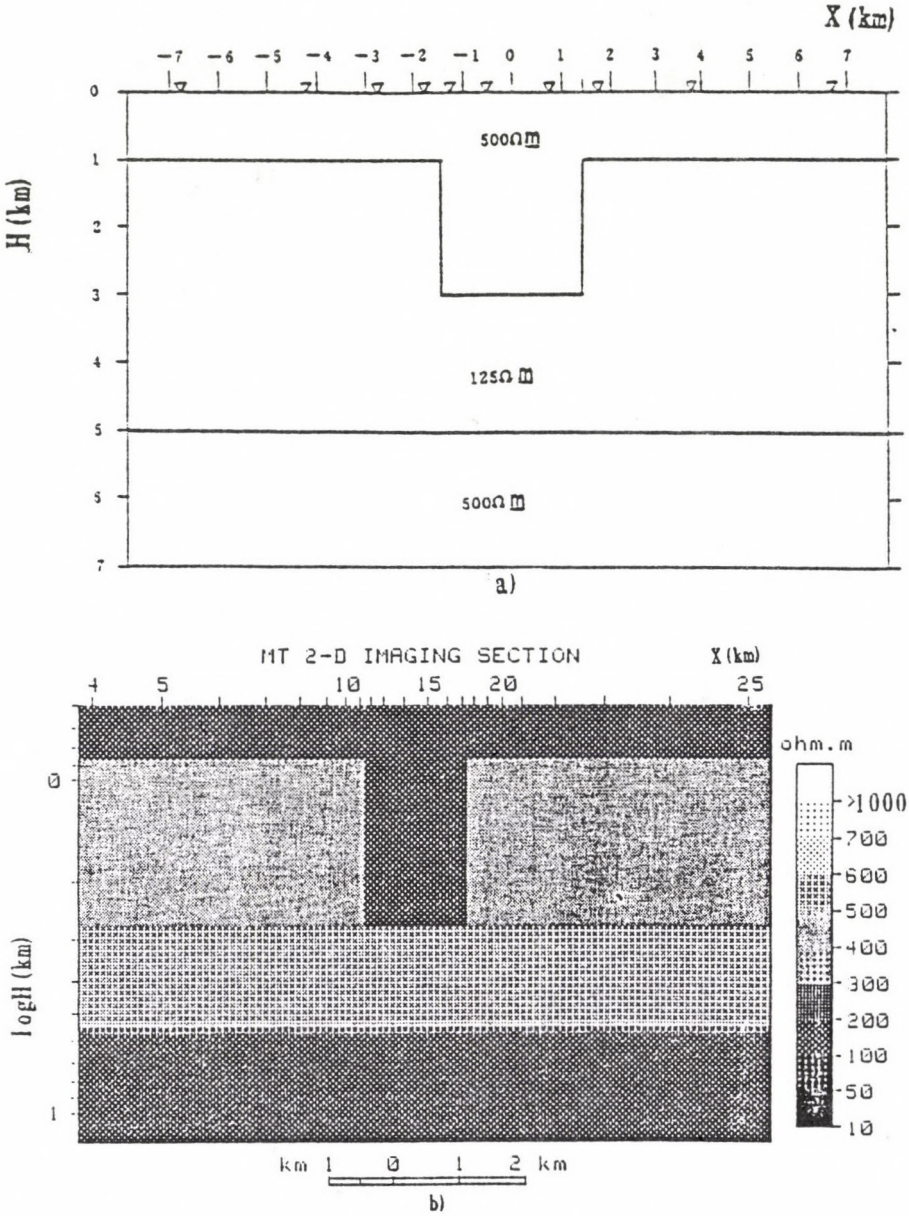

Fig. 1. Three layer section with resistive graben, a) model, b) 2-D MT imaging section

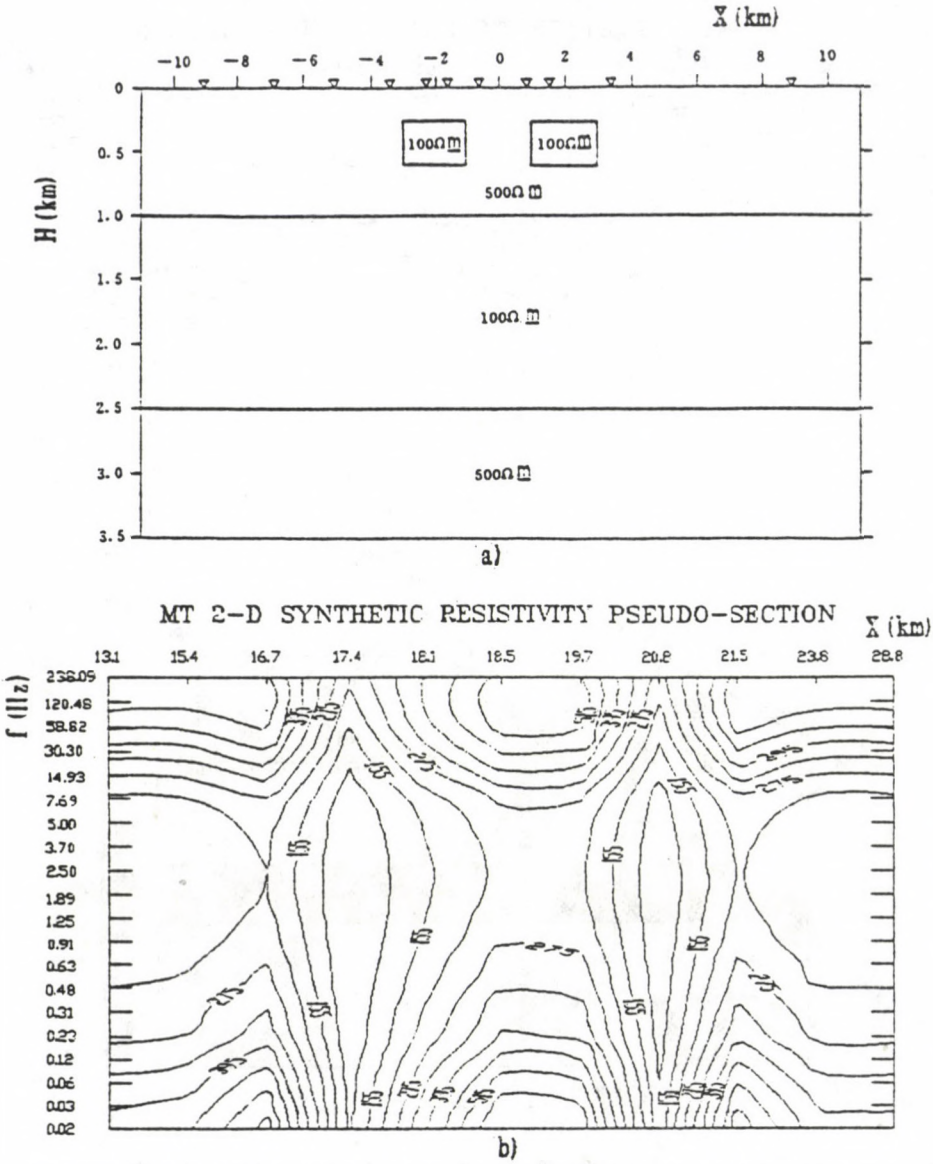

Fig. 2. a. Three layer section with two conductors of 100 Ωm. a) model, b) apparent resistivity ϱ_a pseudosection

Field example

The MT profile Z-K is almost in North-South direction at the southern front of the North China Platform, where the thickness of Quaternary is more than 500 m in average, its resistivity is about 10 Ωm and there is no significant lateral variation in them. The Da-bie mountain, where the resistive Precambrian stratum is widely outcropping, is near to the southern end of the MT profile. This area has undergone

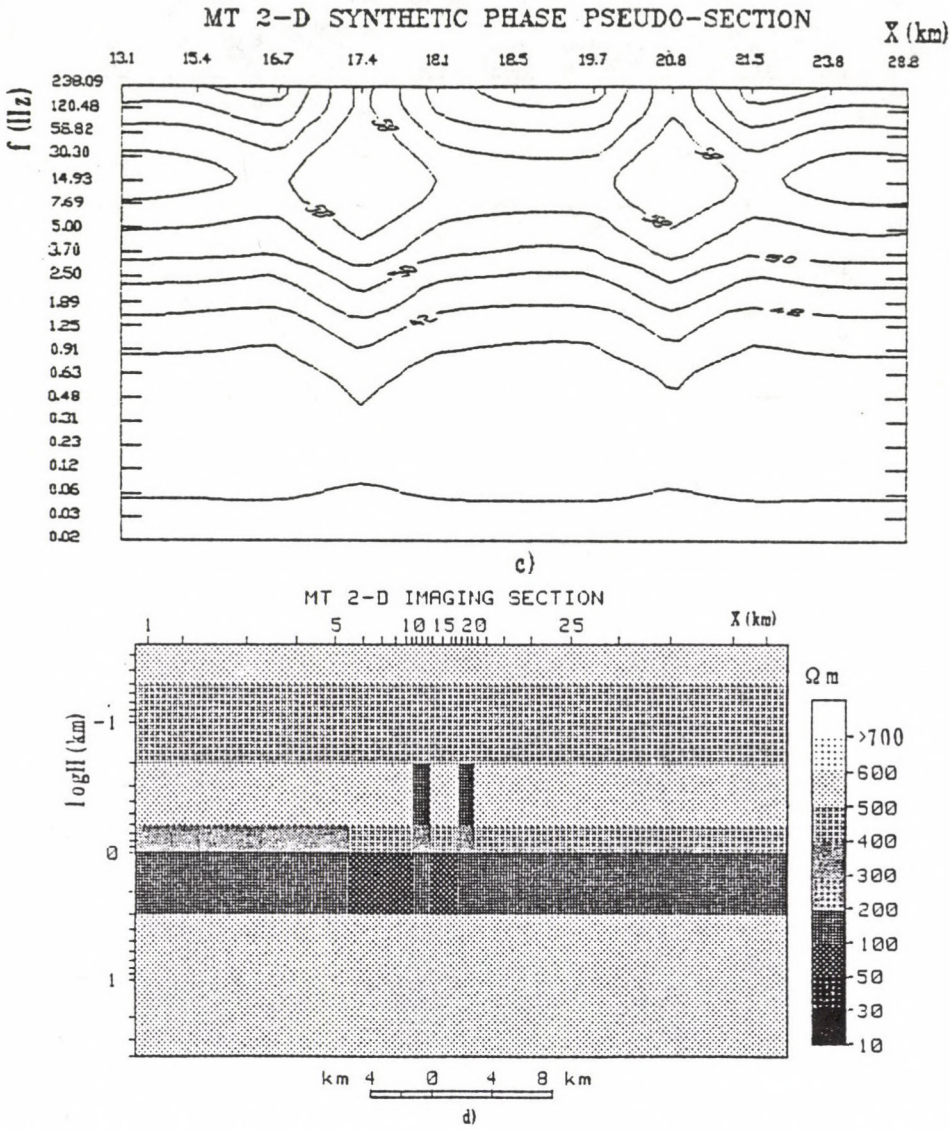

Fig. 2. b. Three layer section with two conductors of 100 Ωm. c) phase pseudosection, d) resulting imaging section

strong tectogenesis from the Indo-China movement time till the Late Yan-shan movement time, as a result, the geological structure is very complicated. In the area of the MT profile the strike of regional geological structure is North-West or North-West-West, approximately normal to the direction of the MT profile. The survey region is located at the joint zone of the North China Platform and of the Yang-zi Platform. Some geologists think that the Da-bie mountain including Proterozoic, Palaeozoic and Mesozoic formation is a block which is overthrust north in Indo-

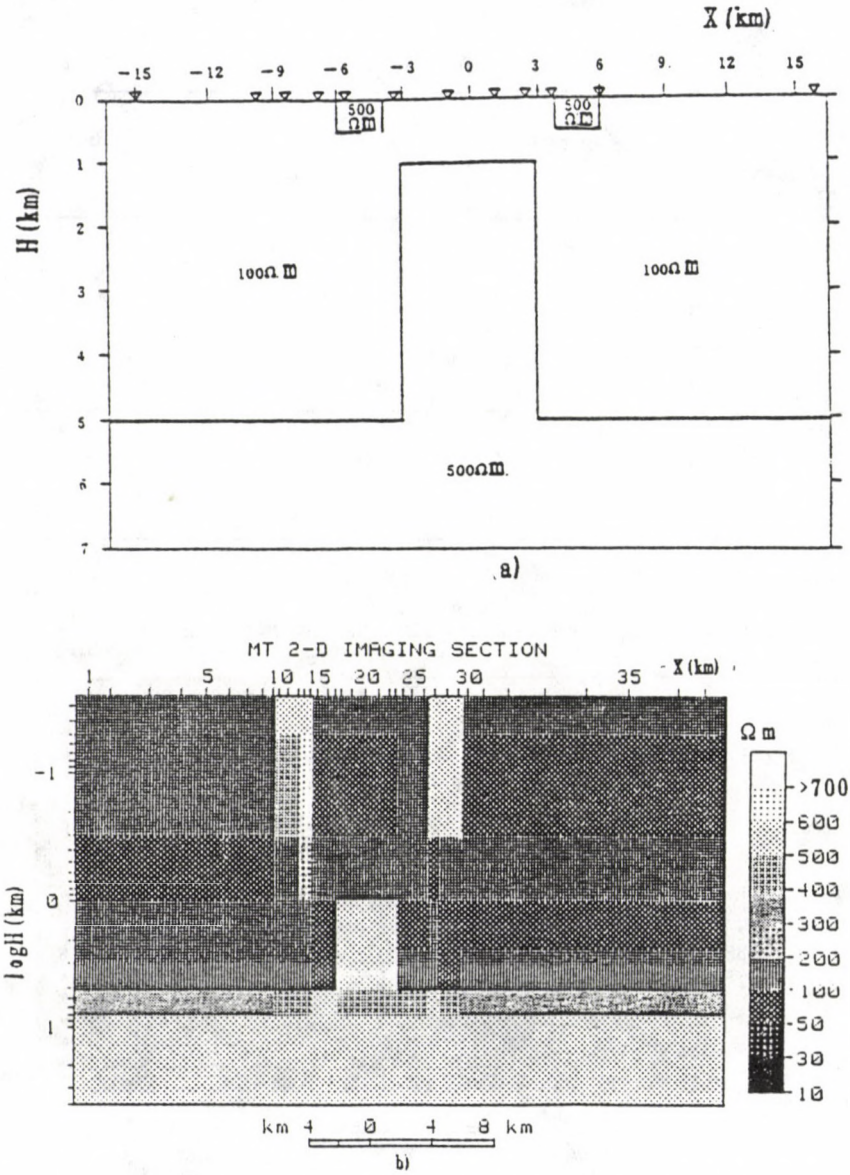

Fig. 3. A model considering of two resistive (500 Ωm) bodies and a horst. a) model, b) resulting imaging section

China movement time. It is very interesting for the study of regional tectonics and mineral resource exploration to investigate whether there exists an overthrust structure.

In the survey region, Cambrian and Proterozoic formations are characterized by high resistivity in the range 300 Ωm to 800 Ωm. There is no significant difference

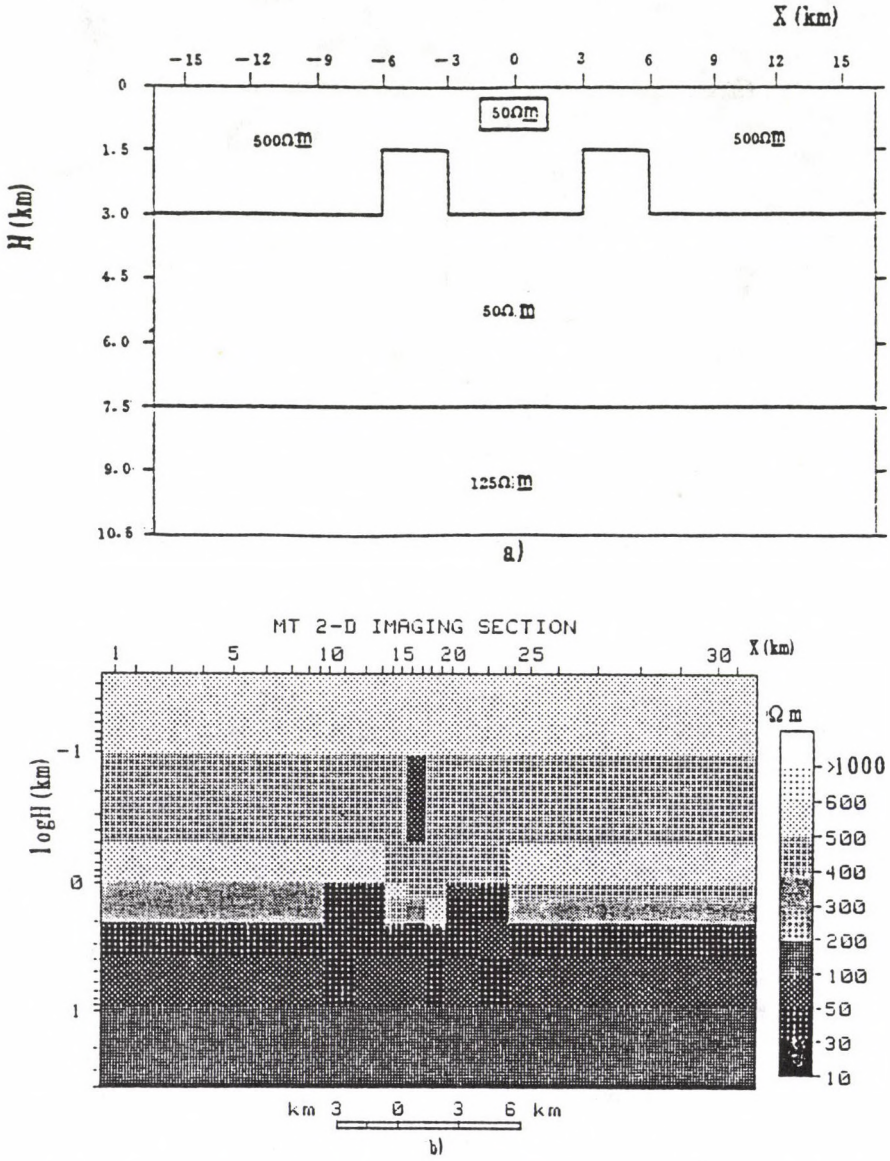

Fig. 4. A model consisting of two horst and a shallow anomalous body. a) model, b) resulting imaging section

between them, but there does exist a significant electrical interface between them and the upper younger formation, the resistivity of latter is some tens ohmmeters or even a few ohmmeters.

A segment of the MT profile is chosen for the test computation. The segment is of geological significance. It is in N-S direction and 70 km long. There are 13 MT

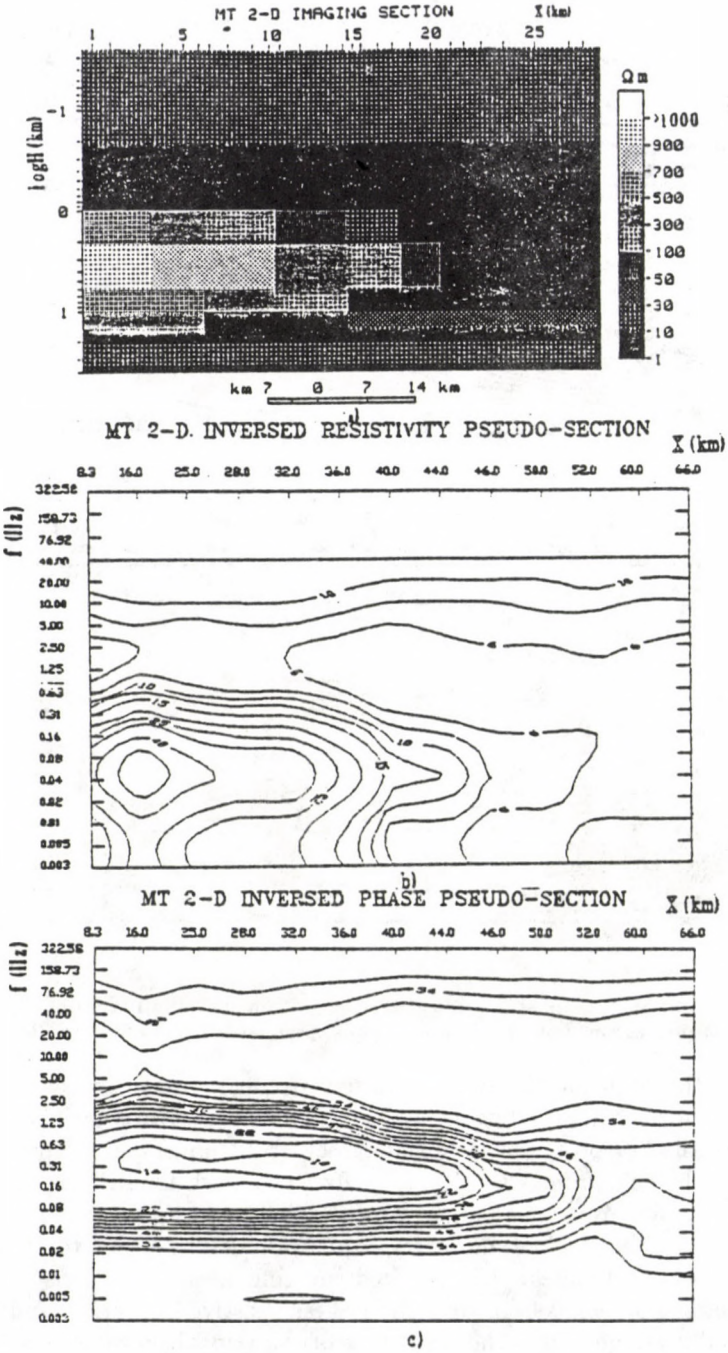

Fig. 5. a. Z-K MT profile at the southern front of North China Platform. a) resulting imaging section, b) apparent resistivity pseudosection of the imaging model, c) phase pseudosection of the imaging model

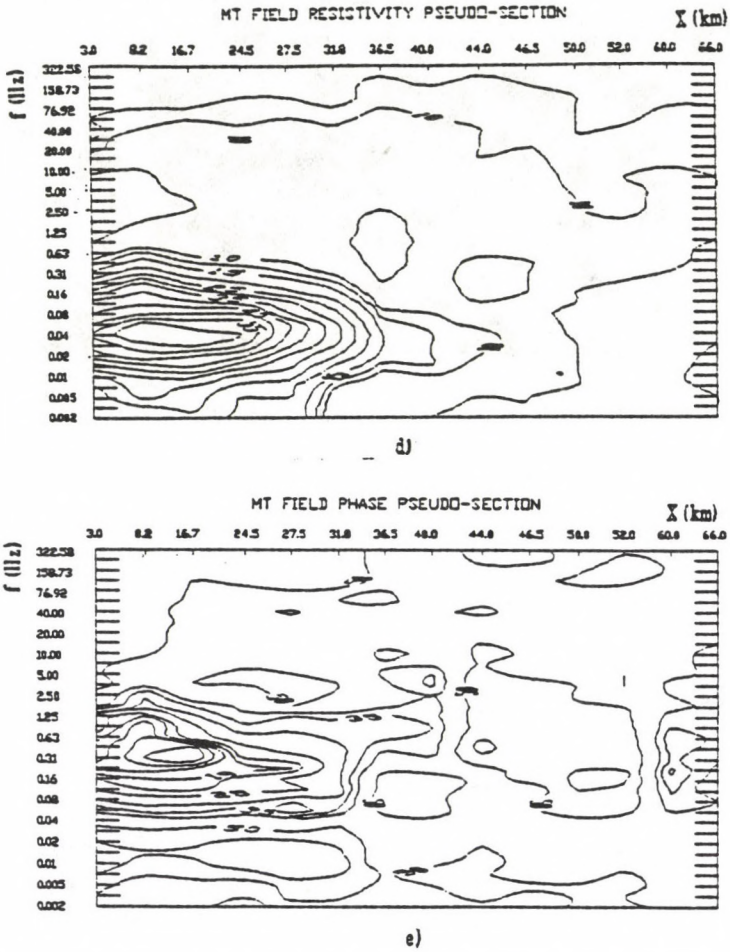

Fig. 5. b. Z-K MT profile at the southern front of North China Platform. d) MT apparent resistivity pseudosection of observed data, e) phase pseudosection of observed data

stations in this segment, the average station spacing is 5 km. On the basis of results of 1-D MT formal interpretation, a geoelectrical structure is delineated along the segment of the MT profile as an initial model for 2-D inversion. 18 frequencies were chosen which range from 320 to $2 \cdot 10^{-3}$ Hz. The mesh for FEM is 27×25. There are 56 blocks for inversion. The error of the observed data is given as 10 percent. After 13 iterations, the variance decreased from 123.9 percent to 16.4 percent, it takes 7 hours 25 minutes. The resulted imaging section (Fig. 5a) shows clearly that a resistive body has been wedged in a less resistive formation, and has thinned north, finally tapered out. The resistivity of the central, most resistive body is up to 800 Ωm, only older Palaeozoic and Proterozoic formation can have such a high resistivity. The resistivity of the underlain layer is only 10–20 Ωm. Apparently the

latter is younger Mesozoic, Cenozoic or even partly upper Palaeozoic formation. It could be inferred that there may be a large scale overthrust structure of gentle slope at the southern front of the North China Platform. The apparent resistivity and phase pseudosection of the resulting model from inversion and observed data are shown in Fig. 5b–e. It is easy to see from those that the appearances of anomalies and the values of isolines are similar or approximate to each other between the corresponding figures, this means that the fit to the original data has been basically performed.

5. Discussion

The results of tests with synthetic data show that the effectiveness of application for E polarization is not as good as for B polarization, since the iterative procedure converged slowly or even diverged. To search the reason for that theoretically, the Frechet derivatives of surface impedance and components of electromagnetic field have been analysed, by which we can measure the variation of surface response caused by variation of the electrical parameters of the substratum, thereby it is termed sensitivity function. It is a useful tool for studying the resolution of electrical and electromagnetic survey (Sasaki 1989). The results show that the Frechet derivative is a product of the gradient of the Green function $G^H(r, r')$ and the magnetic field intensity for B polarization while it is a product of $G^E(r, r')$ and $E(r')$ for E polarization. Therefore, there is a higher resolution for B polarization than for E polarization, this may be the basic reason for the lower effectiveness of application of the GPST technique to E polarization with respect to B polarization. To improve the effectiveness of the GPST technique to 2-D MT imaging, we intend to introduce the idea of removing singular eigenvalues to increase the resolution for E polarization (Lowery et al. 1989).

Acknowledgements

We are grateful for helps from Professor Zeng Hua lin, Associate Professor Deng Qian hui and Hu Jian de.

The project was supported by National Natural Science Foundation of China.

References

Chen Y M 1985: *Geophysics*, 50, 1664–1675.

Li Jian hua, Tang Long ji, Ou Ya fei 1987: *J. Numerical Methods and Computer Applications*, 1, 35–43.

Liu Guo dong, Chen Le shou 1984: Magnetotelluric Studies. Seismological Press, Beijing

Lowery T, Allen M B, Shive P N 1989: *Geophysics*, 54, 766–774.

Sasaki Y 1989: *Geophysical Prospecting*, 37, 395–406.

Tikhonov A N, Arsenin V Y 1977: Solution of ill-posed problems. John Wiley and Sons, New York

Xie Gang quan, Li Jian hua 1988: *Scientia Sinica*, 4, 359–365.

Yang Wen cai 1989: Geophysical inversion and seismic tomography (in chinese). Geological Press, Beijing

Acta Geod. Geoph. Hung., Vol. 29(1–2), pp. 251–254 (1994)

ASYMPTOTIC VARIANCE OF THE MOST FREQUENT VALUE AND OF THE DIHESION

B HAJAGOS[1]

[Manuscript received July 8, 1994]

The paper gives general formulas for the asymptotic scatter of the most frequent value (up to now, formulas were only known for special cases, e.g. for symmetric distributions, as well as for the asymmetric U-distribution).

The two theses given in the paper prove the finiteness of the asymptotic scatter of the most frequent value for cases with often valid premissas, thus these theses have evident significance for the practical use in geophysics and in other fields for the MFV-methods.

Keywords: asymptotic variance; dihesion; most frequent value; robust statistics

1. The formulas used (after Huber 1981)

The most frequently used form of the robust statistics defines the T location- and S scale parameters of a distribution F by the equations

$$\int \psi\left[\frac{x-T(F)}{S(F)}\right]F(dx)=0 \quad \text{and} \quad \int \chi\left[\frac{x-T(F)}{S(F)}\right]F(dx)=0 \tag{1}$$

where $\psi(y)$ is generally an odd and $\chi(y)$ an even function (see e.g. Huber 1981, p. 136).

The asymptotic variance of the location- and scale-parameters are the variances of the corresponding influence-functions:

$$A^2(F,T)=E[IC^2(x;F,T)] \quad \text{and} \quad A^2(F,S)=E[IC^2(x;F,S)]$$

(as the expected value of the IC-influence is 0, see Huber 1981, p. 14).

Using the definition of the influence, let us substitute in Eq. (1) $F \sim F_t = (1-t)F + t\delta_x$. Following derivation after t, let $t=0$. Thus the following set of equations is obtained:

$$\begin{aligned} IC(x;F,T)\int\psi'(y)F(dx)+IC(x;F,S)\int\psi'(y)yF(dx)&=\psi(y)S(F)\\ IC(x;F,T)\int\chi'(y)F(dx)+IC(x;F,S)\int\chi'(y)yF(dx)&=\chi(y)S(F)\end{aligned} \tag{2}$$

where $y=\dfrac{x-T(F)}{S(F)}$ (Huber 1981, p. 136).

[1]Geophysical Institute, University Miskolc, H-3515 Miskolc, Egyetemváros, Hungary

1217-8977/94/$ 4.00

2. General formulas and laws for the asymptotic variances of the most frequent value and of the dihesion

$$\psi(y) = \frac{k^2 y}{k^2 + y^2} \quad \text{and} \quad \chi(y) = \frac{(a+1)y^2 - 1}{(1+y^2)^2} \tag{3}$$

where $y = \frac{x-M}{\varepsilon}$, then the location parameter of the distribution F is the most frequent value M, its scale parameter $S = k\varepsilon$, with ε, the generalized dihesion (Steiner 1988 and 1991).

Let us introduce the following integrals:

$$S_{ik} = \int \left(\frac{k^2}{k^2+y^2}\right)^i F(dx); \quad D_{ik} = \int y \left(\frac{k^2}{k^2+y^2}\right)^i F(dx);$$

$$E_{ik} = \int y \frac{k^2}{k^2+y^2} \left(\frac{1}{1+y^2}\right)^{i-1} F(dx)$$

Thus the equations defining M and ε obtained from Eq. (1) can be written in the following form:

$$D_{1k} = 0 \quad \text{and} \quad (a+1)S_{11} - (a+2)S_{21} = 0 \tag{4}$$

and the coefficients of Eq. (2) result as:

$$\begin{aligned} B_{11} &= 2S_{2k} - S_{1k} & B_{12} &= 2D_{2k} \\ B_{21} &= 2[(2a+4)D_{31} - (a+1)D_{21}] & B_{22} &= 2[(2a+3)S_{21} - (2a+4)S_{31}] \end{aligned}$$

the determinant of this system of equations is: $D = B_{11}B_{22} - B_{12}B_{21}$.

If $D \neq 0$, then the solution is:

$$IC(x; F, T) = \frac{\varepsilon}{D}(\psi(y)B_{22} - \chi(y)B_{21}), \quad IC(x; F, S) = \frac{\varepsilon}{D}(\chi(y)B_{11} - \psi(y)B_{12})$$

For the calculation of the expected value of the square of these expressions the following notations are introduced:

$$G_1 = S_{1k} - S_{2k}, \quad G_2 = (a+1)^2 S_{21} - 2(a+1)(a+2)S_{31} + (a+2)^2 S_{41}$$

further

$$G_3 = (a+1)E_{2k} - (a+2)E_{3k}$$

and the variances have the following form:

$$\begin{aligned} A^2(M) &= \frac{k^2\varepsilon^2}{D^2}(B_{22}^2 G_1 + B_{12}^2 G_2 - 2B_{22}B_{12}G_3) \\ A^2(\varepsilon) &= \frac{\varepsilon^2}{D^2}(B_{21}^2 G_1 + B_{11}^2 G_2 - 2B_{21}B_{11}G_3) \end{aligned} \tag{5}$$

Usually the pair of functions ψ, χ is given as

1. $a = 2$ and k arbitrary (e.g. 1, 2 or 3)
2. $k = 1$ and a arbitrary ($a > 1$).

In the special case, when the distribution F is symmetric, $\psi(y)$ is an odd and $\chi(y)$ an even function, the first equation of the system Eq. (1) is fulfilled for the symmetry point for all $S > 0$ and the second equation may have several solutions, too.

It is advantageous to define the scale parameter as the maximum solution S.

As both $\psi'(y)y$ and $\chi'(y)$ are odd functions, $IC(x; F, T)$ can be expressed from the first equation, $IC(x; F, S)$ from the second, independently of each other, i.e. $A^2(F, T)$ is independent of $A^2(F, S)$, but it depends on S.

In the present case (Steiner 1991) if the distribution F is symmetrical, then:

$$A^2(M) = k^2\varepsilon^2 \frac{G_1}{B_{11}^2} = k^2\varepsilon^2 \frac{S_{1k} - S_{2k}}{(2S_{2k} - S_{1k})^2} .$$

That means that the asymptotic variance of M is finite if $2S_{2k} - S_{1k} \neq 0$.

Theorem 1: If in the case of a symmetric distribution the scale parameter $S = k\varepsilon$ is chosen sufficiently large to get the expected value of the weights $\dfrac{S^2}{S^2 + (x - M)^2}$ as $S_{1k} > 1/2$, then $2S_{2k} - S_{1k} > 0$, and $A^2(M)$ is finite.

Namely for all functions $0 \leq h(x) \leq 1$

$$\int h^2(x)F(dx) \leq \int h(x)F(dx) \quad \text{as} \quad h^2(x) \leq h(x)$$

and Schwartz's inequality is fulfilled:

$$\int h^2(x)F(dx) = \int h(x)F(dx) \int F(dx) \geq [\int h(x)F(dx)]^2 .$$

If $h(x) = \dfrac{S^2}{S^2 + (k + M)^2}$ then

$$S_{1k}^2 \leq S_{2k} \leq S_{1k} .$$

Let us multiply the inequalities by 2 and subtract S_{1k}:

$$2S_{1k}^2 - S_{1k} \leq 2S_{2k} - S_{1k} \leq S_{1k} .$$

It follows that $2S_{2k} - S_{1k} > 0$, if $S_{1k} > \frac{1}{2}$.

Theorem 2: If the distribution F is continuous and its density function is $f(x)$ is unimodal, then:

$$2S_{2k} - S_{1k} = \int_{-\infty}^{\infty} \frac{k^2(k^2 - y^2)}{(k^2 + y^2)^2} f(x)dx > 0 \tag{6}$$

and $A^2(M)$ is finite.

In the proof the symmetry point is chosen at $M = 0$ without a loss of generality, and be $\varepsilon > 0$, $y = ku = \frac{x}{\varepsilon}$, $S = k\varepsilon$. In this case the integral in Eq. (6) is using the symmetry:

$$S \int_{-\infty}^{\infty} \frac{1-u^2}{(1+u^2)^2} f(Su)du = 2S \int_{0}^{\infty} \frac{1-u^2}{(1+u^2)^2} f(Su)du \ .$$

Let us split this integral into two parts:

$$\int_{0}^{1} \frac{1-u^2}{(1+u^2)^2} f(Su)du + \int_{1}^{\infty} \frac{1-u^2}{(1+u^2)^2} f(Su)du$$

with substituting in the latter integral $u \sim \frac{1}{u}$, the two integrals can be added:

$$\int_{0}^{1} \frac{1-u^2}{(1+u^2)^2} \left(f(Su) - f\left(\frac{S}{u}\right) \right) du \ .$$

This integral is positive, as in the interval $0 \leq u \leq 1 \quad 1 - u^2 \geq 0$ holds and $f(Su) \geq f(\frac{S}{u})$, if $f(x)$ is unimodal.

Q.e.d.

Let us observe that in the proof of the theorems, the fulfilment of the second condition of Eq. (4) has not been used. However, if ε is also determined from the measurement results according to the second condition in Eq. (4), then it is necessary that the asymptotic variance

$$A^2(\varepsilon) = \varepsilon^2 \frac{G_2}{B_{22}^2} = \varepsilon^2 \frac{(a+1)^2 S_{21} - 2(a+1)(a+2)S_{31} - (a+2)^2 S_{41}}{4[(2a+3)S_{21} - (2a+4)S_{31}]^2}$$

should be finite, i.e. $B_{22} \neq 0$. Using the condition given in Eq. (4), $A^2(\varepsilon)$ is obtained as:

$$A^2(\varepsilon) = \varepsilon^2 \frac{S_{21}^3 + S_{11}^2 S_{41} - 2S_{11}S_{21}S_{31}}{4(S_{21}^2 + S_{11}S_{21} - 2S_{11}S_{31})^2}$$

too. This has evidently the advantage in numerical computations that based on the values M and ε determined from the data, empiric values of the S_{1k}-s can be determined, too.

References

Huber J 1981: Robust statistics, Wiley, New York

Steiner F 1988: *Geophysical Transactions*, Special issue, 34, 139–260.

Steiner F (Ed.) 1991: The most frequent value. Akadémiai Kiadó, Budapest

Acta Geod. Geoph. Hung., Vol. 29(1–2), pp. 255–257 (1994)

A THESIS ON THE DIFFERENCE OF THE DETERMINATION METHODS OF THE SCALE PARAMETER (CML $\neq$ MFV)

L CSERNYÁK[1]

[Manuscript received July 8, 1994]

It is shown that from among symmetric distributions, only the normal distribution has the property that if used as substituting distribution, it gives the same formula for the scale parameter when the estimations are used after the maximum likelihood method and according to the most frequent value (MFV). That means that the "CML-estimation" in the monograph by Andrews et al. (1972) is not identical with the MFV-estimation (in spite of identical ψ-functions), the resistance of the latter (insensitivity against outliers) is much higher and correspondingly its use is more advantageous in geophysics and in other fields of science.

Keywords: density functions; location parameter; normal distribution; scale parameter

It is shown that among symmetric distributions, only the normal distribution has the property that used as substituting distribution, the same scale parameter is obtained using both the maximum likelihood estimation and a robust estimation method.

Be g a twice differentiable density function. If the scale and location parameters are estimated with the help of g, using the maximum likelihood (ML) method for a random variable ξ with the density function f, then the location of the maximum of the function

$$L(S,T) = \int_{-\infty}^{\infty} \ln\left[\frac{1}{S} g\left(\frac{x-T}{S}\right)\right] f(x)dx$$

is to be found (T is the location and S the scale parameter). That means in case of a given T that the scale parameter is the solution of the equation

$$\int_{-\infty}^{\infty} \left[-\frac{1}{S} - \frac{x-T}{S^2} g'\left(\frac{x-T}{S}\right) \frac{1}{g(\frac{x-T}{S})}\right] f(x)dx = 0\,. \tag{1}$$

(Here and in the following is $g'(x) = \dfrac{dg(x)}{dx}$.)

[1]Institute of Mathematics, University Miskolc, H-3515 Miskolc, Egyetemváros; Highschool for Finances and Accounting, H-1149 Budapest, Buzogány u. 10–12, Hungary

1217-8977/94/$ 4.00

If the estimation is made using the robust method proposed by Steiner (1991), then the scale parameter is obtained for a given location parameter from the following equation:

$$\int_{-\infty}^{\infty} \frac{\partial^2 g(\frac{x-T}{S})}{\partial T^2} \frac{1}{g(\frac{x-T}{S})} f(x)dx = 0 \tag{2}$$

which can be written in the form

$$\frac{1}{S^2} \int_{-\infty}^{\infty} g''\left(\frac{x-T}{S}\right) \frac{1}{g(\frac{x-T}{S})} f(x)dx = 0 .$$

It can be easily seen that if g is the density function of the standard normal distribution, then Eqs (1) and (2) give the well-known scatter, i.e. both methods result in the same scale parameter.

Now the problem emerges whether there is any other g in addition to that of the normal distribution which has the same property. For even function, the answer is negative.

THEOREM: Be g a twice differentiable, even standard density function. If Eqs (1) and (2) yield for a fixed value of T for all density functions f the same scale parameter, then g is the density function of the standard normal distribution.

Proof: We shall prove somewhat more than the theorem: it is sufficient to discuss even (symmetrical) density functions f. It can be supposed that $T = 0$. By introducing in Eqs (1) and (2) the transformation $x = St$, the following equations result:

$$\frac{1}{S} \int_{-\infty}^{\infty} \left[-1 - tg'(t)\frac{1}{g(t)}\right] f(St)dt = 0 , \tag{3}$$

$$\frac{1}{S} \int_{-\infty}^{\infty} g''(t)\frac{1}{g(t)} f(St)dt = 0 . \tag{4}$$

The solutions of Eqs (3) and (4) coincide for all f-s if the multiplier of f in the two equations differ only by a constant factor, i.e. if

$$-1 - tg'(t)\frac{1}{g(t)} = \lambda g''(t)\frac{1}{g(t)} .$$

Following some rearrangement we obtain the differential equation

$$\lambda g''(t) + tg'(t) + g(t) = 0 . \tag{5}$$

a) Be $\lambda > 0$. Using the transformation $u\sqrt{\lambda} = t$, and introducing the notation $g(u\sqrt{\lambda}) = y(u)$, we get from Eq. (5):

$$y''(u) + uy'(u) + y(u) = 0 . \tag{6a}$$

One of the solutions of this equation is $y_1(u) = e^{-\frac{u^2}{2}}$ (in this case g is the density function of the normal distribution).

By using the transformation $y(u) = z(u)e^{-\frac{u^2}{2}}$, the equation

$$z'' - z'u = 0$$

is obtained from Eq. (6), and from this $z'(u) = Ge^{\frac{u^2}{2}}$. Thus the other solution of Eq. (6) is:

$$y_2(u) = z(u)e^{-\frac{u^2}{2}}\ , \quad \text{where} \quad z'(u) = e^{\frac{u^2}{2}}\ .$$

It can be easily seen that y_2 is an odd function disregarding an additive constant, thus it cannot be density function.

It should be remarked that $\lim_{\infty} u \cdot y_2(u) = 1$ (it can be easily shown using the L'Hospital-rule), therefore y_2 tends to zero only in the order of magnitude of $1/u$, that is a further proof that it cannot be density function.

According to these, the general solution of Eq. (6),

$$y(u) = C_1 y_1(u) + C_2 y_2(u)$$

is a density function only if $C_2 = 0$.

b) If $\lambda < 0$, then $-\lambda = \lambda^* > 0$ is introduced.

Following the procedure used in a), we get the equation

$$y''(u) - uy'(u) - y(u) = 0 \tag{6b}$$

with one of the solutions $y_1(u) = e^{\frac{u^2}{2}}$. Following further the steps in a), we get the equation

$$z'' + z'u = 0\ .$$

From this, $z'(u) = K_1 e^{-\frac{u^2}{2}}$, i.e. $z(u) = K_1\Phi(u) + K_2$ (where Φ is the density function of the standard normal distribution).

Thus the general solution of Eq. (6b) is

$$y(u) = e^{\frac{u^2}{2}}(K_1\Phi(u) + K_2)\ .$$

This y cannot be a density function, as

$$\lim_{\infty}(K_1\Phi(u) + K_2) = K_1 + K_2\ ,$$
$$\lim_{-\infty}(K_1\Phi(u) + K_2) = K_2\ ,$$

and therefore y can disappear at $\pm\infty$, only if $K_1 = K_2 = 0$.

c) In the case $\lambda = 0$ it follows from Eq. (5) that $g(t) = c/t$ which cannot be a density function.

The theorem is so proven.

Reference

Steiner F (ed.) 1991: The Most Frequent Value. Akadémiai Kiadó, Budapest

Acta Geod. Geoph. Hung., Vol. 29(1–2), pp. 259–272 (1994)

SOME PROPERTIES OF THE MFV-CALCULATIONS TESTING IN CASE OF $a \to 1$ FOR THE SUPERMODEL $f_a(x)$

F STEINER[1]

[Manuscript received November 14, 1994]

The paper shows that the most frequent value (MFV-) procedures (based on P_k-norms) essentially differ from other, often used reweighted least square algorithms, therefore it would be misleading to regard the latter ones as algorithms with similar properties: the similarity is limited only to the computing techniques. Inversion results (based on the minimization of the P_k-norms) can have negligible error even if more than 50 percent of the data are asymmetrically lying outliers. The location of the densest data is to be accurately determined with the MFV-procedure even if only some percent of data belongs to this interval characterized by maximum probability densities.

Keywords: bias; outliers; resistance; statistical efficiency; statistical norms

Introduction

The $f_a(x)$ supermodel is defined by the standard densities

$$f_a(x) = n(a) \cdot (1 + x^2)^{-a/2} \tag{1}$$

primarily for the open interval $1 < a < \infty$ (see *e.g.* Steiner 1991).

The formula for the norming factor $n(a)$ is the following:

$$n(a) = \frac{\Gamma[a/2]}{\sqrt{\pi} \cdot \Gamma[(a-1)/2]} \,. \tag{2a}$$

In case of $a \to \infty$ (if the variance is constant), $f_a(x)$ tends to the Gaussian density function, its standard form being $f_G(x) = (2\pi)^{-1/2} \exp(-x^2/2)$, and in this sense this type of probability distribution (as limiting case) can be justified considered as a type-member of the supermodel $f_a(x)$. On the contrary, $a = 1$ does not define a probability density function because the norming factor written in the form (using the equation $\Gamma[(a+1)/2] = [(a-1)/2] \cdot \Gamma[(a-1)/2]$)

$$n(a) = \frac{(a-1) \cdot \Gamma[a/2]}{2 \cdot \sqrt{\pi} \cdot \Gamma[(a+1)/2]} \tag{2b}$$

[1]Department of Geophysics, University of Miskolc, H-3515 Miskolc, Egyetemváros, Hungary

obviously shows that $n(a)$ tends to zero if $a \to 1$. Equation (1) immediately indicates that it does not define in case of $a = 1$ a probability density function as the integral $\int_{-\infty}^{\infty} 1/\sqrt{1+x^2}dx$ is divergent.)

In spite of the foregoings the function

$$f(x) = \frac{1}{\sqrt{1+x^2}} \tag{3}$$

is worth investigating as the $f(x)$-values (apart from a constant factor) are arbitraryly near to the $f_a(x)$-values on an arbitraryly long interval about the symmetry point if $a = 1 + \delta$ and δ are sufficiently small. Speaking more mathematically, we can even denote $f(x)$ as limit function of the functions belonging to the $f_a(x)$ family in case of $a \to 1$ in the sense that there exists a monotonous series (c_n) of numbers fulfilling

$$\lim_{n\to\infty} c_n \cdot f_{a_n}(x) = f(x) \qquad (-\infty < x < \infty) \tag{3a}$$

and obviously $c_n = 1/n(a_n)$ is a series with this property if $\lim_{n\to\infty} a_n = 1$. These investigations will turn out to be also instructive for the special properties and advantages of the most frequent value calculation itself.

1. Similarities between $f(x)$ and the Cauchy density function

There are many possible definitions of similarity but we discuss only two of them, namely which are most useful to our purposes.

1.1 Similarity around the maximum $f(x)$-value ($f_{\max}$)

The most simple kind of fitting is shown in Fig. 1 by the continuous and the dashed lines. The first line shows the shape of the const$/\sqrt{1+x^2}$ = const $\cdot f(x)$ function where the (otherwise arbitrary) constant factor was actually chosen to $(\pi \cdot \sqrt{3})^{-1} = 0.1838$. In case of this choice the values of the Cauchy probability density function $f_C(x)$ are in the interval $(-\sqrt{3}, +\sqrt{3}$, *i.e.*, in the interval of the greatest $f(x)$-values (greater than $f_{\max}/2$) very near to the const $\cdot f(x)$ values (if the Cauchy distribution is also symmetrical to the origin and the parameter of scale S equals $\sqrt{3}$). In other words, the "fashion" of the concentration hardly varies from $a = 2$ (this is the Cauchy type) to a-values arbitraryly near to unity.

On the contrary, the flanks are ver different: *e.g.*, already at $x = \pm 3.5 \cdot \sqrt{3}$ $f_C(x)$ is less than a half of the const $\cdot f(x)$-value (we have seen that at $x = \pm\sqrt{3}$ an equality holds: $f_C(\pm\sqrt{3}) = 0.1838 \cdot f(\pm\sqrt{3})$). That means that the simplest comparison belonging the similarity (made in the foregoings) has only taken into consideration the central parts of the functions, — and unquestionably this was the best way to make conclusions for the "fashion" of the accumulation of the data around the symmetry point (*i.e.*, around the parameter of location, denoted by T). — It is useful, however, to measure the similarity in a more sophisticated (and less arbitrary) way, too.

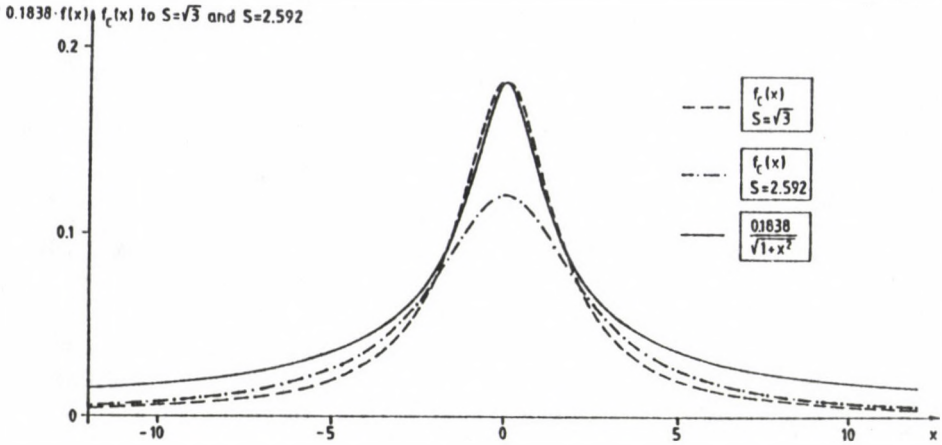

Fig. 1. The curve of the function $f(x)/\pi \cdot \sqrt{3}$ (continuous line) and two Cauchy density functions most similar to it, in one case in the sense of the treatments given in point 1.1. (– – –), in the other case in the sense discussed in point 1.2. (– · – · –)

1.2 A possible definition of the width for functions which have positive values for all x-es in the interval $(-\infty,\ +\infty)$

In case of the density function of the Cauchy distribution

$$f_C(x) = \frac{1}{\pi}\frac{S}{S^2 + (x-T)^2} \tag{4}$$

it lies on hand to accept $2S$ as the width of $f_C(x)$ due to at least three reasons:

a) the interquartile range of $f_C(x)$ equals $2S$ (as $\int\limits_{T-S}^{T+S} f_C(x)dx = 1/2$);

b) outside of the interval $(T-S,\ T+S)$ the $f_C(x)$-values are less than the half of the maximum value of $f_C(x)$ (*i.e.*, of $f_C(T)$), inside of it $f_C(x)$ is greater than $f_C(T)/2$, and this interval has obviously the length $2S$;

c) the dihesion ε as a measure of the dispersion of randomly distributed data (see *e.g.* Steiner 1991) has the value in case of Cauchy-distributed data exactly S, and therefore the length of the interval $(T-\varepsilon,\ T+\varepsilon)$ is also $2S$.

Consequently, we are almost forced by these coincidences to accept $2S$ as the width of $f_C(x)$ if no special requirements of other type are to be taken into consideration.

In general case we can choose to a quadratically integrable and positive $\varphi(x)$ (> 0 in the interval $-\infty < x < \infty$) function that $f_C(x)$-function which is most similar to it and we define the $2S$-value of this $f_C(x)$ function as the width of the $\varphi(x)$ function itself. Let the expression be accepted as a measure of similarity

(according to Csernyák (1991) Eqs (2–3)) the expression

$$W_{\varphi(x)} = \frac{\int\limits_{-\infty}^{\infty} f_C(x) \cdot \varphi(x) dx}{\sqrt{\int\limits_{-\infty}^{\infty} f_C^2(x) dx \int\limits_{-\infty}^{\infty} \varphi^2(x) dx}} . \tag{5}$$

Constant factors play obviously not any role in this expression, the "similarity" is therefore characterized on ground of the "fashions" of the x-dependencies of both functions (neglecting "amplitudes"). Consequently, we can calculate without any difficulty the width of $f(x)$ (given in Eq. (3)), too. (The meaning of this width is according to the foregoings the width of $f_a(x)$ in case of $a \to 1$.) Accordingly, we have to determine the maximum place of

$$W_{f(x)} = \frac{\int\limits_{-\infty}^{\infty} \frac{1}{x^2 + S^2} \cdot \frac{1}{\sqrt{1 + x^2}} dx}{\sqrt{\int\limits_{-\infty}^{\infty} \frac{1}{(x^2 + S^2)^2} dx \cdot \int\limits_{-\infty}^{\infty} \frac{1}{1 + x^2} dx}} . \tag{6a}$$

All integrals figuring in this expression can be given analitically (see *e.g.* the "classical" Gröbner and Hofreiter 1958) and therefore finally only the maximum place of the function

$$\sqrt{\frac{S}{S^2 - 1}} \cdot \ln(S + \sqrt{S^2 - 1}) \tag{6b}$$

is to be determined, resulting in the value

$$S = 2.59204005. \tag{7}$$

The corresponding $f_C(x)$ is to be seen also in Fig. 1 (– · – · –) which is obviously influenced also by the flanks and in this sense it is more similar to const · $f(x)$ as $f_C(x)$ calculated to $S = \sqrt{3}$ (– – –); let us remember that the latter takes only the central part of $f(x)$ into consideration. In spite of all that, the width of $f(x)$ is not too much: according to Eq. (7) it is evidently $2 \cdot 2.59204 = 5.18408$ which value is only 50 percent more than $2 \cdot \sqrt{3} = 3.46410$.

2. Efficiency of MFV-calculations for the supermodel $f_a(x)$ if $a \to 1$

The S-value given in Eq. (7) is the so-called "dihesion" ε of the standard $f_a(x)$ in case of $a \to 1$: on the one hand, Csernyák (1991) proved (see loc. cit. in connection with Eq. (5)) that the (for an arbitrary density function $\varphi(x)$) primarily otherwise defined "dihesion" ε is really the maximum place of the expression given in Eq. (5), and, on the other hand, the function $f(x) = 1/\sqrt{1 + x^2}$ figuring in Eq. (6a) is the limiting function for the $f_a(x)$-supermodel in case of $a \to 1$ in the sense which was defined earlier (see Eq. (3a)).

It can be easily proven (*e.g.* using the maximum likelihood principle) that for $1 < a < \infty$ the best statistical method for determining the location parameter T is the most frequent value- (MFV-) calculation made by the parameter of scale $S = k(a) \cdot \varepsilon$ where $k(a)$ is the reciprocal value of the dihesion of the standard $f_a(x)$ distribution, *i.e.*, of ε_{st} (Steiner 1991). Some well known values of k are: $k(9) = 3$; $k(5) = 2$ and for $a = 2$ (Cauchy-distribution) k equals the unity. Let be supposed that the best k-value is also for the limiting function $f(x)$ the reciprocal of the value given in Eq. (7), *i.e.*, $k(1) = 0.3857965$.

In column 8 of the Table at the end of Steiner's book (1991) are the formulas for the asymptotic variances given for $k = 2$ and $k = 3$; the generalization for an arbitrary k can be trivially made. As in our case $k\varepsilon = 1$ holds, the limit of the asymptotic variance (the square of the asymptotic scatter A) concerning the functions $c_n \cdot f_{a_n}(x) = (1 + x^2)^{-a/2}$ can be written as

$$A^2 = \frac{S_1 - S_2}{[2S_2 - S_1]^2} \tag{8a}$$

where

$$S_1 = \lim_{a_n \to 1} \int_{-\infty}^{\infty} \frac{1}{1+x^2} \frac{1}{(1+x^2)^{a_n/2}} dx = \int_{-\infty}^{\infty} \frac{1}{1+x^2} \frac{1}{\sqrt{1+x^2}} dx = 2 \tag{8b}$$

and

$$S_2 = \lim_{a_n \to 1} \int_{-\infty}^{\infty} \frac{1}{(1+x^2)^2} \frac{1}{(1+x^2)^{a_n/2}} dx = \int_{-\infty}^{\infty} \frac{1}{(1+x^2)^2} \frac{1}{\sqrt{1+x^2}} dx = \frac{4}{3} \tag{8c}$$

(concerning the numerical values see once more Gröbner and Hofreiter 1958). With these values (and using Eq. (2b)) we have for A^2 the following:

$$A^2 = \frac{3}{2} \,. \tag{8d}$$

The definition of the statistical efficiency e is (in percent)

$$e = \frac{A^2_{\min}}{A^2} \cdot 100 \ \% \,, \tag{9}$$

where the minimum variance of the statistical estimates $A^2_{\min}$ (to a given probability distribution) is formally the reciprocal value of the Fisher-information (see *e.g.* Cramér 1945); A^2 denotes the actual variance (*i.e.*, for the actually used statistical procedure and naturally for the same distribution). The very important practical meaning of e is the following: how many percent of data would be enough to the same accuracy if we would apply the (to the probability distribution) optimal method instead of the actually used one.

As for the c_n times of the standard $f_a(x)$ distributions

$$A^2_{\min} = \frac{a+2}{a(a-1)} \cdot \frac{1}{n(a_n)} \tag{10}$$

holds (see *e.g.* Steiner 1988), the statistical efficiency e (Eq. (9)) in case of $k = 0.3857965$ and for $a_n \to 1$ turn out to be (regarding Eqs (8d) and (10))

$$\begin{aligned} e &= \lim_{a_n \to 1} 100 \frac{a_n+2}{a_n(a_n-1)} \cdot \frac{1}{n(a_n)} \cdot \frac{2}{3} = \\ &= \lim_{a \to 1} 100 \frac{(a_n-1)\cdot\Gamma[a_n/2]}{2\cdot\sqrt{\pi}\cdot\Gamma[(a_n+1)/2]} \cdot \frac{2}{3} \cdot \frac{a_n+2}{a_n(a_n-1)} = 100\ \% \end{aligned} \tag{11}$$

(it is well known that $\Gamma(1/2) = \sqrt{\pi}$ and $\Gamma(1) = 1$ hold). The $k_{opt} = 1/\varepsilon_{st}$ relation really holds also for the limiting function $f(x)$ (which represents the case $a = 1$), and not only in the open interval $1 < a < \infty$.

In the practice of the MFV-procedures it is enough to work with "round values" of k (because of the great robustness of these statistical algorithms). Consequently, instead of $k = 0.3857965$ we propose to use simply $k = 1/2$, if very long tailed distributions are to be expected. It follows from numerical calculations that the value of $\overline{A}^2 \equiv (S_1 - S_2)/(2S_2 - S_1)^2 = 3/2$ in Eq. (8d) varies slowly according to Fig. 2 around $k_{opt} = 0.385797$, and therefore the use of $k = 1/2$ causes a neglibigle loss in the value of efficiency: $e = 98.156\ \%$ holds for this k-value being $\overline{A}^2(k_{opt})/\overline{A}^2(1/2) = 0.98156$. (The cause of the rapid increase of $\overline{A}2$ — and consequently of a rapid decrease of the efficiency — for diminishing k-values is evidently the fact that using $S = k\varepsilon$ are more and more values practically neglected from the most dense, most informative group of data.)

A curious question arises: applying $k = 1/2$, the efficiency in case of the Gaussian distribution, *i.e.*, on the opposite end of the studied type-interval ($a \to \infty$) would not have a practically unusable small value? Such a danger does not exist: using $k = 1/2$ the efficiency equals 50.42 %.

The reversed questions are not less important: if the probability distributions are mainly expected in the Cauchy-Gaussian type interval and therefore for k the choices 1, 2 and 3 are proposed (see *e.g.* the Table at the end of Steiner 1991), which for efficiencies can be achieved using these k-values for distributions with extremely heavy flanks, *i.e.*, near to $a = 1$? The approximate answers are given by the curves "$Mk = 3$" and "$Mk = 2$" in Figs 8 and 10 in Steiner and Hajagos (1993) and the curve "MFV" in Fig. 6 in Hajagos and Steiner (1993a); in all three cases the efficiencies are shown in function of $t = 1/(a-1)$, consequently $a \to 1$ corresponds to $t \to \infty$. For $k = 3$ about 30 %, for $k = 2$ near to 50 % and for $k = 1$ more than 70 % seem to be the limiting efficiency in case of $a \to 1$. (Even these asymptotic values attracted the author's attention to a more detailed investigation of the properties of the MFV-procedures in the type interval where also extremely heavy tails occur.) These asymptotic values, however, can also be calculated correctly in similar way which resulted in Eq. (11). Namely, if instead of the specially chosen $k\varepsilon = 1$ (in Eqs (8b) and (8c)) the general expression $k\varepsilon$ is written, *i.e.*, if the following original

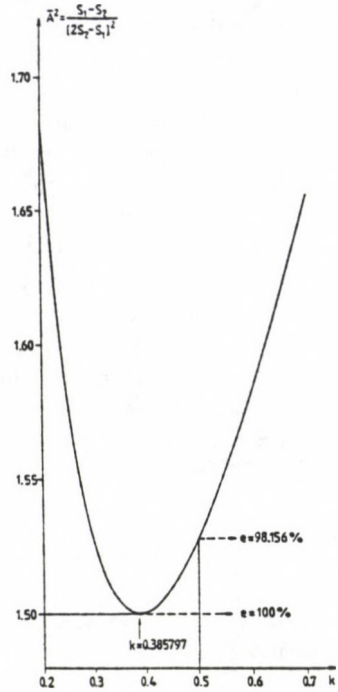

Fig. 2. Auxiliary curve to show the variation of the quotient $\overline{A}^2 \equiv (S_1 - S_2)/(2S_2 - S_1)^2$ (see Eq. (8)) versus the factor k. The small variation around the optimal $k = 0.385797$ convinces us that practically the "round value" $k = 1/2$ can be equally used (causing only an efficiency loss of less than 2 % if $a \to 1$)

definitions of S_1 and S_2 are accepted in case of $f(x)$ (see Eq. (3)):

$$S_1 = \int_{-\infty}^{\infty} \frac{(k\varepsilon)^2}{(k\varepsilon)^2 + x^2} \frac{1}{\sqrt{1+x^2}} dx \tag{12a}$$

and

$$S_2 = \int_{-\infty}^{\infty} \frac{(k\varepsilon)^4}{[(k\varepsilon)^2 + x^2]^2} \frac{1}{\sqrt{1+x^2}} dx \ , \tag{12b}$$

and consequently the asymptotic variance A^2 is to be written with these S_1 and S_2 in the form

$$A^2 = (k\varepsilon)^2 \cdot \frac{S_1 - S_2}{[2S_2 - S_1]^2} \ , \tag{13}$$

the efficiency is to be computed according to the formula

$$\begin{aligned} e \ &= \ \lim_{a \to 1} 100 \frac{(a-1) \cdot \Gamma(a/2)}{2 \cdot \sqrt{\pi} \cdot \Gamma[(a+1)/2]} \cdot \\ &\cdot \frac{[2S_2 - S_1]^2}{(k\varepsilon)^2 (S_1 - S_2)} \cdot \frac{a+2}{a(a-1)} = 150 \cdot \frac{[2S_2 - S_1]^2}{(k\varepsilon)^2 \cdot (S_1 - S_2)} \end{aligned} \tag{14}$$

(compare with Eq. (11)). Using naturally the ε-value 2.59204005 (see Eq. (7)), the correct values of the asymptotic turn out to be the followings: for $k = 3$ $e_{asympt.} = 32.8606$ %; for $k = 2$ $e_{asympt.} = 48.0112$ % and for $k = 1$ $e_{asympt.} = 77.7336$ %. This latter value and the fact that if $k = 1$, the efficiency is for the Gaussian distribution (*i.e.*, for $a \to \infty$) 73.73 %, mean not less that for the whole type interval $1 < a < \infty$ the efficiencies for the norm PC (see Table I) are greater than 73.7 %. Compare this astonishing result with the properties of the procedure based on the L_1-norm (see once more Table I): at the Gaussian distribution the efficiency is well known for the medians to be only $100 \cdot 2/\pi = 63.662$ % and for $a \to 1$ the efficiency tends to zero. Concerning the last statement see the "median" curve in Fig. 8 of Steiner and Hajagos (1993), or compute directly according to the formula $A_{med} = 1/[2\varphi(T)]$ (see *e.g.* Cramér 1945) which has evidently in case of $\varphi(x) = f_a(x)$ the form $A_{med} = 2 \cdot n(a)$. Consequently instead of the formula in Eq. (14) we have for the norm L_1 (see Table I)

$$e = \lim_{a \to 1} 100\, e_{med} = \lim_{a \to 1} 100 \cdot \frac{a+2}{a(a-1)} \cdot \frac{(a-1)^2 \cdot \Gamma^2(a/2)}{\pi \cdot \Gamma^2[(a+1)/2]} \tag{15}$$

which limit value is obviously zero. (A further consequence of Eq. (15): the maximum e_{med}-value for $f_a(x)$ distributions is 83.67 % which is reached at $a = 2.64$.)

3. Comment to the resistance of statistical procedures based on different norms of the residuals

Let X_i be the residual, *i.e.*, the i-th difference between measured and calculated values. Let Table 4 from Hajagos and Steiner (1993b) be cited in a shortened form: the formulas of two L_p-norms and that of four P_k norms are contained in Table I. (The general expressions for both the L_p and P_k norms are given in integral form in the Table at the end of Steiner's (1991) book.) The P-norms are here denoted according to the "eigen-distribution" where the algorithm based on the minimization of the chosen norm optimally works. As $f_a(x)$ in case of $a = 9$ is called "Jeffreys-distribution", the P_k-norm for $k = 3$ is denoted by P_J; similarly, P_C means the P_k-norm for $k = 1$ being optimal for Cauchy-densities of the residuals. In case of $k = 2$ (optimal for $a = 5$, *i.e.*, for the so-called geostatistical distribution) P_k is denoted simply as P (without any index) as the minimization of this norm is considered as the "standard MFV-procedure" recommended for use in cases when no *a priori* information is given concerning the distribution type of the errors. For $k = 1/2$ the P_k-norm was denoted by P_{lt} as this norm works best for very long-tailed distributions of the residuals. (This fact was proven just in the earlier discussions.)

As it was mentioned earlier, the continuous line in Fig. 1 does not represent a real case with respect to the probability theory, the formulas of the MFV-procedures, however, can be applied also for the case if $\varphi(x) = f(x)$. For example, the formula for the dihesion ε (which has a central role in the MFV-procedures, see Hajagos

and Steiner 1993a), namely the (also iteratively used) equation

$$\varepsilon^2 = 3\frac{\int\limits_{-\infty}^{\infty} \frac{X^2}{(\varepsilon^2+X^2)^2}\varphi(X)dX}{\int\limits_{-\infty}^{\infty} \frac{1}{(\varepsilon^2+X^2)^2}\varphi(X)dX} \tag{16}$$

is also satisfied if $\varphi = f$ (see Eq. (3)) and $\varepsilon = 2.59204005$. (By the way, it is perhaps not superfluous to mention that the ε values figuring in the formulas for the P_k-norms in Table I must always statisfy Eq. (16).)

Table I. Different norms discussed in the present paper. The minimization of each norm defines an inversion algorithm

Defining value	Norm	Formula	Eigen-distribution (for this type of error distribution works the norm in question optimally)
$p = 1$	L_1	$\frac{1}{n}\sum\limits_{i=1}^{n}\lvert X_i\rvert$	Laplace
$p = 2$	L_2	$\sqrt{\frac{1}{n}\sum\limits_{i=1}^{n} X_i^2}$	Gauss
$k = 3$	P_J	$\varepsilon\cdot\left\{\prod\limits_{i=1}^{n}\left[1+\left(\frac{X_i}{3\varepsilon}\right)^2\right]\right\}^{\frac{1}{2n}}$	Jeffreys
$k = 2$	P	$\varepsilon\cdot\left\{\prod\limits_{i=1}^{n}\left[1+\left(\frac{X_i}{2\varepsilon}\right)^2\right]\right\}^{\frac{1}{2n}}$	geostatistical
$k = 1$	P_C	$\varepsilon\cdot\left\{\prod\limits_{i=1}^{n}\left[1+\left(\frac{X_i}{\varepsilon}\right)^2\right]\right\}^{\frac{1}{2n}}$	Cauchy
$k = \frac{1}{2}$	P_{lt}	$\varepsilon\cdot\left\{\prod\limits_{i=1}^{n}\left[1+\left(\frac{2X_i}{\varepsilon}\right)^2\right]\right\}^{\frac{1}{2n}}$	(very long-tailed error distributions)

It is useful to realize once more that the maximum zone of $f(x)$ well approximates the maximum zones of the real $f_a(x)$ distributions if $a = 1 + \delta$ (where $\delta \ll 1$). As $f(x)$ itself can be treated with the MFV calculation techniques, this procedure obviously results in the symmetry point as in most frequent value T, even if the probability $\int\limits_{T-\varepsilon}^{T+\varepsilon} \varphi(x)dx$ *i.e.*, the occurrence probability of the densest data in the neighbourhood of T is very small (in case of $\varphi(x) = f_a(x)$ if $\delta(> 0)$ is very near to zero). In these cases the flanks are really heavy, representing a probability between 0.4 and 0.5 each, — the total Fisher-information, however, can be "exhausted" using MFV-procedures, as we have seen it earlier: remember that 100 percent efficiency can be achieved — and even the standard version of the MFV-procedure

(with $k = 2$, see the P-norm in Table I) results in nearly 50 percent efficiencies in these extremal cases. It should be stressed that this property does not characterize all in the practice used robust procedures, on the contrary, Figs 6, 7 and 8 in Steiner and Hajagos (1993) and Fig. 6 in Hajagos and Steiner (1993a) show that the efficiencies of all most cited and used robust procedures tend to zero if $a \to 1$. The name "most frequent value" is in all circumstances justified, even in extremal ones. This important property accentuatedly separates the MFV-procedures from other "reweighted least squares" techniques having very similar algorithms in cases where a linear dependence exists belonging to the unknown model-parameters (or a reasonable linearization can be made). Seemingly there are no great differences among the plenty of reweighted least squares algorithms: only the weight-formulas must be calculated specially. The various weight-formulas, however, are defined in the overwhelming majority of cases only as "ad hoc" proposals without sufficient theoretical background, and, in addition, these weights are often used with an obscure parameter of scale S which is "determined" by looking at the ceiling.

In few, rather exceptional cases the scale parameter of the weights can be approximately chosen resulting in acceptable bias of the modelparameters on ground of the tested measuring error of the used instrument or equipment, *e.g.*, in some simple geodetic tasks or if gravimeter measurements were carried out. In the latter case, however, the residuals of the inversion algorithm can be heavily influenced if the model used in the inversion is a poor approximation of the geological reality. Consequently, alltough there really exist lucky situations where the inversion can be executed with an "*a priori* known" scale parameter in the weights, or, more generally, in the norms, — in the overwhelming majority of cases, however, the value of the scale parameter must be determined statistically. Hajagos and Steiner (1993a) stressed the importance of the correct choice of the algorithm for determining this parameter of scale; the same iteration algorithm with just the same formula for the weights can give fully satisfying results or fully unacceptable, too, (on ground of the same measuring material), depending only upon the choice of the scale parameter.

In the followings it should be treated the resistance of the inversion algorithms based on the minimization of the norms given in Table I. These algorithms will be mathematically (or better: with respect to the computing techniques) in the near future more and more really minimizations of the norms, thanks to the growing computer-possibilities instead to make troublesome linearizations resulting in a sort of "reweighted least squares" algorithm. This process simultaneously will promote that the practitioner will be more and more aware the differences existing among the algorithms based on different norms (and will not lead oneself by the nose by the computing-technical fact that "reweighted least squares" can be carried out by the same program).

Looking at the expression for L_2 in Table I, it can be trivially established that in the presence of outliers some very great residuals X_i can occur and these distort the results of the inversion (*i.e.*, the minimum place of the norm in the parameterspace) or lead to fully unacceptable results. The classical statistical methods, based on the L_2-norm, are therefore in a high degree not resistant.

This statement is in close connection with the fact that in the L_2-formula X_i^2

figures in a simple sum. Consequently, the L_1-norm has not neglibigle advantages to respect of the resistance against outliers as a great $|X_i|$ does not play by no means a such dominant role in the value of the sum defining the L_1 norm as X_i^2 does in the L_2 norm. It should be mentioned, however, that asymmetrically lying outliers which represent a not neglibigle percent of the data, may significantly distort the inversion results even if L_1 is minimized.

And what about the resistance of the MFV-procedures, based on the minimization of P_k-norms? The formulas in Table I give no trivial answers but some investigations were made in the last two decades resulting in the statement: MFV-procedures are really in a high degree resistent. Let be cited an astonishing result from Hajagos and Steiner (1993b) where already direct minimization of the norms were carried out in all cases: even more than 50 percent(!) of the data can be outlier and nevertheless, the inversion result remains acceptable. Figure 13 in loc. cit. shows that using the P_{lt}-norm the distortion is even in this extremal case less than 1 percent; the inversion results determined by the minimization of the L_1-norm are fully unacceptable in the same case. (It should be mentioned too, that one single outlier can cause some ten percent distortion in the inversion result when using the L_2-norm, see loc. cit. Fig. 10.)

In the following the results of an instructive and simple investigation, should be shown concerning the resistance of the MFV-procedures.

Let be denoted by $f_{x_o}(x)$ the following probability density function:

$$f_{x_o}(x) = \begin{cases} n(x_o) \cdot f(x), & \text{if} \quad |x| < x_o \\ 0, & \text{if} \quad |x| \geq x_o \end{cases}, \tag{17}$$

where the norming factor $n(x_o)$ is obviously the reciprocal value of the integral $\int\limits_{-\infty}^{\infty} f_{x_o}(x)dx$, *i.e.*

$$n(x_o) = \frac{1}{\int\limits_{-x_o}^{x_o} \frac{1}{\sqrt{1+x^2}} dx}. \tag{17a}$$

For the determination of MFV-values, weighted means are to be calculated iteratively where the weights are in general case $1/[(k\varepsilon)^2 + (x-T)^2]$. Supposing relatively great values of x_o, *i.e.*, heavy flanks, $k = 1/2$ can be appropriately chosen (this evidently results in the minimum place of the P_{lt}-norm).

Outliers should be in a relatively "infinite" rate among the data, in particular in the interval (x_o, ∞), distributed according to $f(x)$. (We remember that in the reality this infinite rate can only be approximated by $f_a(x)$ with $a = 1 + \delta$ where δ is very small compared to the unity.) For T obviously

$$T = \frac{\int\limits_{-x_o}^{\infty} \frac{x}{(\varepsilon/2)^2 + (x-T)^2} \frac{1}{\sqrt{1+x^2}} dx}{\int\limits_{-x_o}^{\infty} \frac{1}{(\varepsilon/2)^2 + (x-T)^2} \frac{1}{\sqrt{1+x^2}} dx} \tag{18}$$

must hold and therefore the distortion of the symmetry point of the correct data by this plenty of outliers is that T-value from the $(T;\ \varepsilon)$ value-pair, which latter simultaneously fulfils Eq. (18) and

$$\varepsilon^2 = 3\frac{\int\limits_{-x_o}^{\infty} \frac{(x-T)^2}{[\varepsilon^2+(x-T)^2]^2}\frac{1}{\sqrt{1+x^2}}dx}{\int\limits_{-x_o}^{\infty} \frac{1}{[\varepsilon^2+(x-T)^2]^2}\frac{1}{\sqrt{1+x^2}}dx} \tag{19}$$

(compare Eq. (16)). The T-values (here in the meaning of "bias") for $x_o = 100$; 50; 20 and 10 are given in Table II, where also the actual ε-values are given, and the quotients T/ε, too, to show the high resistance of the MFV-procedures.

Table II.

x_o	T	ε	T/ε
100	0.01868	2.5901	0.7 %
50	0.03731	2.5842	1.4 %
20	0.09181	2.5460	3.6 %
10	0.1752	2.4371	7.2 %

(These values can be reached by ping-pong iteration of only 50 steps, even if both start values are very great: 10^7.) It is instructive, too, to compare the ε-values with the value in Eq. (7). (It is perhaps not superfluous to clear in advance the eventually arising doubts concerning the integrals figuring in Eqs (18) and (19), giving a finite upper limit of the integrals, say, the value 10^5 instead of $+\infty$. It can be easily verified that in this case the values of the integrals in the mentioned two equations vary as maximum by a value 10^{-5}, being $\int\limits_{10^5}^{\infty} 1/x^2 dx = 10^{-5}$, see the numerator of the expression in Eq. (18); the resulting difference at the other three integrals are even much less than this value. — To this value 10^5, however, it is instructive to give in the followings a hypotetical but practical representation, too.)

If the asymmetrically occurring outliers lie very far, too, *e.g.*, in the interval $10 < x < 10^5$, distributed according to $c \cdot f(x)$, and the correct data occur in the interval $-10 < x < 10$ according also to $c \cdot f(x)$ (with the same c-value), the ratio of the outliers to the reliable data is more than 3/2:

$$\frac{\text{outliers}}{\text{correct data}} = \frac{\int\limits_{-10}^{10} \frac{c}{\sqrt{1+x^2}}dx}{\int\limits_{10}^{10^5} \frac{c}{\sqrt{1+x^2}}dx} = 1.5553\,.$$

In spite of this terrible ratio of asymmetric outliers, the symmetry point of the distribution $f_{x_o}(x)$ $x_o = 10$ (see Eq. (17)) can be calculated by the MFV method

Table III.

	T	ε		T	ε
		$d = 6571.2$			5.651
0			10		
	$med = 49.93$			0.7596	
		2306.0			3.020
1			15		
	70.29			0.2638	
		846.6			2.576
2			20		
	89.62			0.1946	
		332.9			2.472
3			25		
	43.34			0.1800	
		140.20			2.446
4			30		
	21.13			0.1764	
		63.46			2.439
5			35		
	10.61			0.1755	
		31.25			2.438
6			40		
	5.534			0.1753	
		17.11			2.438
7			45		
	3.033			0.1753	
		10.62			2.437
8			50		
	1.773			0.1752	
		7.402			
9					
	1.117				

using $k = 1/2$ (*i.e.*, minimizing the P_{lt}-norm) with a small bias of 0.175, which is only 7.2 % of the dispersion-characteristics ε (dihesion, see Table II).

As for the $f_{x_o}(x)$ probability densities the symmetry point is the adequate location parameter, naturally the meaning of the resulting T remains "bias" applying other statistical procedures, too. Choosing the L_1-norm minimization, *i.e.*, determining the value of the median for $x_o = 10$ and for the same outliers, the result is $T \approx 50$ (=49.93) and this value represents obviously a very poor estimation of the symmetry point of the correct data (lying in the origin). The distortion of the appropriate error characteristic, namely of the L_1 norm value calculated for $med = 49.93$, (*i.e.*, of the minimum L_1-norm often denoted by d = mean of the absolute residuals,) is even much more: $d = L_1(med) = 6571$ (the correct value is $d = 3.0184$).

The following conclusion: "these values got by using the L_1-norm are unusable", however, would be a false interpretation. In the contrary: both values can be used for starting the double way iteration defined by the Eqs (18) and (19). It would be perhaps not superfluous to give in Table III the detailed results of a ping-pong iteration for this case.

We see that *med* and d, or more generally: the results of the L_1 calculations can be useful as starting values for the most frequent value-iterations, even if a great number of asymmetrical outliers are present. (It should be mentioned, too, that significantly fewer steps of the iteration would be needed to the same accuracy by using more sophisticated procedures of the iteration but it seemed here to be the best choice to show the simplest, *i.e.*, the so-called ping-pong variant: one step according to Eq. (19), one step to Eq. (18), *etc.*)

At last it should be mentioned as a curiosity that the algorithm defined by Eqs (18) and (19), using the generalization that $\varphi(x)$ is written instead of $1/\sqrt{1+x^2}$, can work also if $\varphi(x)$ is an unimodal function but for it the integral $\int_{-\infty}^{\infty} \varphi(x)dx$ is divergent. In this case $\varphi(x)$ is obviously not a probability density function in sense of the probability theory but $\varphi(x)$ is nevertheless able to represent quotients of probabilities, for example $\int_a^b \varphi(x)dx / \int_c^d \varphi(x)dx$ gives that the event $a < x < b$ is how many times more probable than the event $c < x < d$. The location-interval of the densest lying data (*i.e.*, T and ε) can be determined by P_k-norms for such data, too, *i.e.*, if for the data-density model $\varphi(x)$ the relation $\int_{-\infty}^{\infty} \varphi(x)dx < \infty$ does not hold.

Acknowledgements

The author is indebted to thanks to Prof. Dr. L Csernyák, to C. Prof. B Hajagos and to Drs. T Fancsik for fruitful consultations and discussions.

References

Cramér H 1945: Mathematical Methods of Statistics. Almqvist & Wiksells, Uppsala

Csernyák L 1991: In: The Most Frequent Value. Steiner F ed. Chapter II. Akadémiai Kiadó, Budapest

Gröbner W, Hofreiter N 1958: Integraltafel II. (Bestimmte Integrale) Springer, Wien und Innsbruck

Hajagos B, Steiner F 1993a: *Geophysical Transactions*, 38, 211–230.

Hajagos B, Steiner F 1993b: In: Annales Univ. Sci. ELTE (Budapest), *Sectio Geophysica et Meteorologica*, 9, 101–146.

Steiner F 1988: *Geophysical Transactions*, 34, 139–260.

Steiner F ed. 1991: The Most Frequent Value. Akadémiai Kiadó, Budapest

Steiner F, Hajagos B 1993: *Geophysical Transactions*, 38, 193–210.

Acta Geod. Geoph. Hung., Vol. 29(1–2), pp. 273–275 (1994)

Book reviews

H FRÖHLICH: *Begleitende Programme zur Trassierung.* Dümmlers Verlag, Bonn, 1994, 16 pages, 9.80 DM

This booklet is a schematic description of a PC-compatible 3.5" or 5.25" program disk to be purchased for DM 175. The disk contains auxiliary programs to solve the following two-dimensional tasks if the main points and parameters of the trace are known:

SCHNITTE: computation of intersection points between the following geometrical figures: straight line, circle, clothoid and their parallels.

TRABEL (TRassen ABstech ELemente): computation of the basic data, of staking elements of given traces, transformation of the coordinates into local systems.

ZPD (ZwangsPunktDiagnose): diagnosis of reference points. Characterization of the reference points and computation of the shortest distance of these points from straight lines, circles, clothoids.

It can be seen that the main advantage of these routines is that tasks can be solved in addition to standard geometric figures also on a clothoid basis, and further, there is a possibility to compute with these figures in shifted position.

The program package contains two further auxiliary programs, too. The first routine, ZPE has the task to carry out the input of the coordinates and to modify the list of coordinates. The routine TREDIT can be used for the creation of the data base of the main points and of the elements of the trace.

The booklet yields only sufficient information for the installation of the program package and for the use of the printer and it presents 3 examples. There is reference to the detailed documentation of the programs as a text file (180 pages) on the disk thus it is always available and can even be printed. Knowing this, the present booklet cannot be recommended to be purchased. As a product description (on the advertisement level) it is excellent, but expensive, as documentation, it cannot be used.

J Kalmár

F J GRUBER: *Formelsammlung für das Vermessungswesen.* 6th edition. Dümmler Verlag, Bonn, 1994, VIII + 147 pp, 195 fig.

This A5 size vademecum collects formulas occurring in the practice of many fields of surveying in the following sections.

General basic principles, basic mathematical formulas, differential calculus, matrix calculus, plane geometry, plane and spherical trigonometry.

Among basic tasks of surveying, formulas for simple computation of coordinates, calculation and partition of areas are included.

In the chapter on angle measurement, basic formulas for horizontal and vertical angle measuring and angle measuring with gyrotheodolites are summarized in addition to instrumental errors for theodolites.

1217-8977/94/$ 4.00 ©1994 Akadémiai Kiadó, Budapest

In the field of distance measurement, measurements with measuring tape, optical distance measurement and the corresponding correction formulas are summarized.

Point determinations: formulas for polar points, intersection with distances (trilateration), intersection with angles and resection.

Transformations in plane: Helmert and affine transformations.

Altimetric survey: levelling, trigonometric heighting.

Surveying for field engineering: alignement, setting out of arc of a circle, clothoid, calculating of cubic volume.

Basic formulas for adjustment and statistics.

The book is concluded by a selected list of references and by an index.

This most practical collection of formulas will help education, self-teaching and practice in many fields of surveying. Formulas are illustrated by a lot of very clear figures.

Gy Szádeczky-Kardoss

W SCHRÖDER, M COLLACINO collects and eds: *Geophysics: Past Achievements and Future Challanges.* Interdivisional Commission on History, IAGA. Bremen-Roennebeck, Germany, 1994

This volume contains selected papers from the symposia of the Interdivisional Commission on History of the International Association of Geomagnetism and Aeronomy during its Assembly held in Buenos Aires, in September 1993. It continues a series with similar aim and similar form, and even the editors remained. From the 12 papers published in this volume, 7 or 8 deal with the history of science in a more restricted sense, the remaining 4 or 5 use historical data in an analysis of cyclicity and/or connections between climatic parameters and epidemies.

One of the editors, Colacino, and Valensise continue the presentation of the role played by early Italian scientists and scientific societies in the development of geophysics, in the present case dealing with the Cimento Academy and with the Meteorologica Societas Palatina and their contribution to the establishment of modern meteorology. Barreto from Brazil contributes with a biography on Emmanuel Liais, who worked both in Europe and in South-America and established several observatories in Brazil, including a geomagnetic one.

Debarbat evaluates the role played by woman-scientists in different countries in the astronomy, presenting figures on the membership and comparing them by population of these countries and by other factors. Miletits et al. deal with geomagnetic research in the Nagycenk Observatory (Hungary), Schröder and Schneider with early geomagnetic observations near the western South Atlantic coast (abstract). A further paper coauthroed by Schröder and Wiederkehr depicts the early history of weather maps of the Southern Atlantic, during the First Polar Year, a common effort of German and Argentine scientists. Bernhardt contributes to the discovery of the stratosphere, by presenting a proof that it happened in 1894, *i.e.* earlier than supposed (in 1902), due to the fact that Assmann, the responsible scientist was most careful in publishing his results. Lauscher lists significant meteorological events in Europe in form of a Weather Chronic.

Gregori and his group investigate data series on floods, climatic anomalies and explosive volcanism, and try to prove the existence of a four-decade cycle in them. Schröder adds to this paper a list of remarkable floods of the river Tevere from historic sources. Cassiani deals with the connection between weather and plague in 17th century Italian cities, and confirms this correlation. The last paper by Gregori analyses volcanic eruption looking for

cyclic changes in the heat production. This is the most lengthy paper, including roughly one fifth of the volume. It emphasizes the feasibility of such studies, using world-wide data on volcanic activity.

This book is a most valuable contribution both for science historians and for those who investigate historical data concerning possible cyclic changes in meteorological, geomagnetic and thermal data, and connections with biological and epidemical events and cycles.

J Verő

PRINTED IN HUNGARY

Akadémiai Kiadó és Nyomda, Budapest

Manuscripts should include the following components which should be presented in the order listed (tables and illustrations should be separated in case of manuscripts on disk, too).

1. Title, name(s) of the author(s), affiliation, dateline, abstract, keywords
2. Text, acknowledgements
3. References
4. Footnotes
5. Legends
6. Tables
7. Illustrations

1. The *affiliation* should be as concise as possible and should include the complete mailing address of the authors. The *date of receipt* will be supplied by the editors. The *abstract* should not exceed 250 words and should clearly and simply summarize the most important methods and results. 5–10 significant expressions describing the content are used as *keywords*. Authors may recommend these keywords.

2. The *text* should be in English and as short and clear as possible. In case of typewritten manuscripts, please note the following:

— avoid possible confusion between o, O (letters) and 0 (zero), I (letter) and 1 (one), *ν* (Greek nu) and u, v (letters), etc.
— explain ambiguous and uncommon symbols by making marginal notes in pencil
— formulas are to be numbered consecutively with the number in parentheses to the right of the formula. Text references to the equations may then be made by the number in parenthesis. The word equation in this context is to be abbreviated to Eq. and Eqs in the plural
— the International System of Units (SI) should be used.

3. *References* are accepted only in the Harvard system. Citations in the text should be as:

...(Bomford 1971)...or Bomford (1971)...

...(Brosche and Sündermann 1976)...

...(Gibbs et al. 1976b)...

The list of references should contain names and initials of all authors (the abbreviation et al. is not accepted here): for journal articles, year of publication, the title of the journal in abbreviated form, volume number, first and last page. For books or chapters in books, the title is followed by the publisher and place of publication.

All items must appear both in the text and in the list of references.

4. *Footnotes* should be typed on separate sheets.

5. *Legends* should be short and clear. The place of the tables and figures should be indicated in the text, on the margin.

6. *Tables* should be numbered serially with Roman numerals. Vertical lines are not used.

7. All the *illustrations* should contain the figure number and author's name in pencil on the reverse. The most important point with figures is clearness. Photographs and half-tone illustrations should be sharp and well contrasted. Colour photographs will be accepted, but the extra cost of reproduction in colour must be borne by the authors (approx. US$ 260 per page).

Only original papers will be published and a copy of the Publishing Agreement will be sent to the authors of papers accepted for publication. Manuscripts will be processed only after receiving the signed copy of the agreement. Information is sent to the first author if no other wish is expressed. Fifty offprints are supplied free of charge, for further copies please contact the Editorial Office.

Acta Geodaetica et Geophysica Hungarica

VOLUME 29, NUMBERS 3–4, 1994

Akadémiai Kiadó, Budapest

AGG 29 (3–4) 277–458 (1994) HU ISSN 1217–8977

ACTA GEODAETICA ET GEOPHYSICA HUNGARICA
A Quarterly Journal of the Hungarian Academy of Sciences

Acta Geodaetica et Geophysica Hungarica (AGG) publishes original reports and reviews on geodesy and geophysics in English.

AGG is published in yearly volumes of four issues by

AKADÉMIAI KIADÓ
Publishing House of the Hungarian Academy of Sciences
H-1117 Budapest, Prielle K. u. 19–35

Manuscripts and editorial correspondence should be addressed to

AGG Editorial Office
Geodetical and Geophysical Research Institute
H-9401 Sopron, P.O. Box 5
E-mail: actagg@sun10.ggki.hu

Subscription information

Orders should be addressed to

AKADÉMIAI KIADÓ
H-1519 Budapest, P.O. Box 245

Subscription price for Volume 29 (1994) in 4 issues US$ 80.00, including normal postage, airmail delivery US$ 20.00.

INSTRUCTIONS TO AUTHORS

Manuscripts should be sent to the Editorial Office (address see above). Articles not published or submitted for publication elsewhere are only accepted.

Manuscripts should be typewritten in duplicate or on disk (LATEX∗.tex files or Word for Windows Documents or text-only ASCII files) with a written copy, or via E-mail (LATEX∗.tex files or text-only ASCII files).

CONTENTS

Book reviews

Preface

The papers contained in the present issue of Acta Geodaetica et Geophysica Hungarica were presented at the International Winter Seminar on Geodynamics. This seminar was organized by the Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences under the auspices of the International Association of Geodesy, of the European Geophysical Society and of the Central European Initiative's Committee of Earth Sciences.

The Winter Seminar took place in Sopron between the 21th and 25th February 1994, already for the fourth time. The first three aimed to describe different chapters of geodynamics, to give information for geoscientists on the state of art in the research of the structure of our planet and in different fields of physical geodesy. The first Winter Seminar in 1987 delt with problems connected to the Earth's rotation (lecturers were K Bretterbauer, I Abonyi, H Moritz, V Dehant, H Jochmann, P Melchior, M I Yurkina). The next Seminar was in 1989 and the lectures presented there were connected with problems of the deformation of the Earth and its gravity field (the lecturers were H Moritz, E Groten, E W Grafarend, K Arnold, I Abonyi). The Winter Seminar in 1992 concentrated on the questions of the inner structure and dynamics of the Earth (invited lecturers were S M Molodensky, C Denis, V N Zharkov, A Ádám, G F Panza and valuable contributed papers were presented by M Stavinschi – D Dinescu and S Franck).

The 1994 Winter Seminar was slightly different from its predecessors. The seminars of 1987, 1989 and 1992 described different geodynamic phenomena while this fourth Winter Seminar in Sopron concentrated on the techniques and methods used in geodynamical research. This alteration in the organizers' effort is due to the emergence of revolutionary new methods in space and surface geodesy, accuracies of which give new dimension for geodetic measurements and make the study of the temporal variations of geodynamical phenomena more realistic. In the same time some investigators are too optimistic in use of the scientific tools in earthquake prediction, in study of recent crustal movements, in solving open problems of geodynamics *etc.* The realistic estimation of the reliability and accuracy of these new scientific tools is sometimes too optimistic. An understanding of the error sources is also of primary importance. In the same time the new methods of geodynamics need new physical and mathematical background for an effective application.

During the fourth Sopron Winter Seminar two review lectures were connected to the use of VLBI in geodynamical investigations. Another review paper was devoted to the possibilities of satellite and lunar laser ranging. The Global Positioning System (GPS) basically changed the technics of the geodetic measurements carried

Akadémiai Kiadó, Budapest

out on the Earth's surface. The reviewer of this problem showed the possibilities and the limitations of this new method of geodesy.

An important problem of modern geodynamical research is the observation and interpretation of temporal changes in deformations, tilts and gravity variations. Two review papers were connected to the main problems of automation and mathematical methods in geodynamical research.

The Geodetic and Geophysical Reserach Institute is planning to continue the series of biennial Winter Seminars. The next one will be held in February 1996 and its topic will be "Gravity Field and Figure of the Earth".

P Varga

Acta Geod. Geoph. Hung., Vol. 29(3–4), pp. 279–337 (1994)

GEODYNAMICS: THE DEFORMATION OF THE EARTH'S SURFACE

F SANSÓ[1]

Contents:

Keywords: boundary value problem; deformation; elasticity; equilibrium; Hooke's law; stress; viscoelasticity

Foreword

I am extremely grateful to Prof. P Varga who invited me to lecture at the Winter Seminar 1994 in Sopron on such a fascinating subject as Geodynamics-Deformation of the Earth.

Since the success of a Seminar depends from the fact that at least someone has learnt something, I can claim that the Seminar has been successful for sure, as I had to learn a lot in order to give these lectures.

Indeed the subject of deformation of the earth is in itself too wide to hope to give any comprehensive review of the subject in relatively short lecture notes, so I had to make a particular choice of items within this subject, what I did according to the following lines; I tried to be self-consistent so that lecture notes could be used also by people who were not experts in elasticity theory and related topics; I limited the analysis to plane theory which means from the geophysical point of view

[1]DIIAR Politecnico di Milano, Piazza Leonardo da Vinci 32, I–20133 Milano, Italy

1217-8977/94/$ 4.00

to areas of regional extent (*e.g.* with a diameter less than 1000 km) in order to avoid pure geometrical complications; I concentrated the attention on vertical displacements basically thinking in terms of a crust solicited by deep seated phenomena (*e.g.* mantle flow) or surface load (*e.g.* glaciation-deglaciation); I tried to prepare examples which could lead to the natural point of view for a geodesist, namely to see the relation between surface deformation and deep stress as an inverse problem.

Together with elasticity theory, actually its linearized version, the lecture notes tackle more advanced items like plasticity or visco-elasticity; the basic mathematical tool used in this context is the calculus of variations because of its strong connection with the physical concept of energy on one side and its highly advanced potential of analysis in terms of conditions for existence and uniqueness of the solution of the various problem formulated in the text.

I wish that these lecture notes could be useful to someone else in the future beyond myself.

1. Deformation

Deformation is the action of changing the position of "points" of a material body in such a way that the distance between points in the final position is in general different from that in the initial position.

Remark 1.1 Deformation is then different from rigid motion which by definition doesn't alter point distances.

Deformation is described by suitable tensors which we introduce, here in Cartesian coordinates only, as follows: referring to Fig. 1 we can write:

$$\begin{cases} \mathbf{y} &= \mathbf{x} + \mathbf{u}(\mathbf{x}) \\ \mathbf{x} &= \mathbf{y} - \mathbf{v}(\mathbf{y}) \end{cases} \tag{1}$$

where $\mathbf{u}(\mathbf{x})$ is the vector field representing the shift from initial position $\mathbf{x}$ to final position $\mathbf{y}$, while $\mathbf{v}(\mathbf{y})$ is exactly the same displacement field:

$$(\mathbf{x} \leftrightarrow \mathbf{y}) \Rightarrow \mathbf{u}(\mathbf{x}) = \mathbf{v}(\mathbf{y}), \tag{2}$$

but as function of the final point $\mathbf{y}$.

Differentiating Eq. (1) one gets the basic relations (cf. Fig. 2)

$$\begin{cases} d\mathbf{y} &= d\mathbf{x} + U d\mathbf{x} \quad \left(U = \left[\frac{\partial u_i}{\partial x_k}\right]\right) \\ d\mathbf{x} &= d\mathbf{y} - V d\mathbf{y} \quad \left(V = \left[\frac{\partial v_i}{\partial y_k}\right]\right) \end{cases} \tag{3}$$

so that [2]:

$$\begin{cases} ds_y^2 &= d\mathbf{x}^+(I + U^+)(I + U)d\mathbf{x} = d\mathbf{y}^+ d\mathbf{y} \\ ds_x^2 &= d\mathbf{y}^+(I - V^+)(I - V)d\mathbf{y} = d\mathbf{x}^+ d\mathbf{x}. \end{cases} \tag{4}$$

[2] We will use throughout the text the vector, algebraic or tensor symbolism according to convenience.

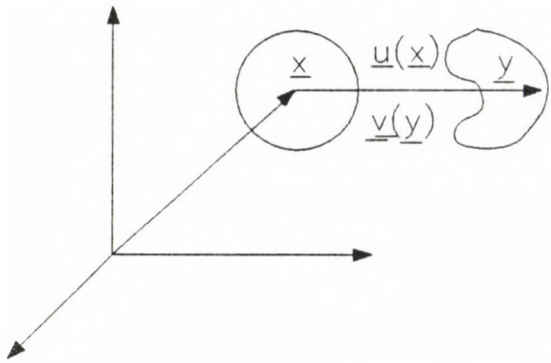

Fig. 1. The displacement fields $\mathbf{u}(\mathbf{x})$, $\mathbf{v}(\mathbf{y})$ (on the figures, vectors are underlined)

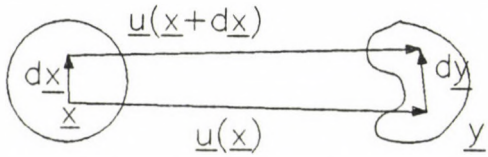

Fig. 2. Differential of the displacement field

We find then as linear rate of deformation, at the point $\mathbf{x}$ in the direction $\boldsymbol{\xi} = \frac{d\mathbf{x}}{ds_x}$, the quantity

$$\frac{ds_y - ds_x}{ds_x} = \frac{ds_y^2 - ds_x^2}{2ds_x^2} = \boldsymbol{\xi}^+ \frac{1}{2}(U^+ + U + U^+U)\boldsymbol{\xi} : \tag{5}$$

the tensor giving rise to the quadratic form Eq. (5), *i.e.*

$$\begin{cases} E & = \frac{1}{2}(U^+ + U + U^+U) \\ E_{ik} & = \frac{1}{2}\left(\frac{\partial u_k}{\partial x_i} + \frac{\partial u_i}{\partial x_k} + \frac{\partial u_\ell}{\partial x_i} \cdot \frac{\partial u_\ell}{\partial x_k}\right), \end{cases} \tag{6}$$

is called the Green strain tensor.

Analogously we define the tensor

$$\begin{aligned} e &= \tfrac{1}{2}(V^+ + V - V^+V) \\ e_{ik} &= \tfrac{1}{2}\left(\frac{\partial v_k}{\partial y_i} + \frac{\partial v_i}{\partial y_k} - \frac{\partial v_\ell}{\partial y_i} \cdot \frac{\partial v_\ell}{\partial y_k}\right), \end{aligned} \tag{7}$$

also called the Almansi strain tensor, as the matrix necessary to describe the linear rate of deformation at $\mathbf{y}$ in the direction $\eta = \frac{d\mathbf{y}}{ds_y}$, namely

$$\frac{ds_y - ds_x}{ds_y} = \frac{ds_y^2 - ds_x^2}{2ds_y^2} = \boldsymbol{\eta}^+ \frac{1}{2}(V^+ + V - V^+V)\boldsymbol{\eta}\,. \tag{8}$$

Remark 1.2 Both tensors, E and e, are symmetric by their very nature.

Remark 1.3 As it was easily predictable the strain tensor vanishes in correspondence to a "rigid motion" displacement: in fact for

$$\begin{gathered} \mathbf{y} = \mathbf{x} + \mathbf{u} = \mathbf{u}_0 + R\mathbf{x} \\ (\mathbf{u}_0 = \text{const.} \quad R = \text{const.} \quad R^+R = I) \end{gathered} \tag{9}$$

one has

$$I + U = R, \quad U = R - I \tag{10}$$

which substituted in Eq. (6) gives

$$E = 0$$

in view of the orthogonality of R.

Small displacements approximation. This is realized when we assume that $\mathbf{u}(\mathbf{x})$ is infinitesimal so that, neglecting 2nd order effects,

$$\begin{cases} E & \cong \frac{1}{2}(U^+ + U) \\ e & \cong \frac{1}{2}(V^+ + V)\,. \end{cases} \tag{11}$$

From Eq. (2) and Eq. (1) we find

$$\begin{cases} \dfrac{\partial u_i}{\partial x_k} & = \dfrac{\partial v_i}{\partial x_k} = \dfrac{\partial v_i}{\partial y_\ell} \cdot \dfrac{\partial y_\ell}{\partial x_k} \\ U & = V(I + U)\,. \end{cases} \tag{12}$$

Since both U and V are infinitesimal, with $\mathbf{u}$ and $\mathbf{v}$, we see that in our linear approximation we can put

$$U \cong V \tag{13}$$

According to Eq. (13) we have that in small displacement theory

$$E \cong e \cong \left[\frac{1}{2}\left(\frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i}\right)\right] \tag{14}$$

and we can refer irrespectively to each of them simply as the "strain tensor".

We shall assume from now on this approximation to hold.

Geometric interpretation of the strain tensor e. To obtain this interpretation let us consider an interval dx_1 of the x_1 axis as incremental vector $d\mathbf{x} = (dx_1, 0, 0)$, as shown in Fig. 3. Let us then take the projection on the (x_1, x_2) plane of the corresponding infinitesimal $d\mathbf{y}$; since

$$\begin{cases} y_1 & = & x_1 + u_2(x_1, 0, 0) \\ y_2 & = & u_2(x_2, 0, 0) \end{cases} \quad (0 \le x_1 \le dx_1) \tag{15}$$

we can write, to the first order (cf. Fig. 3)

$$\begin{cases} y_1 &= u_1(0,0,0) + \left(1 + \dfrac{\partial u_1}{\partial x_1}\right) x_1 = u_1(0,0,0) + (1 + e_{11})x_1 \\ y_2 &= u_2(0,0,0) + \dfrac{\partial u_2}{\partial x_1} x_1 = u_2(0,0,0) + \tan\alpha_1 x_1 \,. \end{cases} \tag{16}$$

From Eq. (16) we derive

$$e_{11} = \frac{dy_1 - dx_1}{dx_1}\,, \tag{17}$$

which shows that e_{11} is the extension of dx_1 in its own direction, as it was obvious from the definition of e; on the other hand if we take Fig. 4, with the help of Eq. (16) we see that

$$\frac{\pi}{2} - \beta = \alpha_1 + \alpha_2 = \frac{\partial u_2}{\partial x_1} + \frac{\partial u_1}{\partial x_2} = 2e_{12}\,. \tag{18}$$

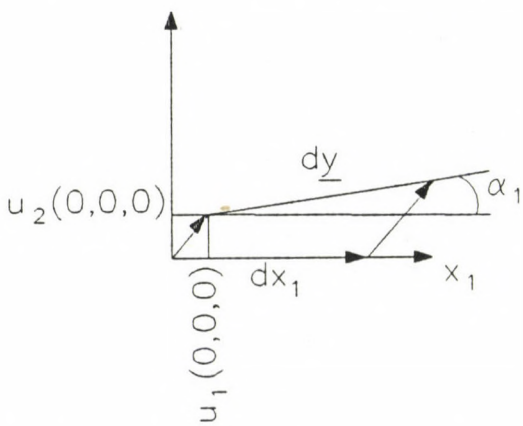

Fig. 3. Deformation of a segment lying on the x_1 axis

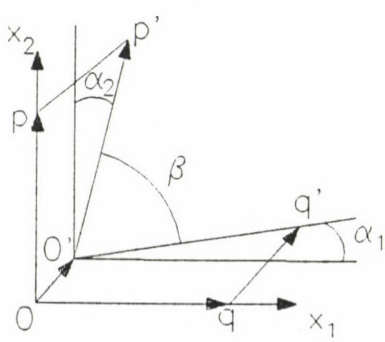

Fig. 4. Deformation of the right angle POQ into the angle $P'O'Q'$ of measure β

Due to Eq. (18) and the analogous equations for the off-diagonal terms $e_{ik}(i \neq k)$, one has called $2e_{ik}$, the "shear angles".

Remark 1.4 As it is clear from Eq. (11) and Eq. (14) the small strain tensor e is essentially the symmetrical part of the Jacobian:

$$U = \frac{\partial \mathbf{u}}{\partial \mathbf{x}} . \tag{19}$$

If we want to consider also the antisymmetric part of U, one is led to define the tensor

$$\begin{cases} \Omega & = \frac{1}{2}(U - U^{+}) \\ \omega_{ik} & = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_k} - \frac{\partial u_k}{\partial x_i}\right) \end{cases} \tag{20}$$

which is skewsymmetric by definition

$$\omega_{ki} = -\omega_{ik} \tag{21}$$

and which is called "rotation tensor".

The knowledge of e alone shouldn't be sufficient to reconstruct U, at first sight, yet it turns out that since U is not an arbitrary matrix, but has to satisfy compatibility conditions, being of the type gradient of the vector field $\mathbf{u}$, then the knowledge of e yields the knowledge of the first derivatives of Ω, $\frac{\partial \Omega}{\partial x_\ell}$. In fact, from Eq. (20),

$$\begin{aligned} \frac{\partial \omega_{ik}}{\partial x_\ell} &= \frac{1}{2}\left[\frac{\partial^2 u_i}{\partial x_\ell \partial x_k} - \frac{\partial^2 u_k}{\partial x_\ell \partial x_i}\right] = \\ &= \frac{1}{2}\left[\frac{\partial^2 u_i}{\partial x_\ell \partial x_k} + \frac{\partial^2 u_\ell}{\partial x_k \partial x_i} - \frac{\partial^2 u_\ell}{\partial x_i \partial x_k} - \frac{\partial^2 u_k}{\partial x_i \partial x_k}\right] = \\ &= \frac{\partial}{\partial x_k}(e_{i\ell}) - \frac{\partial}{\partial x_i}(e_{\ell k}) . \end{aligned} \tag{22}$$

This equation is the key to solve the following problem: given a displacement field $\mathbf{u}$ we can define the (small) strain tensor $e(\mathbf{u})$ through Eq. (14); now given a symmetric tensor e when can we say that it is a strain tensor? In other words, given e, when can we reconstruct a field $\mathbf{u}$ such that $e(\mathbf{u})$ coincides with the given e?

The reasoning runs as follows: if $e(\mathbf{u})$ is computed by Eq. (14) and $\mathbf{u}$ by Eq. (20), then Eq. (22) holds; consider now Eq. (22) for i, k fixed, then it is stated that the right hand side gives for $\ell = 1, 2, 3$ the three components of the (Cartesian) gradient of ω_{ik}; then the following conditions of compatibility necessarily hold, when ω_{ik} is suitably smooth,

$$\frac{\partial}{\partial x_m} \cdot \frac{\partial \omega_{ik}}{\partial x_\ell} = \frac{\partial}{\partial x_\ell} \cdot \frac{\partial \omega_{ik}}{\partial x_m} ,$$

i.e. using the notation $f_{,\ell} = \frac{\partial f}{\partial x_\ell}$

$$e_{i\ell,km} - e_{\ell k,im} = e_{im,k\ell} - e_{mk,i\ell} . \tag{23}$$

The equations Eq. (23) are known as conditions of compatibility of St. Venant, and we have just proved that if $e(\mathbf{u})$ is a strain tensor, then necessarily Eq. (23) holds true.

Viceversa let's assume that a (regular) symmetric tensor e is given, satisfying conditions Eq. (23), in a simply connected domain; since Eq. (23) implies that (for i, k fixed) the vector $e_{i\ell,k} - e_{\ell k,i}$ $(\ell = 1, 2, 3)$ is in fact of gradient type, then we know that the system Eq. (22) has a solution ω_{ik} which is fixed up to an arbitrary constant tensor, $\overline{\omega}_{ik}$ [3].

Let us now consider the new system

$$u_{i,k} = e_{ik} + \omega_{ik}\,; \tag{24}$$

the r.h.s. $e_{ik} + \omega_{ik}$, for i fixed and $k = 1, 2, 3$, is as a matter of fact of gradient type since it satisfies the sufficient conditions

$$e_{ik,\ell} + \omega_{ik,\ell} = e_{i\ell,k} + \omega_{i\ell,k} \tag{25}$$

as it becomes evident as soon as we substitute Eq. (22) (from which ω_{ik} have been derived) into Eq. (25). But then Eq. (24) has a solution for each i, uniquely determined up to a constant vector $\mathbf{u}_i$. We have so shown that St. Venant conditions are sufficient to allow to retrieve the original displacement field $\mathbf{u}$ up to an arbitrary transformation of the type

$$u_{0i} = \overline{u}_i + \overline{\omega}_{ik} x_k\,, \tag{26}$$

which, considering that $\overline{u}_i, \overline{\omega}_{ik}$ have to be small since we are by definition developing a theory of small displacements, represents basically an infinitesimal rigid motion.

Remark 1.5 As it will be useful in the sequel we introduce a notation in which e is split into a pure volume deformation (i.e. isotropic) called dilatation and a pure volume preserving deformation.

Namely we put

$$\begin{cases} e &= \epsilon I + e' \\ \epsilon &= \frac{1}{3} Tre = \frac{1}{3}(e_{ii}) \end{cases} \tag{27}$$

and we see that

$$Tre' = 0\,. \tag{28}$$

To see that $3 \cdot \epsilon$ represents a cubic dilatation it is sufficient to consider that Tre is invariant under coordinate rotations; so we can use a reference system in which $e_{ik} = 0$ $(i \neq k)$, i.e. one in which the axes are eigenvectors of e with eigenvalues e_{11}, e_{22}, e_{33} respectively. But then under the action of V a cube with

[3] That the tensor ω_{ik} so determined is skewsymmetric descends directly from the shape of the equations Eq. (22), but it has to be imposed as a condition on the constants $\overline{\omega}_{ik}$.

sides dx_1, dx_2, dx_3 is transformed into a parallelepiped with sides

$$\begin{aligned} dy_1 &= (1+e_{11})dx_1 \\ dy_2 &= (1+e_{22})dx_2 \\ dy_3 &= (1+e_{33})dx_3 \,. \end{aligned} \tag{29}$$

Accordingly the volume $dV = dx_1 dx_2 dx_3$ is deformed to the first order into

$$dV' = (1+e_{11})(1+e_{22})(1+e_{33})dV \cong dV + (e_{11}+e_{22}+e_{33})dV \,,$$

neglecting higher order terms. This shows that

$$3 \cdot \epsilon = e_{11} + e_{22} + e_{33} = \frac{dV' - dV}{dV} \,, \tag{30}$$

justifying the name of cubic dilatation for Tre.

2. Equilibrium and stress

Before entering into our subject we want to underline that, from a theoretical mechanical point of view to start from "equilibrium" in analyzing the physical behaviour of a new force field, is not an intolerable restriction to the description of a frozen world, because dynamical laws can always be derived from statics by introducing inertia forces, i.e. mass times acceleration, to close the balance of otherways unbalanced static forces: this in fact is the deep meaning of 2nd and 3rd Newton's laws.

The field we want to investigate here is the field of forces internal to a solid body, arising as a consequence of an external action which causes a deformation of the body itself. The bodies we will consider are solid in the sense that, when isolated from any other external influence, we know that they will assume a configuration determined by the purely internal interatomic forces (even neglecting gravitational interaction) and they will keep it in time.

This is what is called "free" equilibrium configuration. This behaviour is indeed different from that of liquid bodies in which the atoms are not kept in a particular relative position but, rather, can run one on the other with the only bound that interpenetration is forbidden; typically a free liquid body can find an equilibrium configuration only under selfgravitation.

Between these two extrema, matter can adopt different physical behaviours displaying intermediate properties, like plasticity or viscoelasticity on which we will dwell later because it is specially this behaviour which can account for many geodynamical phenomena.

The basic assumption on the nature of this internal field is that each particle (or atom if you prefer) of the body interacts only with the neighbouring particles; this can happen for instance if the forces are generated by interatomic potentials,

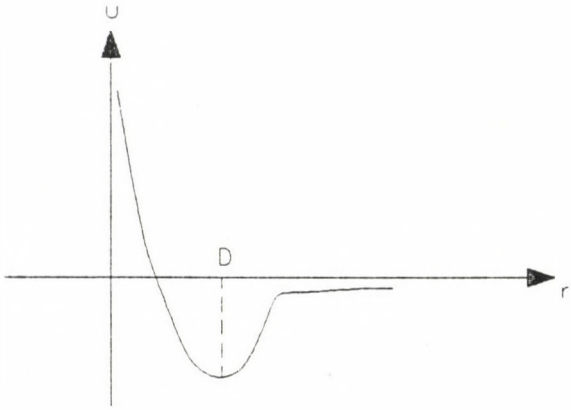

Fig. 5. Qualitative behaviour of an interatomic potential with D characteristic distance between atoms

basically functions of the distances only, like the one shown in Fig. 5, where one can see that already at distance 2D the effect of the potential becomes negligible.

Remark 2.1 When we say that the nature of internal forces is depending only on interatomic distances, we also imply that possible momentum or spin interactions are negligible with respect to these forces so that the body will not support and propagate torques which are independent from the pure position forces and their moment.

Equilibrium conditions. Let us consider now a body B which is in general in a strained configuration due to the action of external forces. These external forces can act on B either through the surface S, in which case they will be described by a vector field $\mathbf{T}$, of Cartesian components T_i, representing the density of force per unit surface, or they can affect directly the internal points of B, in which case they will be described by a vector field $\mathbf{F}$, with Cartesian components F_i, representing the density of force per unit volume of B.

An example of surface force is for instance the load created on the earth's crust because of the ice accumulation during a glaciation period; another example is hanging a load at the free end of a spring, the other end being fixed. An example of body force is the gravity field acting on a volume V occupied by a mass M, with density $\rho(\mathbf{x})$ in V.

By hypothesis we shall assume that B is in equilibrium and we consider a general subdomain B_0, all contained in B.

The normal $\boldsymbol{\nu}$ of B_0 is conventionally oriented into $B - B_0$ (cf. Fig. 6).

By hypothesis B_0 is in equilibrium so all internal forces in B_0 will balance. In fact if the atom i is subject to the force $\mathbf{f}_{ik}$ under the influence of the neighbouring atom k (both internal to B_0), then k is subject to a force

$$\mathbf{f}_{ki} = -\mathbf{f}_{ik} \tag{31}$$

due to Newton's third law, so when we add all the forces acting on the atoms of

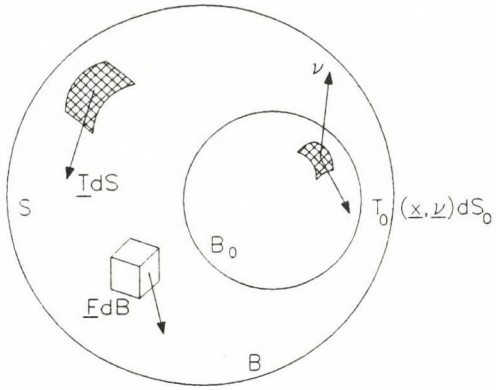

Fig. 6. The body B strained by external forces on S ($\mathbf{T}dS$) and in B ($\mathbf{F}dB$); the internal forces are such that through the (internal) area element dS_0, with normal $\boldsymbol{\nu}$, the outer part of B, $B - B_0$ (lying on the positive side of $\boldsymbol{\nu}$), acts on B_0 by means of the force $\mathbf{T}_0(\mathbf{x}, \boldsymbol{\nu})dS_0$

B_0 we are left with the sum of the body force and the specific forces which act on atoms of B_0 close to S_0, due to the atoms in $B - B_0$ also close to S_0.

Over an area element dS_0, oriented by its normal $\boldsymbol{\nu}$, we will have a force which in general is function of the position $\mathbf{x}$ and of the normal $\boldsymbol{\nu}$, $\mathbf{T}_0(\mathbf{x}, \boldsymbol{\nu})dS_0$.

Always because of Newton's 3rd law we must have

$$\mathbf{T}_0(\mathbf{x}, -\boldsymbol{\nu}) = -T_0(\mathbf{x}, \boldsymbol{\nu}), \tag{32}$$

but we will need the exact dependence of $\mathbf{T}_0$ on $\boldsymbol{\nu}$, and we shall show in a while that $\mathbf{T}_0$ is linear in $\boldsymbol{\nu}$, *i.e.*

$$T_{0i}(\mathbf{x}, \boldsymbol{\nu}) = \sigma_{ik}(\mathbf{x})\nu_k \quad , \tag{33}$$

where the convention of summation over repeated indexes is adopted.

The tensor $\sigma_{ik}(\mathbf{x})$ is by definition the stress tensor.[4]

If Eq. (33) is true, by using the above results and reasonings, we see that the equilibrium condition for B_0 reads

$$\int_{B_0} F_i \, dB + \int_{S_0} T_{0i}(\mathbf{x}, \boldsymbol{\nu}) dS_0 = 0 \quad i = 1, 2, 3$$

or

$$\int_{B_0} F_i \, dB + \int_{S_0} \sigma_{ik}(\mathbf{x})\nu_k \, dS_0 = 0 \, . \tag{34}$$

[4]The relation Eq. (33) is in Cartesian coordinates, yet the tensor character of σ_{ik} is easily argued from the fact that its saturation with the vector field ν_k gives rise to another vector field; hence at the cost of properly defining the components σ_{ik} a similar relation holds in general coordinates.

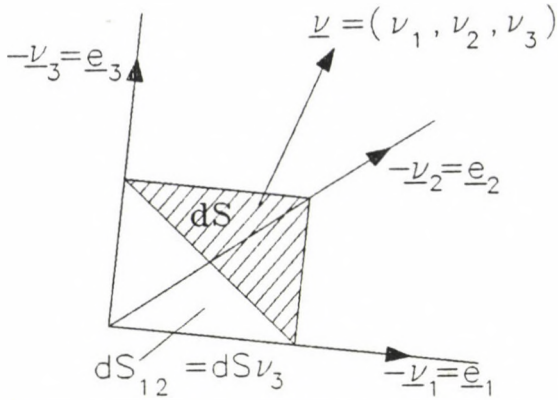

Fig. 7. A material tetrahedron. Note that the normals to the three orthogonal faces are opposite to the unit vectors of coordinate axes

On the other hand, by using the Gauss theorem,

$$\int_{B_0} \frac{\partial}{\partial x_k} \sigma_{ik} \, dB = \int_{S_0} \sigma_{ik} \nu_k \, dS_0$$

we see that Eq. (34) can be written

$$\int_{B_0} \left(F_i + \frac{\partial}{\partial x_k} \sigma_{ik} \right) \, dB = 0 \,, \tag{35}$$

i.e. considering that B_0 is arbitrary,

$$F_i(\mathbf{x}) + \frac{\partial}{\partial x_k} \sigma_{ik}(\mathbf{x}) = 0 \,; \tag{36}$$

these are the indefinite equilibrium conditions which have to hold throughout the body B.

To conclude our proof, we need to verify Eq. (33).

Lemma 2.1: the internal surface force-density $\mathbf{T}_0(\mathbf{x}, \boldsymbol{\nu})$ depends linearly on $\boldsymbol{\nu}$.

In fact let us take a tetrahedron with three faces parallel to coordinate planes as in Fig. 7.

As we see the equilibrium condition for this is expressed by

$$\begin{aligned} \mathbf{T}_0(\mathbf{x}, \boldsymbol{\nu}) dS \quad &+ \mathbf{T}_0(\mathbf{x}, \boldsymbol{\nu}_3) dS_{12} + \mathbf{T}_0(\mathbf{x}, \boldsymbol{\nu}_2) dS_{13} + \\ &+ \mathbf{T}_0(\mathbf{x}, \boldsymbol{\nu}_1) dS_{23} + \mathbf{F}(\mathbf{x}) dV = 0 \,. \end{aligned} \tag{37}$$

Since

$$dS_{12} = dS\nu_3, \quad dS_{13} = dS\nu_2, \quad dS_{23} = dS\nu_1 \tag{38}$$

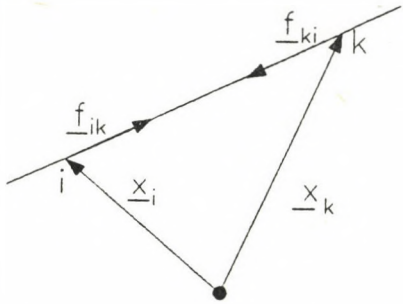

Fig. 8. Internal forces generale no couples

and since dV is an infinitesimal of higher order than dS, also recalling Eq. (32), we see that

$$\begin{aligned} \mathbf{T}_0(\mathbf{x},\boldsymbol{\nu}) &= \mathbf{T}_0(\mathbf{x},e_3)\nu_3 + \mathbf{T}_0(\mathbf{x},e_2)\nu_2 + \mathbf{T}_0(\mathbf{x},e_1)\nu_1 = \\ &= \mathbf{T}_0(\mathbf{x},e_3)\boldsymbol{\nu}\cdot e_3 + \mathbf{T}_0(\mathbf{x},e_2)\boldsymbol{\nu}\cdot e_2 + \mathbf{T}_0(\mathbf{x},e_1)\boldsymbol{\nu}\cdot e_1 \,. \end{aligned} \tag{39}$$

This relation expresses in fact the linear dependence of $\mathbf{T}_0$ on $\boldsymbol{\nu}$ and coincides with Eq. (33) if we put

$$[\mathbf{T}_0(\mathbf{x},e_k)]_i = \sigma_{ik} \,. \tag{40}$$

Lemma 2.2: Symmetry of the stress tensor. We can observe that according to our previous reasoning not only the internal forces balance each other but, as far as two atoms i and k belong to the same domain $B_0 \subset B$, due to Eq. (31) they also have a total zero momentum as it is easily recognized by inspecting Fig. 8 where one sees that

$$\mathbf{x}_i \wedge \mathbf{f}_{ik} + \mathbf{x}_k \wedge \mathbf{f}_{ki} = (\mathbf{x}_i - \mathbf{x}_k) \wedge \mathbf{f}_{ik} = 0 \,. \tag{41}$$

So reasoning as before we find that, since B_0 has to be in equilibrium, we must have

$$\int_{B_0} \mathbf{x} \wedge \mathbf{F} dB + \int_{S_0} \mathbf{x} \wedge \mathbf{T}_0(\mathbf{x},\boldsymbol{\nu}) dS \equiv 0 \,, \tag{42}$$

or recalling Eq. (33) and using a component notation

$$\begin{aligned} &\int_{B_0} (x_i F_k - x_k F_i) dB + \int_{S_0} (x_i \sigma_{k\ell} - x_k \sigma_{i\ell}) \nu_\ell dS_0 \equiv \\ &\equiv \int_{B_0} \left\{ (x_i F_k - x_k F_i) + \frac{\partial}{\partial x_\ell}(x_i \sigma_{k\ell} - x_k \sigma_{i\ell}) \right\} dB \equiv 0 \,. \end{aligned}$$

Since B_0 is arbitrary the expression in parentheses has to be identically zero; since $F_k = \frac{\partial \sigma_{k\ell}}{\partial x_\ell}$ and $F_i = \frac{\partial \sigma i\ell}{\partial x_\ell}$ this expression reduces to

$$\frac{\partial x_i}{\partial x_\ell}\sigma_{k\ell} - \frac{\partial x_k}{\partial x_\ell}\sigma_{i\ell} \equiv \sigma_{ki} - \sigma_{ik} \equiv 0 \,; \tag{43}$$

namely it turns out that the stress tensor σ has to be symmetric.

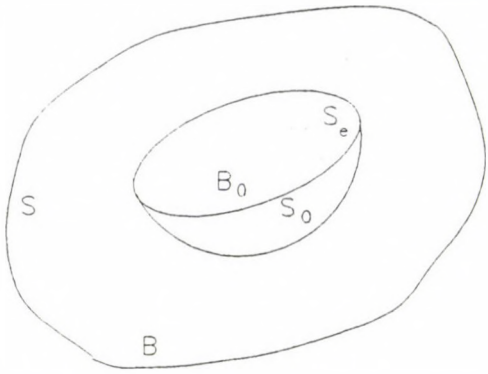

Fig. 9. Subdomain $B_0 \subset B$ with part of common boundary S_e

Remark 2.2 (surface stress). We still have to see the relation existing between the stress tensor and the external forces acting through the external surface of the body (stress vector).

To this aim let us consider a subdomain B_0, subset of B, which has a part of boundary, S_e, in common with S, ($S_e \subset S$) (cf. Fig. 9); we call S_0 the part of boundary of B_0 which is internal to B.

The equilibrium conditions of B_0 write

$$\int_{B_0} F_i \, dB + \int_{S_e} T_i \, dS + \int_{S_0} T_{0i} \, dS = 0 \,, \tag{44}$$

where T_i are the components of the external stress tensor, $T_{0i} = \sigma_{ik}\nu_k$ on S_0 according to Eq. (33) and $F_i = -\frac{\partial \sigma_{ik}}{\partial x_k}$ at points internal to B_0 which are internal to B too.

Whence

$$\int_{B_0} -\frac{\partial \sigma_{ik}}{\partial x_k} \, dB + \int_{S_e} T_i dS + \int_{S_0} \sigma_{ik}\nu_k \, dS = 0 \,,$$

so that making use of Gauss theorem applied to the first term, we find

$$\int_{S_e} \{T_i - \sigma_{ik}\nu_k\} \, dS \equiv 0 \,. \tag{45}$$

Since the domain S_e is arbitrary in S, we find

$$T_i(\mathbf{x}) = \sigma_{ik}(\mathbf{x})\nu_k(\mathbf{x}) \tag{46}$$

which is the sought boundary relation.

We note that in particular when the surface S, or part of it, is free of stresses, the boundary condition

$$\sigma_{ik}\nu_k = 0 \tag{47}$$

has to be satisfied on it.

Remark 2.3 The quantity

$$N(\mathbf{x}) = \nu_i \sigma_{ik} \nu_k \,, \tag{48}$$

i.e. the quadratic form associated with stress tensor at $\mathbf{x}$ and to a "normal" direction $\boldsymbol{\nu}$, represents the density of stress through the surface element dS with normal $\boldsymbol{\nu}$, again projected along the normal; from here the name of *normal stress* usually given to $N(\mathbf{x})$.

Naturally there will be 3 principal directions $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \boldsymbol{\nu}_3$ mutually orthogonal, such that the stress vector $\sigma_{ik}\nu_{1k}$ is again directed as $\boldsymbol{\nu}_1$ *etc.*, in this case

$$\begin{aligned} \sigma\boldsymbol{\nu}_1 &= \sigma_{11}\boldsymbol{\nu}_1 \\ \sigma\boldsymbol{\nu}_2 &= \sigma_{22}\boldsymbol{\nu}_2 \\ \sigma\boldsymbol{\nu}_3 &= \sigma_{33}\boldsymbol{\nu}_3 \end{aligned} \tag{49}$$

and $\sigma_{11}, \sigma_{22}, \sigma_{33}$, represent the normal stresses in the three direction $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \boldsymbol{\nu}_3$.

The tensor σ, in a coordinate system with axes parallel to the principal directions, will have a purely diagonal form; in particular, if we assume that

$$\sigma_{11} = \sigma_{22} = \sigma_{33} = -p$$

we have

$$\sigma_{ik} = -p\delta_{ik} \,. \tag{50}$$

Corresponding to Eq. (50) we see that the normal stress is

$$N(\mathbf{x}) = \nu_i \sigma_{ki} \nu_k = -p\nu_i\nu_i = -p \,; \tag{51}$$

this shows that Eq. (50) represents a situation of isotropic compression.

As we did for the strain tensor, also the stress tensor is sometimes usefully split into a purely isotropic part and a stress deviation with zero trace

$$\sigma = -pI + \sigma' \tag{52}$$

with

$$-p = \frac{1}{3} Tr\sigma \,, \quad Tr\sigma' = 0 \,. \tag{53}$$

Remark 2.4 (statically admissible stresses). It is natural to put the question whether the equilibrium conditions Eq. (36), which form a system of partial differential equations, together with boundary conditions Eq. (46) can determine univocally the stress field. The answer is in the negative.

As a matter of fact it is enough to choose

$$\sigma_{ik} = F_{ik}(x_\ell, x_j) \,,$$

$$\left(\begin{array}{ll} k\ell j & = \text{ permutation of } (1\ 2\ 3) \\ i & \geq k \\ \sigma_{ki} & = \sigma_{ik} \end{array} \right)$$

and one immediately realizes that the homogeneous equilibrium conditions are satisfied

$$\frac{\partial \sigma_{ik}}{\partial x_k} \equiv 0 .$$

A little more refined analysis runs as follows: let us take the strain tensor e related to an arbitrary deformation field $\mathbf{u}$ in B and let us consider the space of all second order, square integrable (in B) symmetric tensors $L^2_{2,S}(B)$ with the usual scalar product

$$(T, S) = \int_B T_{ik} S_{ik} dB ; \tag{54}$$

then tensors $e(\mathbf{u})$ cannot be a dense subset in $L^2_{2,S}$. In fact $e(\mathbf{u})$ are basically constructed by means of 3 fields only (u_i, $i = 1, 2, 3$) while tensors in $L^2_{2,S}$ can be constructed by means of 6 arbitrary fields; this is as a matter of fact the deep reason why St. Venant conditions have to hold.

Consequently there are $\sigma \in L^2_{2,S}$ such that

$$(\sigma, e(\mathbf{u})) \equiv 0 \quad \forall \mathbf{u} . \tag{55}$$

Accordingly, also recalling that σ_{ik} has to be symmetric,

$$\begin{aligned} 0 \; &\equiv \textstyle\int_B \sigma_{ik}(\partial_i u_k + \partial_k u_i) dB = 2 \int_B \sigma_{ik} \partial_k u_i dB = \\ &= 2 \textstyle\int_S \nu_k \sigma_{ik} u_i dS - 2 \int_B \partial_k \sigma_{ik} u_i dB \end{aligned}$$

which, due to the arbitrariness of u_i, imply

$$\partial_k \sigma_{ik} = 0 \qquad \text{in } B$$

$$\sigma_{ik} \nu_k = 0 \quad \text{on } S ,$$

showing that for a given configuration of body forces F_i and surface stress vector T_i there are many stress σ_{ik} satisfying the equilibrium conditions

$$\begin{cases} \dfrac{\partial \sigma_{ik}}{\partial x_k} + F_i = 0 \\ \\ \sigma_{ik} \nu_k = T_i ; \end{cases} \tag{56}$$

the class of all such tensors is called the class of *statically admissible stress fields.*

We note explicitely that in Eq. (56) the second condition could be given not over the whole boundary S, but just on a part S_T of it.

3. Hooke's law and its generalizations

(*"Ut tensio sic vis"* Robert Hooke 1678)

The original form of Hooke's law was stating that the extension ϵ of a spring or of wire of elastic material with a fixed extreme

$$\epsilon = \frac{\delta l}{l}$$

is proportional to the load T applied to the other extreme

$$\epsilon = \frac{T}{E},$$

the constant E being the so called Young modulus.

Following Cauchy we say that the body considered is in elastic regime when the relation between stress and strain tensors is linear, i.e. in general

$$\sigma_{ij} = c_{ijk\ell} e_{k\ell} \,. \tag{57}$$

In general $c_{ijk\ell}$ has to satisfy symmetry conditions like

$$c_{ijk\ell} = c_{jik\ell} = c_{ij\ell k} \tag{58}$$

related to the analogous symmetries of σ and e: energetic and thermodynamical considerations suggest that the further property

$$c_{ijk\ell} = c_{k\ell ij} \tag{59}$$

holds too.

Here we shall work with isotropic bodies only. This implies first of all that there is a system of coordinates which diagonalizes contemporarily both σ and e. In fact let us make the following elementary consideration: a uniform compression of a small cube of an isotropic material will also produce a uniform contraction in all directions, for otherwise we would have a means to identify a privileged direction in our material.

Now let us take a small cube centered at the origin and imagine first to apply a pressure P on the two faces orthogonal to the x_1 axis; this will produce a contraction along x_1, but on the same time a dilation in the x_2, x_3 directions because the atoms getting closer than the free equilibrium position, tend to repulse each other (cf. the qualitative behaviour of the potential in Fig. 5).

Now we add a further uniform compression p in order to annul the dilation in the x_2, x_3 directions; as a final situation we will have a pressure $P+p$ in the x_1 direction and pressure p in the x_2, x_3 directions corresponding to a pure contraction along x_1 (i.e. $u_1 = -\epsilon x_1$) and to zero displacements along x_2 and x_3 (i.e. $u_2 \equiv u_3 \equiv 0$).

The sequence is illustrated in Fig. 10.

Accordingly, with a strain tensor of the form

$$e = \begin{pmatrix} e_{11} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad (e_{11} = -\epsilon_1) \tag{60}$$

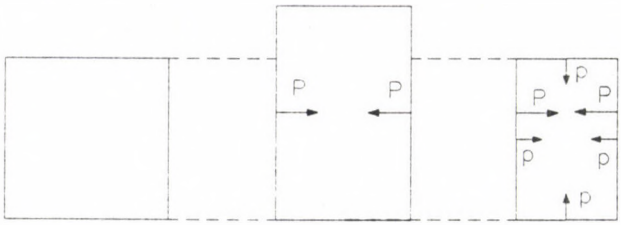

Fig. 10. Sequence producing a pure x_1 contraction

we have a stress of the form

$$\sigma = \begin{pmatrix} -(P_1 + p_1) & 0 & 0 \\ 0 & -p_1 & 0 \\ 0 & 0 & -p_1 \end{pmatrix} ; \tag{61}$$

moreover under the hypothesis of a linear dependence of σ on e we expect that

$$\begin{aligned} -P_1 &= 2\mu e_{11} \\ -p_1 &= \lambda e_{11} , \end{aligned} \tag{62}$$

noting that, since $e_{11} < 0$, we can also say that

$$\lambda > 0, \quad \mu > 0 . \tag{63}$$

Now if we have a purely diagonal strain

$$e = \begin{pmatrix} e_{11} & 0 & 0 \\ 0 & e_{22} & 0 \\ 0 & 0 & e_{33} \end{pmatrix} , \tag{64}$$

i.e. the sum of three tensors representing pure contractions (or dilations) along x_1, x_2, x_3 the corresponding stress must be

$$\sigma = \begin{pmatrix} P_1 + \sum p_k & 0 & 0 \\ 0 & P_2 + \sum p_k & 0 \\ 0 & 0 & P_3 + \sum p_k \end{pmatrix} , \tag{65}$$

where due to the isotropy of the body we must have

$$\begin{aligned} -P_k &= 2\mu e_{kk} \\ -p_k &= \lambda e_{kk} , \end{aligned} \tag{66}$$

with the same (positive) constants λ, μ for *all directions.*

Formulas Eq. (64) and Eq. (65) show that to a diagonal strain tensor corresponds a diagonal stress; in particular we see that Eq. (65) and Eq. (66) can be written as

$$\sigma = 2\mu \; e + \lambda (Tre) I , \tag{67}$$

and since this is a tensorial relation it has hold in any other coordinate system, in particular in any Cartesian system rotated with respect to the one in which e is diagonal.

We write Eq. (67) in components as

$$\sigma_{ik} = 2\mu e_{ik} + \lambda(e_{\ell\ell})\delta_{ik} \,. \tag{68}$$

The constants (λ, μ) are called the Lamé coefficients and in particular μ is the so called shear modulus.

Recalling Eq. (27) we see that Eq. (67) can be written also as

$$\begin{gathered} \sigma = (2\mu + 3\lambda)\epsilon I + 2\mu e' \\ (\epsilon = \tfrac{1}{3}Tre, Tre' = 0) \end{gathered} \tag{69}$$

in particular, recalling Eq. (52), we have

$$-p = \frac{1}{3}Tr\sigma = (2\mu + 3\lambda)\epsilon \,. \tag{70}$$

Moreover since $3 \cdot \epsilon$ has the meaning of rate of volume variation (naturally ϵ is negative when p is positive), we see that the constant

$$k = \lambda + \frac{2}{3}\mu \,, \tag{71}$$

called bulk modulus, represents the rate of volume decrease per unit of uniform pressure p. The constant k has also to be positive for physical reasons.

Hooke's law can therefore be written as

$$\begin{gathered} \sigma = k(Tre)I + 2\mu e' \,, \\ \sigma_{ik} = k(\textstyle\sum e_{\ell\ell})\delta_{ik} + 2\mu \left\{ e_{ik} - \tfrac{1}{3}(\textstyle\sum e_{\ell\ell})\delta_{ik} \right\} \,. \end{gathered} \tag{72}$$

Other constants often used in elasticity theory are the already mentioned Young's modulus E

$$E = \frac{2\mu \cdot k}{\lambda + \mu} \tag{73}$$

which has to be a positive constant too, and the Poisson ratio

$$\nu = \frac{\lambda}{2(\lambda + \mu)} \,; \tag{74}$$

from the positivity of λ and μ, Eq. (74) implies

$$0 < \nu < \frac{1}{2} \,.$$

In terms of these constants Hooke's law writes

$$\sigma_{ik} = \frac{E}{1+\nu}\left[e_{ik} + \frac{\nu}{1-2\nu}\left(\sum e_{\ell\ell}\right)\delta_{ik} \right] \tag{75}$$

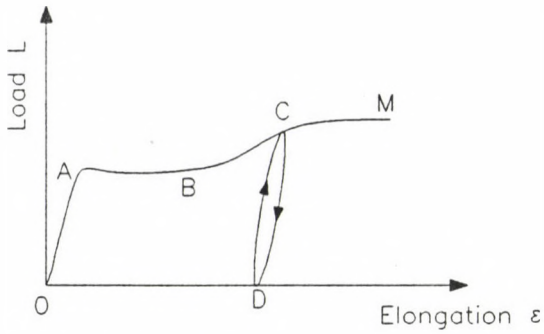

Fig. 11. Load-elongation diagram for mild steel

Just to have an idea of the order of magnitude of these constants let us report a table of values determined by laboratory experiments; E is given in Newton per meter squared, while ν is adimensional.

	$E(Nm^{-2})$	ν
Magnesium	$0.37 \cdot 10^{11}$	0.35
Aluminium	$0.73 \cdot 10^{11}$	0.34
Copper	$1.18 \cdot 10^{11}$	0.35
Iron	$2.13 \cdot 10^{11}$	0.28
Steel	$2.15 \cdot 10^{11}$	0.29

One important feature of the parameter ν is that its range of variability is quite small, say between 0.25 an 0.35 for most materials, so that sometimes assuming that it is constant can be not too bad an approximation. By the way we remark that $\nu = 0.25$ corresponds to $\lambda = \mu$.

Let us also observe that the relation Eq. (75) can be easily inverted to give e as function of σ, *i.e.*

$$e_{ik} = \frac{1}{E}\left\{(1+\nu)\sigma_{ik} - \nu\left(\sum \sigma_{\ell\ell}\right)\delta_{ik}\right\} \tag{76}$$

Plasticity. In a one-dimensional load-elongation experiment with a mild steel wire one would find a behaviour like that in Fig. 11.

We can distinguish on this diagram the following different phases:
$\overline{OA}$: linear elastic regime $\epsilon = \frac{L}{SE}$ (S = section);
$\overline{AB}$: plastic behaviour; keeping the load constant one has a material flow $\left(\frac{d}{dt}\epsilon \neq 0\right)$;
$\overline{BM}$: work hardening behaviour up to the strength limit L_M.

The point A at which plastic behaviour starts is called the yield point; slightly below A there is a limit point E corresponding to the maximum load L_E that one can apply, returning to the original length when it is removed. At any point (*e.g.* C in Fig. 11) beyond A one would unload the wire, one recovers only a small part of elongation, corresponding to the purely elastic component, like CD in Fig. 11; reloading brings back to the same point C.

It is important to remark that to describe a phenomenon of this kind it is not enough to generalize Hooke's relations in the sense of establishing non-linear

constitutive equations, these in fact are not able to account for the phenomenon of the residual deformation which is left in the body even upon complete unloading.

Such a behaviour can be described only by the so called incremental theory.

The idea is that at each point $\mathbf{x}$ if we follow the history of $\sigma_{ik}(\mathbf{x}, t)$ in time we have no plastic behaviour until a so called "yield condition" is reached, typically written in the form

$$f(\sigma(t)) = k^2 . \tag{77}$$

So, for $f(\sigma) < k^2$ we have no plasticity and this is switched on only when Eq. (77) is verified; plastic flow then continues to occur as long as Eq. (77) is fulfilled and it stops only in case that f drops below k^2. When k^2 is a pure constant we say that we have pure plasticity; if on the contrary k^2 depends on previous plastic deformation we have a so called work hardening behaviour. As for the relation between stress and strain we can maintain that

$$e_{ij} = e^e_{ij} + e^p_{ij} \tag{78}$$

where e^e, the elastic component, follows an ordinary Hooke's law, *e.g.*

$$e^e_{ij} = \frac{1}{E}\left\{(1+\nu)\sigma_{ik} - \nu\left(\sum \sigma_{\ell\ell}\right)\delta_{ik}\right\} ,$$

while e^p_{ij}, the plastic component, satisfies the differential equation

$$e^p_{ij} = \frac{\alpha}{\eta} \cdot \frac{\partial f}{\partial \sigma_{ij}} \tag{79}$$

with

$$\begin{cases} \alpha = 0 & \text{if } f(\sigma) < k^2 , \quad \forall \dot{f} \\ & \text{or } f(\sigma) = k^2 , \quad \dot{f} < 0 \\ \alpha = 1 & \text{if } f(\sigma) = k^2 , \quad \dot{f} = 0 . \end{cases} \tag{80}$$

One important remark is that experience suggests that plasticity is a pure shear phenomenon, i.e. there is no plastic volume change; therefore the yield function f is usually taken as function of the stress deviation $\sigma' = \sigma - \frac{1}{3}(Tr\sigma)$. The simplest of such yield functions is the one by von Mises

$$f(\sigma') = Tr(\sigma')^2 , \tag{81}$$

providing through Eq. (79) a plastic flow e^p basically proportional to σ'; in this case the proportionality constant η has the same meaning and dimensions as a viscosity coefficient.

Viscoelasticity. The theory of plasticity with its basic non-linear character is difficult to be dealt with. Yet in many applications it is important to be able to describe flow phenomena which produce unrecoverable deformation: in particular this is certainly true in geodynamical applications of the theory.

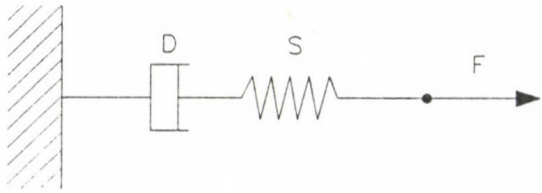

Fig. 12. Schematic representation of a Maxwell body with spring S and dashpot D in series

This is probably the reason why viscoelasticity or the theory of linear hereditary bodies, as they were called by Vito Volterra, has become so popular nowadays.

The idea here is that instead of having a flow only after a certain threshold has been reached, we have a flow at any moment simply proportional to stress.

This is illustrated in one dimension in Fig. 12, where a so called Maxwell model is presented, according to which the elongation is the combination of two effects, one purely elastic represented by a spring, the other viscous represented by a dashpot allowing a flow proportional to the load F.

The two effects can be combined in this case into the unique equation

$$\dot{u} = \frac{\dot{F}}{a} + \frac{F}{\eta} . \tag{82}$$

More generally a linear hereditary body is one for which deformation at time t is a linear function of stress at present and past times.

If this relation is further assumed to be homogeneous in time, the above description can be conveniently cast in mathematical form as

$$e_{ik}(\mathbf{x}, t) = \int_{-\infty}^{t} C_{ik\ell m}(\mathbf{x}, t - \tau) d\sigma_{\ell m}(\mathbf{x}, \tau) \tag{83}$$

where the tensor $C_{ik\ell m}$ is called the "creep function". If, as we shall assume, there is a certain time before which neither strain nor stress were present in the body, we can take this time as origin and we can write

$$e_{ik}(t) = \int_{0}^{t} C_{ik\ell m}(t - \tau) d\sigma_{\ell m}(\tau) \tag{84}$$

where we will implicitely admit that e and σ are continuous tensors so that, *e.g.*

$$e_{ik}(0) = 0 , \quad \sigma_{em}(0) = 0 . \tag{85}$$

Equation (84), when inverted, will give a relation of the same type

$$e_{ik}(t) = \int_{0}^{t} R_{ik\ell m}(t - \tau) de_{\ell m}(\tau) \tag{86}$$

with the tensor $R_{ik\ell m}$ called in this case the "relaxation function".

If we make the further assumption that the body in question is isotropic, we can, by analogy with the general Hooke's law, conclude that

$$e_{ik}(t) = \int_0^t C_1(t-\tau)d\sigma_{ik}(\tau) + \sum_\ell \int_0^t C_2(t-\tau)d\sigma_{\ell\ell}(\tau)\delta_{ik} \tag{87}$$

$$\sigma_{ik}(t) = \int_0^t R_1(t-\tau)de_{ik}(\tau) + \sum_\ell \int_0^t R_2(t-\tau)de_{\ell\ell}(\tau)\delta_{ik}\,. \tag{88}$$

Moreover, by suitably exploiting the conditions Eq. (85) one realizes that

$$\int_0^t C(t-\tau)d\sigma_{ik}(\tau) = \partial_t \int_0^t C(t-\tau)\sigma_{ik}(\tau)d\tau = \partial_t C * \sigma_{ik} \tag{89}$$

where the customary notation for the convolution product has been used.

With this notation we find that the relations Eq. (87), Eq. (88) can be written concisely as

$$e = \partial_t \{C_1 * \sigma + (C_2 * Tr\sigma)I\} \tag{90}$$

$$\sigma = \partial_t \{R_1 * e + (R_2 * Tre)I\}\,. \tag{91}$$

Remark 3.1. We want to stress that a relation of the type Eq. (82) can indeed be represented in the form

$$u(t) = \partial_t \{C * F\}$$

with

$$C(t-\tau) = \frac{1}{a} + \frac{t-\tau}{\eta}$$

as it can be verified by a direct calculation.

Remark 3.2. Due to the particular shape of hereditary constitutive equations we are pushed to use a Laplace transform approach, which simplifies considerably Eqs (90), (91).

In fact, recalling that by definition

$$\mathcal{L}(f)(s) = \int_0^{+\infty} e^{-st} f(t)dt\,, \tag{92}$$

the following properties are well known and easy to check:

$$\mathcal{L}(\dot{f})(s) = -f(0) + s\mathcal{L}(f)(s) \tag{93}$$

$$\mathcal{L}(f * g)(s) = \mathcal{L}(f) \cdot \mathcal{L}(g) . \tag{94}$$

By using these properties and recalling the conditions Eq. (85), representing with an over-bar the operation of Laplace transformation

$$\overline{f}(s) = \mathcal{L}(f)(s) , \tag{95}$$

we get from Eq. (90), Eq. (91)

$$\overline{e} = (s\overline{C}_1) \cdot \overline{\sigma} + (s\overline{C}_2) Tr\overline{\sigma} I \tag{96}$$

$$\overline{\sigma} = (s\overline{R}_1) \cdot \overline{e} + (s\overline{R}_2) Tr\overline{e} I . \tag{97}$$

Based on this last remark one can really arrive to formulate an equivalence principle between elasticity and viscoelasticity theory:

Equivalence principle any viscoelastic problem can be solved by formulating the corresponding static elastic problem in which the usual constants λ, μ or E, ν are substituted by functions of s, in the Laplace transform domain, $\overline{\lambda}, \overline{\mu}$ or $\overline{E}, \overline{\nu}$ related to the creep functions C_1, C_2 or relaxation functions R_1, R_2 by the equations

$$\begin{cases} \overline{\mu}(s) & = \frac{1}{2} s\overline{R}_1 \\ \overline{\lambda}(s) & = s\overline{R}_2 \end{cases} \tag{98}$$

$$\begin{cases} \overline{E}(s) & = \frac{1}{s(\overline{C}_1 + \overline{C}_2)} \\ \overline{\nu}(s) & = -\frac{\overline{C}_2}{\overline{C}_1 \overline{C}_2} . \end{cases} \tag{99}$$

We will see at the end of the next paragraph a discussion justifying this principle.

4. Boundary value problems in elastic and viscoelastic deformation theory

Elastic small displacement theory. We start this paragraph by treating *equilibrium problems.*

We first of all observe that the combination of the following three elements

1. the strain-displacement relation

$$e = e(\mathbf{u}) , \tag{100}$$

i.e. in small displacement theory

$$e_{ik} = \frac{1}{2} \left(\frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i} \right) ; \tag{101}$$

2. the general equilibrium conditions, *i.e.*

$$\frac{\partial \sigma_{ik}}{\partial x_k} + F_i = 0 \tag{102}$$

$$\sigma_{ik}\nu_k = T_i \tag{103}$$

where Eq. (102) holds inside the body B, while Eq. (103) holds on the surface S;

3. constitutive equations specifying the type of body we are considering and expressing a relation between σ and e, *e.g.* Hooke's law for elastic isotropic bodies

$$\sigma = 2\mu e + \lambda(Tre)I \tag{104}$$

provides a natural means of expressing deformation phenomena in terms of partial differential equations.

In fact using Eq. (101) in Eq. (104) and the results in Eq. (102) we get, assuming that Lamé parameters can be considered as constants,

$$\sigma_{ik} = \mu\left(\frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i}\right) + \lambda\left(\sum_j \frac{\partial u_j}{\partial x_j}\right)\delta_{ik} \tag{105}$$

$$\mu\sum_k \frac{\partial^2 u_i}{\partial x_k} + (\lambda+\mu)\frac{\partial}{\partial x_i}\sum_k \frac{\partial u_k}{\partial x_k} + F_i = 0 \tag{106}$$

or, in vector notation

$$\mu\Delta\mathbf{u} + (\lambda+\mu)\nabla(\nabla\cdot u) + \mathbf{F} = 0\,; \tag{107}$$

this is known as Lamé equation.

Sometimes Eq. (107) is rewritten in the equivalent form

$$(2\mu+\lambda)\nabla(\nabla\cdot\mathbf{u}) - \mu\nabla\wedge(\nabla\wedge\mathbf{u}) + \mathbf{F} = 0\,. \tag{108}$$

Formally, *i.e.* disregarding problems of the data assigned at the boundary and problems of regularity, we can see that $\nabla\cdot\mathbf{u}$ and $\nabla\wedge\mathbf{u}$ could be determined separately.

In fact by taking $\nabla\cdot$ and $\nabla\wedge$ of Eq. (108) after rearrangements we find

$$\begin{cases} (2\mu+\lambda)\,\Delta\,(\nabla\cdot\mathbf{u}) & = -\nabla\cdot\mathbf{F} \\ \mu\,\Delta\,(\nabla\wedge\mathbf{u}) & = \nabla\wedge\mathbf{F}\,; \end{cases} \tag{109}$$

in particular when $\mathbf{F}$ is the gradient of a harmonic field, as it is the case with gravity for instance, both $\nabla\cdot\mathbf{u}$ and $\nabla\wedge\mathbf{u}$ are harmonic.

Now that we have a partial differential equation for $\mathbf{u}$ we must think of the subsidiary conditions which help in identifying the particular solution sought.

As we know these auxiliary conditions (boundary values, initial conditions, etc.) give rise to properly posed problems, depending on the nature of the differential operator; for instance an elliptic differential operator should be associated with B.V.P. (Boundary Value Problems). So before going into details we establish the following Lemma.

Lemma 4.1: the Lamé differental operator

$$L = \mu(\triangle)I + (\lambda + \mu) \nabla \nabla^{+} \tag{110}$$

is elliptic, and more precisely strictly negative definite, on a space of vector functions $\mathbf{u}$ sufficiently regular and such that

$$\mathbf{u} \mid_S \equiv 0\,, \tag{111}$$

on condition that

$$\mu \geq \mu_0 > 0\,. \tag{112}$$

The point here is just to evaluate the quadratic form

$$(\mathbf{u}, L\mathbf{u}) = \int\limits_B \{\mu(\mathbf{u} \cdot \Delta\mathbf{u}) + (\lambda + \mu)\mathbf{u} \cdot \nabla(\nabla \cdot \mathbf{u})\}\, dB\,;$$

by applying Gauss theorem and condition Eq. (111) we have

$$\begin{cases} (\mathbf{u}, L\mathbf{u}) = -\int\limits_B \left\{\mu TrUU^{+} + (\lambda + \mu)(TrU)^2\right\} dB \\ U = \left[\dfrac{\partial u_i}{\partial x_k}\right] = \nabla\mathbf{u} \end{cases} \tag{113}$$

so that

$$(\mathbf{u}, L\mathbf{u}) \leq -\mu_0 \int\limits_B TrUU^{+} dB \leq 0\,. \tag{114}$$

In particular we can never have

$$(\mathbf{u}, L\mathbf{u}) = 0 \tag{115}$$

for $\mathbf{u} \neq 0$ because this would imply

$$TrUU^{+} = 0 \quad \rightarrow U = \left[\frac{\partial u_i}{\partial x_k}\right] \equiv 0 \rightarrow u_i = \text{const.}$$

what is not possible because $u_i \mid_S = 0$ implying that $u_i \equiv 0$ throughout B.

According to the Lemma 4.1 just proved we would expect that Lamé equation should be complemented by boundary value data. More precisely we can state the following B.V.P.'s:

1. we give on the boundary S the displacement vector $\mathbf{u}$; for technical reasons this is best done in the following way: we define a sufficiently regular function $\mathbf{u}_0$ which on S attains the wanted behaviour, but it is otherwise arbitrary and then we stipulate that
$$\mathbf{u} - \mathbf{u}_0 \,|_S = 0\,; \tag{116}$$

2. we give on the boundary S the stress vector $\mathbf{T}$
$$\begin{gathered} \sigma\boldsymbol{\nu} \,|_S = \mathbf{T} \\ (\sigma_{ik}\nu_k \,|_S = T_i)\,, \end{gathered} \tag{117}$$
or, with the same reasoning as in 1, defined a suitable $\mathbf{T}_0$
$$\sigma\boldsymbol{\nu} - \mathbf{T}_0 \,|_S = 0\,;$$

3. we split the boundary S into two complementary parts
$$S = S_u \cup S_T\,, \quad S_u \cap S_T = 0$$
and, having defined functions $\mathbf{u}_0$, $\mathbf{T}_0$ which on S_u, S_T attain wanted given values, we stipulate
$$\begin{cases} \mathbf{u} - \mathbf{u}_0 \,|_{S_u} & = 0 \\ \sigma\boldsymbol{\nu} - \mathbf{T}_o \,|_{S_T} & = 0\,. \end{cases} \tag{118}$$

We will give some information in the next paragraph on the way these problems can be analyzed from the mathematical point of view, here we shall restrict ourselves to a couple of remarks in particular to clarify the deep relation existing between elasticity theory and the biharmonic equation.

Remark 4.1. If we take the first of Eq. (109) into account and apply a Laplace operator to Eq. (107), we find
$$\mu\Delta^2\mathbf{u} = -\mathbf{F} + \frac{\lambda+\mu}{2\mu+\lambda}\,\nabla\,(\nabla\cdot\mathbf{F}) \tag{119}$$
showing that the elastic displacement field of a body not submitted to body forces is always biharmonic. Naturally Eq. (119) is a higher order P.D.E. so if we want to use it practically to solve a B.V.P. we must add to the boundary conditions of type 1, 2 or 3 above, the equation Eq. (107) as a new boundary condition.

Remark 4.2. A classical tool to solve P.D.E. of mathematical physics is the use of potentials.

In elasticity theory there exists a large variety of such potentials so we shall restrict here to mention only potentials for the displacement field $\mathbf{u}$.

Namely we can use the classical Helmolz representation for any field as

$$\mathbf{u} = \nabla\phi + \nabla \wedge \mathbf{A} \tag{120}$$

where ϕ is the scalar potential and $\mathbf{A}$ a vector potential, submitted to the gauge condition

$$\nabla \cdot \mathbf{A} = 0\,. \tag{121}$$

Let us note that according to a well known theorem any vector field enjoys a representation like Eq. (120), so also for $\mathbf{F}$ we will have

$$\begin{gathered} \mathbf{F} = \nabla\psi + \nabla \wedge \mathbf{B} \\ (\nabla \cdot \mathbf{B} = 0)\,. \end{gathered} \tag{122}$$

Substituting Eq. (120) and Eq. (122) in Eq. (108) and separating the gradient and rotational part we get

$$\begin{cases} \nabla\{(2\mu + \lambda)\Delta\phi + \psi\} = 0 \\ \nabla \wedge \{-\mu\Delta\mathbf{A} + \mathbf{B}\} = 0\,. \end{cases} \tag{123}$$

For an indefinite solid this means

$$\begin{cases} (2\mu + \lambda)\Delta\phi + \psi = 0 \\ -\mu\Delta\mathbf{A} + \mathbf{B} = 0\,. \end{cases} \tag{124}$$

In particular if $\mathbf{F}$ is a harmonic field of gradient type, *i.e.* $\mathbf{F} = \nabla\psi$, $\Delta\psi = 0$, then Eq. (123) implies even for a finite body

$$(2\mu + \lambda)\Delta\phi + \psi = \text{const.}\,,$$

i.e.

$$\Delta^2\phi = 0\,. \tag{125}$$

Another classical tool often used to simplify the solution of P.D.E., is the so called "reduction of dimensionality", *i.e.* the attempt of exploiting symmetries and invariances in order to eliminate one or more independent variables and one or more unknown fields.

Of particular interest for the geophysical applications we shall see later, is the case of two-dimensional problems, where we imagine to have a Cartesian system with z pointing upward, x horizontal and the physical quantities generally independent from y. When convenient, like here we shall switch to unabridged notation. It is interesting to observe that there are two distinct types of plane symmetries, namely one in which we assume that the displacement field $\mathbf{u}$ is always in the

(x, z) plane, i.e. $u_y \equiv 0$ and also u_x, u_z are independent from y, so that $\frac{\partial u_x}{\partial y} \equiv 0$, $\frac{\partial u_z}{\partial y} \equiv 0$, this is known as the *plain strain* problem; the other one in which we rather assume that the stress vector through any $y =$ const. plane is always zero, namely $\sigma_{xy} = \sigma_{yy} = \sigma_{zy} \equiv 0$, which is known as the *plain stress* problem.

Let us first consider the case of plain strain; due to the defining conditions we have in this case immediately

$$e_{xy} = e_{yy} = e_{zy} \equiv 0\,,$$

so that the following "reduced" equations can be written

$$e_{xx} = \frac{\partial u_x}{\partial x}\,, \quad e_{zz} = \frac{\partial u_z}{\partial z}\,, \quad e_{xz} = \frac{1}{2}\left(\frac{\partial u_x}{\partial z} + \frac{\partial u_z}{\partial x}\right) \tag{126}$$

$$\begin{cases} \sigma_{xx} &= 2\mu e_{xx} + \lambda(e_{xx} + e_{zz}) \\ \sigma_{zz} &= 2\mu e_{zz} + \lambda(e_{xx} + e_{zz}) \\ \sigma_{xz} &= 2\mu e_{xz} \end{cases} \tag{127}$$

$$\begin{cases} \frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \sigma_{xz}}{\partial z} + F_x &= 0 \\ \frac{\partial \sigma_{xz}}{\partial x} + \frac{\partial \sigma_{zz}}{\partial z} + F_z &= 0\,. \end{cases} \tag{128}$$

It is remarkable that if we compute the y components of the stress tensor, not all of them are zero, but $\sigma_{xy} = \sigma_{zy} = 0$ and

$$\sigma_{yy} = \lambda(e_{xx} + e_{zz})\,. \tag{129}$$

This is easily understandable from a physical point of view thinking for instance of the case in which we compress the body in z direction; as a reaction we expect a dilation in both x and y directions and if we want to stop the y displacement we must add one further stress in y.

Conversely the case of plain stress will lead to configurations where all the fields related to x, z components do not depend on y and further $e_{xy} = e_{zy} = 0$, but

$$e_{yy} = -\frac{\nu}{E}(\sigma_{xx} + \sigma_{zz})\,. \tag{130}$$

Naturally equilibrium conditions Eq. (128) continue to hold.

Remark 4.3. It is interesting to underline that in many cases we are more interested in computing the stress tensor throughout the body rather than the displacement field; for instance we might be willing to verify whether σ reaches somewhere the plastic threshold.

When the boundary conditions are favourable, *e.g.* if we consider a B.V.P. of type 2, it is possible to devise a treatment based on the determination of σ only.

Let us see how it works in plane case. We note that equations Eq. (128) are indeed not enough as we have 3 field unknowns, namely $\sigma_{xx}, \sigma_{xz}, \sigma_{zz}$.

On the other hand we must remember that the tensor e has to satisfy St. Venant conditions to be an "acceptable" strain and since Hooke's law has to hold we derive analogous conditions for σ.

So we find that in any case the following equation has to hold

$$\partial_x^2 e_{zz} + \partial_z^2 e_{xx} = 2\partial_{xz} e_{xz} \,. \tag{131}$$

Now the inverse of Hooke's law has two distinct forms in case of

a) plain strain

$$\begin{cases} e_{xx} = \frac{1+\nu}{E}\left\{(1-\nu)\sigma_{xx} - \nu\sigma_{zz}\right\} \\ e_{zz} = \frac{1+\nu}{E}\left\{(1-\nu)\sigma_{zz} - \nu\sigma_{xx}\right\} \\ e_{xz} = \frac{1+\nu}{E}\sigma_{xz} \end{cases} \tag{132}$$

or of

b) plain stress

$$\begin{cases} e_{xx} = \frac{1}{E}\left\{\sigma_{xx} - \nu\sigma_{zz}\right\} \\ e_{zz} = \frac{1}{E}\left\{\sigma_{zz} - \nu\sigma_{xx}\right\} \\ e_{xz} = \frac{1+\nu}{E}\sigma_{xz} \,. \end{cases} \tag{133}$$

If for instance we take Eq. (133) and substitute in Eq. (131) we find

$$(\partial_x^2 - \nu\partial_z^2)\sigma_{zz} + (\partial_z^2 - \nu\partial_x^2)\sigma_{xx} = 2(1+\nu)\partial_{xz}\sigma_{xz} \,, \tag{134}$$

which is the equation to be added to Eq. (128) in order to have a closed system.

If we further assume that the body force field is conservative, *i.e.*

$$F_x = \frac{\partial V}{\partial x} \,, \quad F_z = \frac{\partial V}{\partial z}$$

we see that Eq. (128) can be identically satisfied by introducing a suitable potential (Airy stress function) Φ, such that

$$\begin{gathered} \sigma_{xx} + V = \partial_z^2\Phi \quad \sigma_{zz} + V = \partial_x^2\Phi \\ \sigma_{xz} + V = -\partial_{xz}\Phi \,. \end{gathered} \tag{135}$$

Substituting Eq. (135) into Eq. (134) and rearranging we find

$$\Delta^2\Phi = (1-\nu)\Delta V \,, \tag{136}$$

i.e. again a bi-harmonic equation in Φ when V is harmonic.

Repeating an analogous reasoning for the plain strain case, and introducing the Airy function as in Eq. (135) we derive the equation

$$\Delta^2\Phi = \frac{(1-2\nu)}{1-\nu}\Delta V\,. \tag{137}$$

Dynamic equations in the linear elasticity range

The equations of motion of an elastic body further submitted to a body force field $\mathbf{F}$, descend as a straightforward application of d'Alambert principle, maintaining that a mass m submitted to an acceleration $\mathbf{a}$ generates an inertia force $-m\mathbf{a}$ balancing all the other external forces acting on m: in other words if we have an equilibrium theory in which the mass element m is in equilibrium if

$$\mathbf{F}(m) = 0\,, \tag{138}$$

then motion equations are derived directly from Eq. (138) by introducing an inertia force and setting

$$\mathbf{F}(m) - m\mathbf{a} = 0\,. \tag{139}$$

For an elastic body, by isolating an infinitesimal volume dV, we have the equilibrium condition

$$\frac{\partial\sigma_{ik}}{\partial x_k}dV + F_i dV = 0 \tag{140}$$

expressing the balance of the external force $F_i\ dV$ with the internal stress field. When these two forces are not balanced a motion will rise and if dm is the mass contained in dV, we will have

$$-dma_i + \frac{\partial\sigma_{ik}}{\partial x_k}dV + F_i dV = 0$$

or, considered that $a_i = \ddot{u}_i$, and denoting with $\rho = \frac{dm}{dV}$ the mass density

$$\rho\ddot{u}_i = \frac{\partial\sigma_{ik}}{\partial x_k} + F_i\,. \tag{141}$$

If $\mathbf{u}$ is a small displacement, $e = e(\mathbf{u})$ the corresponding strain tensor and $\sigma = \sigma(e)$ the stress

$$\sigma = 2\mu e + \lambda(Tre)I\,, \tag{142}$$

then Eq. (141) takes the form

$$\rho\ddot{\mathbf{u}} = \mu\Delta\mathbf{u} + (\lambda+\mu)\nabla(\nabla\cdot\mathbf{u}) + \mathbf{F}\,, \tag{143}$$

which is known as Navier equation.

The system of equations Eq. (143) is hyperbolic and we expect it to possess solutions which display the typical behaviour of waves.

In fact since the Lamé operator L is definite negative, we expect it to have eigenfunctions $\mathbf{u}_n$ (normal modes) such that

$$L\mathbf{u}_n = -\omega_n^2\mathbf{u}_n\,;$$

furthermore $\mathbf{u}_n$ have to be orthogonal, in the sense that

$$(\mathbf{u}_n, \mathbf{u}_m) = \int_B \mathbf{u}_n \cdot \mathbf{u}_m dB = \delta_{nm} ,$$

due to self-ajointness of L.

Then expanding $\mathbf{u}$ in a normal modes series

$$\mathbf{u} = \sum c_n(t)\mathbf{u}_n$$

we find from Eq. (143)

$$\sum \ddot{c}_n \mathbf{u}_n = -\sum \omega_n^2 \mathbf{u}_n + \mathbf{F} ,$$

i.e.

$$\ddot{c}_n + \omega_n^2 c_n = (\mathbf{u}_n, \mathbf{F}) \tag{144}$$

showing that the time component $c_n(t)$ behaves like a harmonic oscillator, of free frequency ω_n, forced by $(\mathbf{u}_n, \mathbf{F})$.

Accordingly we see that a proper set of conditions defining uniquely a solution of Eq. (143) is that of "initial conditions", *e.g.*

$$\mathbf{u}\,|_{t=0} = \mathbf{u}_0 , \quad \dot{\mathbf{u}}\,|_{t=0} = \dot{u}_0 \tag{145}$$

with $\mathbf{u}_0$, $\dot{\mathbf{u}}_0$ given functions.

Also, assuming ρ to be constant, it is easy to verify that Eq. (143), in an unbounded space and in absence of the forcing term $\mathbf{F}$, can have as solutions plane waves, like

$$\mathbf{u} = \mathbf{A} e^{ik(\mathbf{e}\cdot\mathbf{x} - vt)} \tag{146}$$

where

$\mathbf{e}$ = unit vector in the propagation direction;
k = wave-number = $\frac{2\pi}{L}$;
L = wave-length;
$kv = \omega = \frac{2\pi}{T}$ = pulsation;
T = period;
v = propagation velocity.

Substituting Eq. (146) into Eq. (143) one finds the relation

$$\rho\omega^2 \mathbf{A} = k^2 \left[\mu \mathbf{A} + (\lambda + \mu)\mathbf{e}(\mathbf{e} \cdot \mathbf{A}\right] \tag{147}$$

from which the two cases $\mathbf{e} \parallel \mathbf{A}$ (longitudinal waves) and $\mathbf{e} \perp \mathbf{A}$ (shear waves) can be easily treated, deriving the two velocities v_L, v_S

$$\begin{cases} v_L = \frac{\omega}{k} = \sqrt{\frac{2\mu+\lambda}{\rho}}, & (\mathbf{e}(\mathbf{e}\cdot\mathbf{A}) = \mathbf{A}) \\ v_S = \frac{\omega}{k} = \sqrt{\frac{\mu}{\rho}}, & (\mathbf{e}\cdot\mathbf{A} = 0) . \end{cases} \tag{148}$$

Remark 4.4. Very roughly speaking we can say that for the upper layers of the earth the velocity v can vary between a few up to 10 km/sec. Therefore even for wavelengths of the order of 10^3 km we cannot have periods longer than 10^3 sec i.e. in the range of minutes.

These motions are therefore completely temporally separated from viscous phenomena: on the other hand as we shall see in a minute also the forces involved are of a completely different magnitude so that we can maintain that the two phenomena are practically independent.

Dynamic equations and viscoelasticity. In order to introduce a viscoelastic behaviour in the dynamic equation Eq. (141) one has basically only to take the viscoelastic "Hooke's law" Eq. (131) into account.

In order to stress the fact that Eq. (131) will certainly include a purely elastic component, corresponding to the simple choice

$$R_1^e = 2\mu\ , \quad R_2^e = \lambda\ ,$$

and that the really viscoelastic part will act only on the deviator e'_{ik}, to supply a deviatoric stress σ', we make the position

$$\begin{cases} R_1 & = 2\mu + 2\eta(t) \\ R_2 & = \lambda - \frac{2}{3}\eta(t) \end{cases} \tag{149}$$

so that the stress-strain relation now becomes

$$\begin{aligned} \sigma_{ik}(t) \quad &= 2\mu e_{ik}(t) + \lambda \textstyle\sum_e e_{\ell\ell}(t)\delta_{ik} + \\ &+ \int_0^t 2\eta(t-\tau)\left\{de_{ik}(\tau) - \tfrac{1}{3}\textstyle\sum de_{\ell\ell}(\tau)\delta_{ik}\right\}\ . \end{aligned} \tag{150}$$

Remark 4.5. We have already seen in § 3 an example of a so-called Maxwell body schematically represented by a spring S and a dashpot D in series.

Maybe it's time now to mention the so-called Voigt body, where S and D are put in parallel like in Fig. 13.

The corresponding stress strain relation is, in the simplified case of Fig. 13,

$$F = \mu u + \eta \dot{u}\ .$$

This equation can be integrated with the condition $u(0) = 0$, giving

$$u = \frac{F}{\mu}\left(1 - e^{-\frac{\mu}{\eta}t}\right)\ , \tag{151}$$

which shows that the final equilibrium is controlled by the spring, while the transitory is damped by the dashpot.

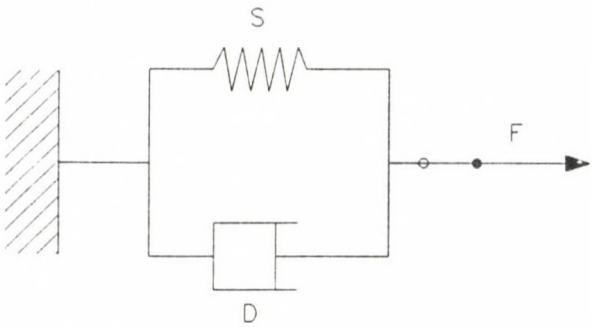

Fig. 13. Schematic of the Voigt body; S = spring, D = dashpot

The Voigt model in general would be

$$\begin{aligned} \sigma_{ik}(t) &= 2\mu e_{ik}(t) + \lambda \textstyle\sum_i e_{ll}(t)\delta_{ik} + \\ &+ 2\eta_0 \left\{ \dot{e}_{ik}(t) - \tfrac{1}{3} \textstyle\sum_l \dot{e}_{ll}(t)\delta_{ik} \right\} \end{aligned} \tag{152}$$

where η_0 has the meaning of a viscosity coefficient. The Laplace transform of Eq. (151) is simply

$$\overline{\sigma} = 2\mu\overline{e} + \lambda T_r(\overline{e})I + 2\eta s \left\{ \overline{e} - \frac{1}{3}(Tr\overline{e})I \right\} .$$

Let us now write explicitely the dynamic equation; we have, by using Eq. (150),

$$\begin{aligned} \rho\ddot{\mathbf{u}} &= \mu\Delta\mathbf{u} + (\lambda+\mu) \nabla (\nabla \cdot \mathbf{u}) + \\ &+ \int_0^t \eta(t-\tau) \left[\Delta\dot{\mathbf{u}} + \tfrac{1}{3} \nabla (\nabla \cdot \dot{\mathbf{u}}) \right] d\tau + \mathbf{F} . \end{aligned} \tag{153}$$

For the Voigt body for instance this becomes

$$\begin{aligned} \rho\ddot{\mathbf{u}} &= \mu\Delta\mathbf{u} + (\lambda+\mu) \nabla (\nabla \cdot \mathbf{u}) + \\ &+ \eta_0 \left[\Delta\dot{\mathbf{u}} + \tfrac{1}{3} \nabla (\nabla \cdot \dot{\mathbf{u}}) \right] + \mathbf{F} . \end{aligned} \tag{154}$$

The introduction of the viscous term in Eq. (153) produces an exponential damping of the solution in a way very similar to the elementary case Eq. (151). Just to verify it on a very simple ground one could try to find a plane wave solution of Eq. (154) when $\mathbf{F} = 0$; if we put

$$\mathbf{u} = \mathbf{A} e^{i(k\mathbf{e}\cdot\mathbf{x} - \omega t)} \tag{155}$$

and assuming for the sake of simplicity to be in the case of transverse waves, *i.e.* $\mathbf{e} \cdot \mathbf{A} = 0$, we obtain an equation relating ω to k as

$$\omega^2 = k^2(\mu - i\omega\eta_0) . \tag{156}$$

The idea of solving this equation with respect to ω^2 corresponds to searching a solution that at $t = 0$ has the shape

$$\mathbf{u}_0 = \mathbf{A}e^{ik\mathbf{e}\cdot\mathbf{x}}$$

Now since $\mu > 0$, $\omega\eta_o > 0$ we see that the root ω of Eq. (156) is such that

$$\omega = k(a - ib), \qquad a > 0,\ b > 0$$

so that

$$\mathbf{u}_t = \mathbf{A}e^{ik(\mathbf{e}\cdot\mathbf{x}-at)-kbt} \tag{157}$$

clearly displaying an exponential decay of the wave in time. Since this is the solution when $\mathbf{F}$ is absent, we may expect that for a time which is large with respect to the free periods of the body we have a solution where oscillations are died and only the final equilibrium position is reached. The same will happen if $\mathbf{F}$ is varying with a characteristic time much longer than the free periods.

Remark 4.6. To be more definite let us perform a very rough computation of the order of magnitude of the various forces entering into a realistic geodynamic phenomenon. For instance let us imagine that a disk of crust with radius $L = 100$ km starts moving upward, increasing its velocity from zero to 10 cm/yr.

This implies an acceleration roughly of

$$a \sim 10^{-14}\text{cm sec}^{-2}$$

and therefore an inertial force (per unit/volume)

$$|\,\rho a\,| \sim 3 \cdot 10^{-14}\text{g cm}^{-2}\ \text{sec}^{-2}\,.$$

On the other hand the body force (per unit volume) acting on the crust is of the order

$$|\,\rho g\,| \sim 3 \cdot 10^{3}\text{g cm}^{-2}\ \text{sec}^{-2}\,;$$

as we know most of this weight is supported by an elastic reaction (stress) in unperturbed conditions. It is so very important to evaluate the viscous term.

If we have a displacement of 10 cm over $L = 100$ km we have

$$|\,e\,| \sim 10^{-6}$$

so that

$$|\,\dot{e}\,| \sim \frac{|\,e\,|}{1yr} = 3 \cdot 10^{-14}\,.$$

By using a very prudential estimate of the coefficient η for the crust (cf. Artynshkov 1983)

$$\eta \sim 3 \cdot 10^{20}\text{g cm}^{-1}\ \text{sec}^{-2}$$

and noting that the viscous force is

$$|\,\eta(\mathrm{div}\dot{e})\,| \sim \eta\frac{|\,\dot{e}\,|}{L} \sim 1\mathrm{g\ cm}^{-2}\ \mathrm{sec}^{-2}\,.$$

We see that this term is in any way much bigger than the inertial forces, and such would remain even taking $L = 1000$ km.

The conclusion that can be drawn from these estimates is that even for a phenomenon much larger and much quicker, we have a difference of many orders of magnitude between inertial forces and viscous forces in geodynamics, at least for the upper layers; therefore we can ignore in Eq. (153) the term $\rho\ddot{\mathbf{u}}$ and put it to zero, so justifying the hypotheses which were below the statement of the equivalence principle reported at the end of § 3.

5. Elements of mathematical analysis of B.V.P.'s in elastostatics

In this paragraph we try to examine shortly the two basic questions of existence and uniqueness of the solution of the B.V.P.'s 1, 2, 3, presented at the beginning of § 4.

We underline that by the results of the discussion at the end of the last paragraph and in force of the equivalence principle we are treating at the same time the viscoelastic case.

We start with a few notes on the question of uniqueness and we observe that essentially Lemma 4.1, is already a proof of the uniqueness of the solution of the B.V.P. 1. We repeat the argument in such a way that it covers also the case of B.V.P. 3.

Let's assume that we have two solutions $\mathbf{u}_1, \mathbf{u}_2$ of the same B.V.P.

$$\begin{cases} \mu\Delta\mathbf{u} + (\lambda+\mu)\nabla(\nabla\cdot\mathbf{u}) = \mathbf{F} & \\ \mathbf{u} - \mathbf{u}_0 = 0 & S_u \\ \sigma(\mathbf{u})\boldsymbol{\nu} - \mathbf{T}_0 = 0 & S_T\,; \end{cases} \tag{158}$$

then the difference

$$\mathbf{v} = \mathbf{u}_2 - \mathbf{u}_1$$

satisfies the corresponding homogeneous problem

$$\begin{cases} \mu\Delta\mathbf{v} + (\lambda+\mu)\nabla(\nabla\cdot\mathbf{v}) = 0 & \\ \mathbf{v} = 0 & S_u \\ \sigma(\mathbf{v})\boldsymbol{\nu} = 0 & S_T\,. \end{cases} \tag{159}$$

But then

$$\begin{aligned} 0 &= \int_B \mathbf{v} \cdot \{\mu \Delta \mathbf{v} + (\lambda + \mu) \nabla (\nabla \cdot \mathbf{v})\} \, dB = \\ &= \int_B v_i \partial_k \sigma_{ik}(\mathbf{v}) dB = \\ &= \int_S v_i \sigma_{ik}(\mathbf{v}) \nu_k dS - \int_B (\partial_k v_i) \sigma_{ik}(\mathbf{v}) dB \,. \end{aligned} \tag{160}$$

On the other hand in Eq. (160) we have

$$\int_S v_i \sigma_{ik} \nu_k dS = \int_{S_u} v_i \sigma_{ik} \nu_k dS + \int_{S_T} v_i \sigma_{ik} \nu_k dS = 0 \tag{161}$$

because $v_i = 0$ on S_u and $\sigma_{ik}\nu_k = 0$ on S_T.

Moreover since σ_{ik} is symmetric one has, using the notation $V = [\partial_k v_i]$,

$$Tr\sigma V = Tr\sigma \frac{1}{2}(V + V^t) + Tr\sigma \frac{1}{2}(V - V^t) = Tr\sigma \frac{1}{2}(V + V^t) = Tr\sigma e(\underline{v}) \,.$$

Then

$$\begin{aligned} \int_B \partial_k v_i \sigma_{ik}(v_k) dB &= \int_B Tr\sigma e \; dB = \int_b Tr\,[2\mu e + \lambda (Tre) I]\, e \; dB = \\ &= 2\mu \int_B Tre^2 \; dB + \lambda \int_B (Tre)^2 dB \,. \end{aligned} \tag{162}$$

Therefore, from Eqs (160), (161), (162) we would have

$$2\mu \int Tre^2 dB + \lambda \int (Tre)^2 dB = 0$$

which, if the conditions

$$\lambda \geq 0 \,, \quad \mu \geq \mu_0 > 0$$

are fulfilled, implies

$$Tre^2 = 0 \quad \Rightarrow \quad e = 0 \,. \; (in \; B)$$

On the other hand, according to our discussion of § $e(\mathbf{v}) \equiv 0$ implies (cf. Eq. (26))

$$\mathbf{v} = \mathbf{a} + \boldsymbol{\omega} \wedge \mathbf{x} \,; \tag{163}$$

but Eq. (163) cannot hold on a piece of a regular surface S_u, of positive measure, without being $\mathbf{a} = 0$, $\boldsymbol{\omega} = 0$.

In fact let $\mathbf{x}_0$ be a point internal to S_u, so that we can identify two independent tangent vectors $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2$ ($\tau_1 \wedge \tau_2 \neq 0$); then we have

$$\mathbf{v} - \mathbf{v}_0 = \boldsymbol{\omega} \wedge (\mathbf{x} - \mathbf{x}_0) \equiv 0$$

which certainly holds around $\mathbf{x}_0$; this implies

$$\boldsymbol{\omega} \wedge (t_1 \boldsymbol{\tau}_1 + t_2 \boldsymbol{\tau}_2) \equiv 0 \quad \forall t_1, t_2$$

i.e.

$$\boldsymbol{\omega} \wedge \boldsymbol{\tau}_1 = 0\,, \quad \boldsymbol{\omega} \wedge \boldsymbol{\tau}_2 = 0\,,$$

which is possible only if $\boldsymbol{\omega} = 0$. But in this case we have just $\mathbf{v} = \mathbf{a}$ and since $\mathbf{v} = 0$ on S_u, also $\mathbf{a} = 0$, as it was to be proved.

Remark 5.1. If we consider now the B.V.P. 2, namely

$$\sigma(\mathbf{u})\boldsymbol{\nu} = T_0 \quad on\ S\,,$$

we see that the above reasoning (and in particular Eqs (160), (161), (162)) continues to hold so that the conclusion that $e = 0$ and

$$\mathbf{v} = \mathbf{a} + \boldsymbol{\omega} \wedge \mathbf{x} \tag{164}$$

is still valid. On the other hand in this case we see that this corresponds to a true non-uniqueness of the solution since

$$e(\mathbf{u} + \mathbf{v}) = e(\mathbf{u}) + e(\mathbf{v}) = e(\mathbf{u})$$

and therefore

$$\sigma(\mathbf{u} + \mathbf{v}) = \sigma(\mathbf{u})\,,$$

so that if $\mathbf{u}$ satisfies the Lamé equation and the pure traction B.V.P., which involve only σ, so does $\mathbf{u} + \mathbf{v}$ too, when $\mathbf{v}$ has the form Eq. (164).

A non-uniqueness of the solution calls the attention to the possible existence of conditions that data $\mathbf{F}$, $\mathbf{T}_0$ have necessarily to satisfy in order that a solution could exist (compatibility conditions); in fact we know that often B.V.P.'s like the one we are studying enjoy the so called Fredholm alternative, stating that there must be as many constraints on the data as many linearly independent solutions exist for the homogeneous problem.

In our case we can observe that if

$$\partial_k \sigma_{ik}(\mathbf{u}) + F_i \equiv 0$$

then

$$\begin{aligned} 0 &= \int_B (v_i \partial_k \sigma_{ik} + v_i F_i) dB = \\ &= \int_S (v_i \sigma_{ik} \nu_k) dS - \int_B \partial_k v_i \sigma_{ik} dB + \int_B v_i F_i dB\,. \end{aligned} \tag{165}$$

On the other hand, as we have already observed

$$\int_B \partial_k v_i \sigma_{ik} dB = \int_B Tre(\mathbf{v}) \sigma(\mathbf{u}) dB = 0$$

because by definition

$$e(\mathbf{v}) \equiv 0\,.$$

So, also recalling that $\sigma_{ik}\nu_k = T_i$, from Eq. (165) we receive

$$\begin{aligned} 0 &= \int_S \mathbf{v}\cdot\mathbf{T}dS + \int_B \mathbf{v}\cdot\mathbf{F}dB = \\ &= \mathbf{a}\cdot\left\{\int_S \mathbf{T}dS + \int_B \mathbf{F}dB\right\} - \boldsymbol{\omega}\cdot\left\{\int_S \mathbf{x}\wedge T\,dS + \int_B \mathbf{x}\wedge\mathbf{F}\,dB\right\}. \end{aligned} \tag{166}$$

Due to the arbitrariness of $\mathbf{a}$ and $\boldsymbol{\omega}$ we see that if a solution of the given B.V.P. exists, than $\mathbf{T}, \mathbf{F}$ have to satisfy the two conditions

$$\int_S \mathbf{T}dS + \int_B \mathbf{F}dB = 0$$

$$\int_S \mathbf{x}\wedge\mathbf{T}dS + \int_B \mathbf{x}\wedge\mathbf{F}dB = 0$$

expressing basically the annihilation of the resultant of external forces as well as of their total momentum.

Now we come to existence problems; as it is known such problems are generally more difficult to be dealt with and require a deeper knowledge of functional analysis. Here we choose the variational approach because it allows us to study the energy functional and its properties, what is interesting in itself also from the physical point of view.

Let us start by defining a functional representing the internal energy.

We imagine that $\mathbf{F}$ is given in B, $\mathbf{T}_0$ is given on S_T and $\mathbf{u}_0$ is given on S_u and we also assume that in such a situation B has a configuration specified by the displacement field $\mathbf{u}$; now let us think of giving to $\mathbf{u}$ a variation $\mathbf{v}$ $(\mathbf{u}+\mathbf{v})$ in such a way that the assigned displacement on S_u doesn't change; this requires

$$\mathbf{v}\,|_{S_u} = 0\,. \tag{167}$$

Let us calculate the "virtual work" δL done by the external forces $(\mathbf{F}, \mathbf{T})$ during this infinitesimal displacement indeed we have, recalling Eq. (167) and repeating the same arguments of Eqs (161), (162)

$$\begin{aligned} \delta L &= \int_B \mathbf{F}\cdot\mathbf{v}\;dB + \int_{S_T} \mathbf{T}_0\cdot\mathbf{v}\;dS = \\ &= -\int_B \partial_k\sigma_{ik}v_i\;dB + \int_{S_T} T_{0i}v_i\;dS = \\ &= -\int_S \sigma_{ik}\nu_k v_i\;dS + \int_{S_T} T_{0i}v_i\;dS + \int_B \sigma_{ik}\partial_k v_i\;dB = \\ &= \int_B Tr\sigma(\mathbf{u})e(\mathbf{v})\;dB\,. \end{aligned} \tag{168}$$

If this virtual work is physical, *i.e.* if the displacement $\mathbf{u}+\mathbf{v}$ is physically acceptable (*i.e.* if it is obtained by modifying the external forces according to $\delta\mathbf{F} = -\nabla\cdot\sigma(\mathbf{v}), \delta\mathbf{T} = \sigma(\mathbf{v})\boldsymbol{\nu}$) then we expect that this work transfers (if it is positive) energy to the body which accumulates it in the form of "internal" (or elastic) energy.

If we imagine to go from $t\mathbf{u}$ to $(t+dt)\mathbf{u}$, [5] then, due to linearity of σ and e in $\mathbf{u}$,

$$\delta L = t\ dt \int_B Tr\sigma(\mathbf{u})e(\mathbf{u})\ dB$$

so that letting t to range from 0 to 1 we find a total work

$$L = \frac{1}{2}\int_B Tr\sigma(\mathbf{u})e(\mathbf{u})\ dB\,; \tag{169}$$

this can be assumed as definition of internal energy.

We now observe also that the work Eq. (168) was done by the external forces so that the body in the modified configuration has a different potential energy which is exactly equal in absolute value and opposite in sign to δL.

So we can define the total energy of the body as

$$E(\mathbf{u}) = \frac{1}{2}\int_B Tr\sigma(\mathbf{u})e(\mathbf{u})\ dB - \int_B \mathbf{F}\cdot\mathbf{u}\ dB - \int_{S_T} \mathbf{T}_0\cdot\mathbf{u}\ dS\,. \tag{170}$$

We expect the body to be in equilibrium when any virtual variation $\mathbf{u}\to\mathbf{u}+\mathbf{v}$ would increase $E(\mathbf{u})$, i.e. would require one further external agent supplying energy to our system.

Indeed let us prove that if $\mathbf{u}$ is the physical displacement corresponding to $\mathbf{F}, \mathbf{T}_0$ and satisfying the boundary condition on S_u which without loss of generality can be put in the form

$$\mathbf{u}\,|_{S_u} = 0\,, \tag{171}$$

then E attains a minimum at $\mathbf{u}$.

In fact let $\mathbf{v}$ be any other vector also satisfying Eq. (171), then

$$\begin{aligned} E(\mathbf{u}+\mathbf{v}) &= \int_B \tfrac{1}{2}\{Tr\sigma(\mathbf{u})e(\mathbf{u}) + 2Tr\sigma(\mathbf{u})e(\mathbf{v}) + Tr\sigma(\mathbf{v})e(\mathbf{v})\}\,dB + \\ &- \int_B \mathbf{F}\cdot\mathbf{u}dB - \int_{S_T}\mathbf{T}_0\cdot\mathbf{u}dS - \int_B \mathbf{F}\cdot\mathbf{v}dB - \int_{S_T}\mathbf{T}_0\cdot\mathbf{v}dS = \\ &= E(\mathbf{u}) + \left\{\int_B Tr\sigma(\mathbf{u})e(\mathbf{v})dB - \int_B \mathbf{F}\cdot\mathbf{v}dB - \int \mathbf{T}_0\cdot\mathbf{v}dS\right\} + \\ &+ \tfrac{1}{2}\int Tr\sigma(\mathbf{v})e(\mathbf{v})\,. \end{aligned} \tag{172}$$

[5] Let's assume for the sake of simplicity that $\mathbf{u}\,|_{S_u} = 0$, so that $\mathbf{v}\,|_{S_u} = dt\mathbf{u}\,|_{S_u} = 0$.

On the other hand the midterm in the right hand side of Eq. (172) is identically zero, whatever is $\mathbf{v}$, because of Eq. (168), so

$$E(\mathbf{u}+\mathbf{v}) = E(\mathbf{u}) + \frac{1}{2}\int Tr\sigma(\mathbf{v})e(\mathbf{v})dB\,.$$

Since

$$\int_B Tr\sigma(\mathbf{v})e(\mathbf{v})dB \geq 0$$

we conclude

$$E(\mathbf{u}+\mathbf{v}) \geq E(\mathbf{u})\,,$$

i.e. E attains a minimum a $\mathbf{u}$.

We have thus proved the following Theorem 5.1.

Theorem 5.1. if for given $\mathbf{F}, \mathbf{T}_0$ there is a physical displacement corresponding to the equilibrium of the body and such that $E(\mathbf{u}) \leq +\infty$, then E attains at $\mathbf{u}$ a minimum among all other fields for which the energy is finite and satisfying Eq. (171).

It is important to understand that we have not proved that a minimum exists but rather that if a solution to the equilibrium problem exists then it is an extremal field for $E(\mathbf{u})$; we can generalize the concept of "solution" of our equilibrium problem by saying that if a minimum of E exists then this is the sought solution.

Remark 5.1. This "variational" definition coincides with the so called weak solution because if Eq. (167) has to hold $\forall\mathbf{v}$, then also Eq. (168) has to hold $\forall\mathbf{v}$ which is precisely the weak formulation.

The proof of the existence of the minimum relies on number of results which we report here after in form of Lemmas some of which without proof.

Lemma 5.1: the space $\overline{H}$ of (vector) functions $\mathbf{u}$ which belong to $L^2(B)$ together with their first derivatives and such that

$$\mathbf{u}\,|_{S_u} = 0\,, \tag{173}$$

is a closed subspace of $H^{1,2}$, the space endowed with the norm

$$\|\mathbf{u}\|_{1,2}^2 = \int_B (TrUU^+ + |\,\mathbf{u}\,|^2)dB$$
$$\left(U = \left[\frac{\partial u_i}{\partial x_k}\right]\right)\,. \tag{174}$$

In fact restricting the attention to a single component u of $\mathbf{u}$, let us define in a generalized sense the boundary integral

$$\int \nu\Phi\; u\; dS \overset{Def}{=} \int (\Phi \bigtriangledown u + u \bigtriangledown \Phi)dB \tag{175}$$

for any $u \in \overline{H}$ and any Φ sufficiently smooth.

The condition Eq. (173) can then be expressed with the help of Eq. (175) as

$$\int (\Phi \bigtriangledown u + u \bigtriangledown \Phi) dB = 0 \quad \forall \Phi : \Phi \mid_{S_T} = 0 \,. \tag{176}$$

Now if $u_n \to \tilde{u}$ in $H^{1,2}$ ($u_n \in \overline{H}$) we have indeed, $\forall \Phi : \Phi \mid_{S_T} = 0$,

$$\int (\Phi \bigtriangledown \tilde{u} + \tilde{u} \bigtriangledown \Phi) dB = \lim \int (\Phi \bigtriangledown u_n + u_n \bigtriangledown \Phi) dB = 0$$

which means that not only $\tilde{u} \in H^{1,2}$ but also it satisfies Eq. (176), *i.e.* Eq. (173) so that $\tilde{u} \in \overline{H}$.

Lemma 5.2: (Korn's theorem) the energy norm

$$||\mathbf{u}||^2 = \int Tr\sigma(\mathbf{u}) e(\mathbf{u}) dB \tag{177}$$

is equivalent in $\overline{H}$ to the $H^{1,2}$ norm Eq. (174), *i.e.* there is a constant A such that

$$A^{-1} ||\mathbf{u}||_{1,2}^2 \leq ||\mathbf{u}||^2 \leq A ||\mathbf{u}||_{1,2}^2 \,. \tag{178}$$

We note that Eq. (177) is a true norm, *i.e.* $||\mathbf{u}|| = 0$ implies $\mathbf{u} = 0$ (recall condition Eq. (173)) as we have already observed discussing uniqueness, yet the equivalence Eq. (178) needs to be proved directly.

The right inequality in Eq. (178) follows from observing that

$$Tr\sigma e = 2\mu Tre^2 + \lambda (Tre)^2 \leq (2\mu + 3\lambda) Tre^2$$

and

$$\begin{aligned} \int Tre^2 dB &= \tfrac{1}{2} \int TrUU^+ dB + \tfrac{1}{2} \int TrU^2 dB \,, \\ \int TrU^2 dB &= \int \textstyle\sum_{i,k} \frac{\partial u_i}{\partial x_k} \frac{\partial u_k}{\partial x_i} dB \leq \\ &\leq \int \left(\textstyle\sum_{i,k} \frac{\partial u_i^2}{\partial x_k} \right)^{\frac{1}{2}} \left(\textstyle\sum_{i,k} \frac{\partial u_k^2}{\partial x_i} \right)^{\frac{1}{2}} dB = \\ &= \int Tr(UU^+) dB \,. \end{aligned} \tag{179}$$

So, summarizing we have (also recalling Eq. (174))

$$||\mathbf{u}||^2 \leq (2\mu + 3\lambda) ||\mathbf{u}||_{1,2}^2 \quad . \tag{180}$$

The left inequality in Eq. (178) however is much more difficult to be established; so we prove Eq. (178) completely, only in the restrictive case that

$\mathbf{u}\,|_S = 0$, *i.e.* $S_u \equiv S$. In this case in fact we have

$$\begin{aligned}\int_B TrU^2 dB &= \int_B \sum_{i,k} \frac{\partial u_i}{\partial x_k}\frac{\partial u_k}{\partial x_i} dB = \\ &= -\int_B \sum_{i,k} \frac{\partial^2 u_i}{\partial x_i \partial x_k} u_k dB = \int_B \sum_{i,k} \left(\frac{\partial u_i}{\partial x_i}\right)\left(\frac{\partial u_k}{\partial x_k}\right) dB = \\ &= \int_B (TrU)^2 dB \geq 0\,.\end{aligned}$$

So we get

$$\begin{aligned}\|\mathbf{u}\|^2 &= 2\mu \int_B Tre^2 dB + \lambda \int_B (Tre)^2 dB \geq \\ &\geq \mu \int_B TrUU^+ dB + \mu \int_B TrU^2 dB \geq \\ &\geq \mu \int_B TrUU^+ dB = \mu \int \sum_{i,k} \left(\frac{\partial u_i}{\partial x_k}\right)^2 dB\,.\end{aligned} \tag{181}$$

The last step to be done is to prove that for a function $\mathbf{u} \in \overline{H}$, in particular satisfying Eq. (173), we have for a suitable constant C

$$\int_B |\,\mathbf{u}\,|^2\, dB \leq C \int_B \sum_{i,k} \left(\frac{\partial u_i}{\partial x_k}\right)^2 dB\,. \tag{182}$$

Here again we cannot give a general proof of Eq. (182) but to try to give at least a hint of the reason why it holds. We consider a simplified example in which B is a cube of side 1 (cf. Fig. 14) and the set S_u, on which $\mathbf{u} = 0$, is the face lying on the x, y plane. In this case

$$u(x,y,z) = \int_0^z u_z(x,,y,\zeta) d\zeta \tag{183}$$

so that, by using the Schwarz inequality,

$$\begin{aligned}\int u^2 dxdydz &\leq \int dxdydz \left|\int_0^1 u_z(x,y,\zeta) d\zeta\right|^2 \leq \\ &\leq \int dxdyd\zeta u_z(x,y,\zeta)^2\,;\end{aligned} \tag{184}$$

this inequality implies Eq. (182).

So summarizing Eq. (181) and Eq. (182) we see that, *e.g.*

$$\|\mathbf{u}\|^2 \geq \frac{\mu}{2} \int_B TrUU^+ dB + \frac{\mu}{2C} \int_B |\,\mathbf{u}\,|^2\, dB\,,$$

i.e. Eq. (178) is proved.

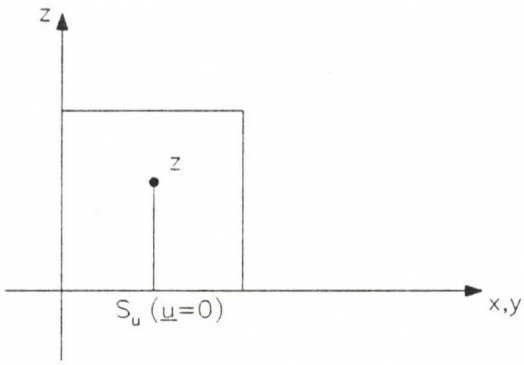

Fig. 14. Proof of norm equivalence in simplified geometry

Lemma 5.3: the trace of a function $\mathbf{u} \in \overline{H}$ on S_T is a square integrable function on S_T.

Again we take the simplified situation of Fig. 14 and, from Eq. (183), we see that

$$\begin{aligned} \int u^2(x,y,1)dxdy &= \int dxdy \left(\int_0^1 u_z(x,y,\zeta)d\zeta \right)^2 \leq \\ &\leq \int dxdyd\zeta u_z(x,y,\zeta)^2 \leq A\|\mathbf{u}\|^2 . \end{aligned} \tag{185}$$

So u is square integrable over the face $z = 1$; on the other hand it is obvious that also the trace of u on any regular surface which has a one to one projection onto $0 \leq x \leq 1, 0 \leq y \leq 1$, is again square integrable on it. With one further integration in x or y we find that $\mathbf{u}$ has a square integrable trace over all the faces.

Lemma 5.4: if we denote the scalar product in $\overline{H}$

$$\langle \mathbf{u}, \mathbf{v} \rangle = \int_B Tr\sigma(\mathbf{u})e(\mathbf{v})dB , \tag{186}$$

such that

$$\|\mathbf{u}\|^2 = \langle \mathbf{u}, \mathbf{u} \rangle , \tag{187}$$

we have for any $\mathbf{F} \in L^2(B)$, $\mathbf{T}_0 \in L^2(S_T)$

$$\begin{cases} \Phi(\mathbf{u}) = \int_B \mathbf{F} \cdot \mathbf{u} \, dB = \langle K\mathbf{F}, \mathbf{u} \rangle \\ \Psi(\mathbf{u}) = \int_{S_T} T_0 \cdot \mathbf{u} \, dS = \langle H\mathbf{T}_0, \mathbf{u} \rangle . \end{cases} \tag{188}$$

This is a straightforward application of the Riesz representation theorem stating that if $\Phi(\mathbf{u})$ is a bounded linear functional in $\overline{H}$ then there must be a

representative $\mathbf{f} \in \overline{H}$ such that

$$\Phi(\mathbf{u}) = \langle \mathbf{f}, \mathbf{u} \rangle \,. \tag{189}$$

In our case since, (recall Eqs (181), (182))

$$| \Phi(\mathbf{u}) |^2 \leq \int_B | \mathbf{F} |^2 \, dB \cdot \int_B | \mathbf{u} |^2 \, dB \leq C \cdot \int_B | F |^2 \, dB \cdot ||\mathbf{u}||^2$$

then the first of Eq. (188) has to hold.

Moreover since (recall Lemma 5.3)

$$| \Psi(\mathbf{u}) |^2 \leq \int_{S_T} | T_0 |^2 \, dS \cdot \int_{S_T} | \mathbf{u} |^2 \, dS \leq C \cdot \int_{S_T} | \mathbf{T}_0 |^2 \, dS \cdot ||\mathbf{u}||^2 \,,$$

then also the second of Eq. (188) has to hold.

The operators K, H are indeed continuous operators respectively from $L^2(B)$, $L^2(S_T)$ into $\overline{H}$.

By exploiting all these Lemmas we can easily prove the following main theorem.

Theorem 5.2. the variational problem

$$E(\mathbf{u}) = \min \qquad \mathbf{u} \in \overline{H}$$

has always a solution in $\overline{H}$ on condition that S_u has positive measure, so that Lemma 5.2 holds.

In fact going back to Eq. (170) we see that

$$E(\mathbf{u}) = \frac{1}{2}||\mathbf{u}||^2 - (\mathbf{F}, \mathbf{u})_{L^2(B)} - (\mathbf{T}_0, \mathbf{u})_{L^2(S_T)} \,,$$

and, considering Eq. (188), we can write

$$E(\mathbf{u}) = \frac{1}{2}||\mathbf{u}||^2 - \langle (K\mathbf{F} + H\mathbf{T}), \mathbf{u} \rangle \,. \tag{190}$$

Also, calling

$$\mathbf{f} = K\mathbf{F} + H\mathbf{T} \in \overline{H} \,, \tag{191}$$

we can rewrite Eq. (190) as

$$E(\mathbf{u}) = \frac{1}{2}||\mathbf{u} - \mathbf{f}||^2 - \frac{1}{2}||\mathbf{f}||^2 \,. \tag{192}$$

It is obvious that the minimization of Eq. (192) has a solution and precisely

$$\mathbf{u} = \mathbf{f} \,, \tag{193}$$

so that our theorem is proved.

We can note that, as the Ritz representative of a functional Φ is unique, we have also proved uniqueness of the solution.

This holds for the B.V.P.'s of types 1 and 3.

Remark 5.2. As for the case of B.V.P. 2 (pure tractions) the situation is a little bit more complicated, because of lack of uniqueness.

However one could prove that when conditions Eq. (166), *i.e.*

$$\int_B \mathbf{F}\, dB + \int_{S_T} \mathbf{T}_0\, dS = 0$$

$$\int_B \mathbf{x} \wedge \mathbf{F}\, dB + \int_{S_T} \mathbf{x} \wedge \mathbf{T}_0\, dS = 0$$

are satisfied then there is always one and only one solution $\mathbf{u}$ orthogonal to the functions of type $\mathbf{a} + \boldsymbol{\omega} \wedge \mathbf{x}$, i.e. such that

$$\int \mathbf{u}\, dB = 0$$

$$\int \mathbf{x} \wedge \mathbf{u}\, dB = 0\,;$$

this solution $\mathbf{u}$ is in $H^{1,2}(B)$.

6. Examples and exercises

We present in this paragraph a few examples and propose a few exercises to let the reader to come closer to geophysical applications of the general theory described in the preceding paragraphs.

Most of the examples will be referred to a flat earth model in which the upper boundary is the x, y plane and z points upward, while the body B is in the half space $z \leq 0$. The possible contribution of an irregular topographic layer can then be described as a pressure $p = -\rho_c g h$ with $\rho_c =$ density of the crust, $g =$ gravity, $h =$ topographic height. Naturally the earth is more spherical than flat, but switching from flat to spherical earth is more an exercise of differential calculus; moreover for tectonic phenomena which involve an area say of 1000 km diameter on the surface we can still draw meaningful hints from on a flat earth model.

To be definite we will work in the so called plain strain hypothesis, *i.e.* assuming $u_y \equiv 0$ and u_x, u_z independent from y.

Example 6.1: as a first example we consider the case of a homogeneous (λ, μ const.) halfspace ($z \leq 0$) on which we impose a displacement in the z direction only

and function of x only; namely

$$\begin{cases} u_x(x,y,0) & \equiv 0 \\ u_y(x,y,0) & \equiv 0 \\ u_z(x,y,0) & = f(x)\,. \end{cases} \tag{194}$$

Let us also agree that $u_y \equiv 0$, i.e. we are in the case of plain strain (in the plain x, z).

Moreover we assume that there are no body forces, i.e. $\mathbf{F} \equiv 0$, so that the equilibrium conditions write

$$\begin{cases} \dfrac{\partial \sigma_{xx}}{\partial x} + \dfrac{\partial \sigma_{xz}}{\partial z} & = 0 \\ \dfrac{\partial \sigma_{xz}}{\partial x} + \dfrac{\partial \sigma_{zz}}{\partial x} & = 0 \end{cases} \tag{195}$$

and the relevant Hooke's relations become

$$\begin{cases} \sigma_{xx} = (2\mu+\lambda)e_{xx} + \lambda e_{zz} = (2\mu+\lambda)\partial_x u_x + \lambda \partial_z u_z \\ \sigma_{zz} = (2\mu+\lambda)e_{zz} + \lambda e_{xx} = (2\mu+\lambda)\partial_z u_z + \lambda \partial_x u_x \\ \sigma_{xz} = 2\mu e_{xz} = \mu(\partial_z u_x + \partial_x u_z)\,. \end{cases} \tag{196}$$

To simplify matters we perform a Fourier transform of all relations in the variable x, getting

$$u_x(x,z) = \int_{-\infty}^{+\infty} e^{-ipx}\hat{u}_x(p,z)dp \tag{197}$$

and so on.

Our problem now becomes[6]

$$\begin{cases} -ip\hat{\sigma}_{xx} + \partial_z \hat{\sigma}_{xz} = 0 \\ -ip\hat{\sigma}_{xz} + \partial_z \hat{\sigma}_{zz} = 0 \end{cases} \tag{198}$$

with

$$\begin{cases} \hat{\sigma}_{xx} = -ip(2\mu+\lambda)\hat{u}_x + \lambda\partial_z \hat{u}_z \\ \hat{\sigma}_{zz} = (2\mu+\lambda)\partial_z \hat{u}_z - ip\lambda\hat{u}_x \\ \hat{\sigma}_{xz} = \mu\partial_z \hat{u}_z - ip\mu\hat{u}_z\,. \end{cases} \tag{199}$$

[6] Let us observe for later use that if body forces were present then we would just have Eq. (198) with known terms $-\hat{F}_x, -\hat{F}_z$.

If we write $\lambda + \mu = \zeta$ to shorten the notation we get the system of differential equations

$$\begin{cases} \mu\partial_z^2\hat{u}_x - p^2(\mu+\zeta)\hat{u}_x - ip\zeta\partial_z\hat{u}_z = 0 \\ -ip\zeta\partial_z\hat{u}_x + (\mu+\zeta)\partial_z^2\hat{u}_z - p^2\mu\hat{u}_z = 0\,. \end{cases} \tag{200}$$

This can be solved with the initial conditions

$$\begin{cases} \hat{u}_x(p,0) = 0 \\ \hat{u}_z(p,0) = \hat{f}(p)\,. \end{cases} \tag{201}$$

As Eq. (200) is a system of the 4th order the conditions Eq. (201) cannot be enough; in fact we must complement them with *boundedness* of $\hat{u}_x, \hat{u}_z$ when $z \to -\infty$.

The simplest way of solving Eq. (200) is to reduce it to a single equation, *e.g.* in u_z.

Differentiating once the second of Eq. (200) and substituting $\partial_z^2 u_x$ in the first we get

$$-ip^3\zeta(\mu+\zeta)\hat{u}_x + p^2(\zeta^2-\mu^2)\partial_z\hat{u}_z + \mu(\mu+\zeta)\partial_z^3\hat{u}_z = 0\,. \tag{202}$$

Differentiating again this equation and substituting $\partial_z\hat{u}_x$ from the second of Eq. (200) we receive the very simple equation

$$\partial_z^4\hat{u}_z - 2p^2\partial_z^2\hat{u}_z + p^4\hat{u}_z = 0\,. \tag{203}$$

The general solution of this equation is

$$\hat{u}_z = (A+Bz)e^{+|p|z} + (C+Dz)e^{-|p|z}$$

where

$$C = D = 0$$

if the boundedness condition has to be satisfied.

With the second of Eq. (201) we see that

$$\hat{u}_z = (\hat{f} + Bz)e^{+|p|z}\,. \tag{204}$$

Finally Eq. (204) has to be used in Eq. (202) to get, with the help of the boundary condition $\hat{u}_x(x,0) = 0$, the final result

$$\hat{u}_z = \hat{f}\left(1 - |\,p\,|\,z\frac{\zeta}{\zeta+2\mu}\right)e^{+|p|z}\,. \tag{205}$$

The expression Eq. (205) can be back-Fourier transformed to supply the solution in the x domain

$$u_z(x,z) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} K(z, x-\xi)f(\xi)d\xi \tag{206}$$

with the kernel

$$K(z,x) = \frac{2z}{z^2+x^2} + 2\frac{\lambda+\mu}{\lambda+3\mu} z \frac{z^2-x^2}{(z^2+x^2)^2} \,. \tag{207}$$

The component $\hat{u}_x$, and then u_x, can be similarly derived making use of the equation Eq. (202).

Exercise 6.1 The Boussinesq problem is to determine the deformation of a semi-infinite space ($z \leq 0$) when a load P is applied to the upper surface, concentrated at the origin

$$\sigma_{xz} = 0\,, \quad \sigma_{yz} = 0\,, \quad \sigma_{zz} = -P\delta(x)\delta(y) \tag{208}$$

where $\delta(x)\delta(y)$ are Dirac's deltas.

The body is supposed to be free of body forces.

The explicit solution of this problem is known *e.g.* in terms of displacements:

$$\begin{cases} u_x &= -\frac{P_0}{4\pi\mu}\left[\frac{xz}{r^3} - \frac{(1-2\nu)x}{r(r+z)}\right] \\ u_y &= -\frac{P_0}{4\pi\mu}\left[\frac{yz}{r^3} - \frac{(1-2\nu)y}{r(r+z)}\right] \\ u_z &= -\frac{P_0}{4\pi\mu}\left[\frac{2(1-\nu)}{r} + \frac{z^2}{r^3}\right] \,. \end{cases} \tag{209}$$

Use this "elementary" solution to find u_z along the z axis ($x=0,\ y=0,\ z<0$) when the load, $-P$, is uniformly spread over a disk of radius R on the x, y plane and show that

$$u_z(0,0,z) = -\frac{P}{2\mu} z \left\{ 2(1-\nu)\left[\sqrt{1+\left(\frac{R}{z}\right)^2} - 1\right] - \left[\frac{1}{\sqrt{1+\left(\frac{R}{z}\right)^2}} - 1\right]\right\} \,. \tag{210}$$

Let us observe that, to compare Eq. (210) with Eq. (209), one has to take into account that $P_0 = P \cdot \pi R^2$.

As we can see at the origin ($z \to 0$) the solution Eq. (210) attains a finite value, namely

$$u_z = -\frac{P(1-\nu)R}{\mu}$$

while the solution Eq. (209) diverges.

At a depth $z \sim R$, Eq. (210) is about one fourth of Eq. (209), while for large negative z, both Eq. (209) and Eq. (210) give the same asymptotic behaviour

$$u_z(0,0,z) \sim -\frac{P}{2\mu}(1,5-\nu)\frac{R^2}{z} \,.$$

This is in a way an illustration of *St. Venant principle* stating that if two system of forces are distributed on the same area S with the same resultant and momentum, than they give rise to the same displacements far away from S.

Example 6.2 (viscoelastic Boussinesq problem). We consider a semi-infinite body with a viscoelastic behaviour of the Voigt type, so that according to Remark 4.5

$$\begin{cases} \overline{\mu}(s) = \mu + \eta s \\ \overline{\lambda}(s) = \lambda - \frac{2}{3}\eta s \,. \end{cases} \tag{211}$$

We assume now that a load $-P$, concentrated of the origin, is suddenly applied at $t = 0$ and then taken constant.

In other words we apply the boundary conditions

$$\begin{cases} \sigma_{xz} = \sigma_{yz} = 0 \\ \sigma_{zz} = -P_0\theta(t)\delta(x)\delta(y) = -P(t)\delta(x)\delta(y) \\ \theta(t) = \begin{cases} 0 & t \leq 0 \\ 1 & t > 0 \end{cases} , \end{cases} \tag{212}$$

and we want to derive the time evolution of the displacement vector, in particular of u_z.

To this aim we can use the already known elastic solution Eq. (209) and the equivalence principle (cf. § 3 at the end) stating that for instance

$$\overline{u}_z = -\frac{\overline{P}}{4\pi\overline{\mu}}\left[\frac{2(1-\overline{\nu})}{r} + \frac{z^2}{r^3}\right] \tag{213}$$

where an overbar denotes a timelike Laplace transform.

We note that

$$\overline{P} = \frac{P_0}{s}$$

$$\frac{1-\overline{\nu}}{\overline{\mu}} = \frac{2}{\eta}\cdot\frac{\alpha+s}{(\beta+s)(\delta+s)}$$

$$\frac{1}{2\overline{\eta}} = \frac{1}{2\eta}\cdot\frac{1}{\delta+s}$$

$$\alpha = \frac{3(2\mu+\lambda)}{4\eta}$$

$$\beta = \frac{3(\mu+\lambda)}{\eta}$$

$$\delta = \frac{\mu}{\eta}\,.$$

With the usual decomposition into simple fractions one has

$$\frac{1-\overline{\nu}}{s\overline{\mu}} = \frac{2(\alpha+s)}{s(\beta+s)(\delta+s)} = \frac{2\alpha}{\eta}\left[\frac{1}{\beta\delta}\cdot\frac{1}{s} + \frac{1}{\beta(\beta-\delta)(\beta+s)} - \frac{1}{\delta(\beta-\delta)(\delta+s)}\right]$$

$$\frac{1}{s\overline{\mu}} = \frac{1}{\eta\delta}\left[\frac{1}{s} - \frac{1}{\delta+s}\right] . \tag{214}$$

Now it is enough to remember that

$$\mathcal{L}(1) = \frac{1}{s}$$

$$\mathcal{L}(e^{-at}) = \frac{1}{s+a} ,$$

so that using Eq. (214) in Eq. (213) we can find immediately the inverse Laplace transform

$$u_z(x,y,z,t) = -\frac{P_0}{2\pi}\left\{\frac{2\alpha}{\eta r}\left[\frac{1}{\beta\delta} + \frac{e^{-\beta t}}{\beta(\beta-\delta)} - \frac{e^{-\delta t}}{\delta(\beta-\delta)}\right] + \frac{z^2}{2\mu r^3}\left[1-e^{-\delta t}\right]\right\} .$$

It is interesting to observe that when $t \to \infty$, u_z tends to the pure elastic equilibrium solution (cf. Eq. (209)); this is typical of a Voigt body for which after a long time the dashpot effect faints and the equilibrium is purely supported by the spring (cf. Fig. 13).

Furthermore we can observe that the two characteristic delay times of our solution, *i.e.*

$$T_1 = \frac{1}{\beta}, \quad T_2 = \frac{1}{\delta},$$

are both proportional to η so that the higher is the viscosity, the longer the time required to reach the asymptotic equilibrium.

Exercise 6.2 Assume a layer of constant density ρ_c is given in (cf. Fig. 15)

$$-M \le z \le 0 ; \tag{215}$$

this layer is supported by a semi-infinite space, $z \le -M$, occupied by an incompressible fluid.

The upper surface of the layer ($z = 0$) is free of tractions, while the body is submitted to a body force, acting in the negative z direction,

$$F_z = -\rho_c g(z) .$$

Find the displacement field in the case that

1. the body is elastic and (constant gravity)

$$g(z) = g_0$$

2. the body is elastic and

$$g(z) = g_0 - Gz$$

3. the body is viscoelastic and

$$g(z) = g_0$$

$$\overline{\mu} = \mu + \eta s$$

$$\overline{\lambda} = \lambda - \frac{3}{2}\eta s\,;$$

moreover the force $g(z)$ starts acting only at $t = 0$ and it was zero before.

In solving this elementary problem one can take into account its particular symmetry and put straightforwardly

$$u_x \equiv 0\,, \quad u_y \equiv 0\,;$$

moreover u_z will depend on z only, so that the only useful constitutive equation is

$$\sigma_{zz} = (2\mu + \lambda)\partial_z u_z\,.$$

As boundary conditions one can use

$$z = 0 \quad \sigma_{zz} = (2\mu + \lambda)\partial_z u_z = 0$$

$$z = -\eta \quad u_z = 0\,,$$

the second being a consequence of the hypothesis of incompressibility of the substratum.

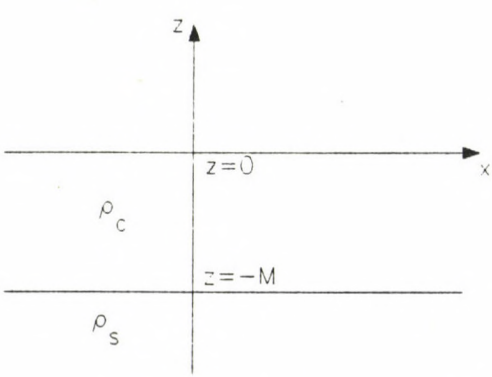

Fig. 15. Schematic representation of a plane layer (crust) of density ρ_c, width M, floating on a liquid substratum of density ρ_S

The solutions are:

1. $u_z = \frac{1}{2}\rho_c \frac{g_0}{\Lambda}(z^2 - \eta^2)$
   $(\Lambda = 2\mu + \lambda)$,
2. $u_z = \frac{1}{2}\rho_c \frac{g_0}{\Lambda}(z^2 - \eta^2) - \frac{1}{6}\rho_c \frac{G}{\Lambda}(z^3 + \eta^3)$,
3. $u_z = \frac{\rho_c \cdot g_0}{\Lambda\eta\gamma}(z^2 - \eta^2) - \{1 - e^{-\gamma t}\}$
   $\left(\gamma = \frac{2\Lambda}{\eta}\right)$.

Example 6.3 We refer to an unperturbed situation as it was discussed in the previous Exercise 6.2, in case 1, *i.e.* for an elastic body of density ρ_c, submitted only to a constant gravity g_0 as external body force.

However we assume now that a load, or an unload (imagined to represent a situation of glaciation or post-glacial rebound) is added to the upper surface.

We use an approach to the solution of this problem, which is due to Bakus and Gilbert and is easily generalizable to multilayered bodies.

Let us first of all linearize our equations around the reference configuration.

The equilibrium equations (plain strain case) are

$$\begin{cases} \dfrac{\partial \sigma_{xx}}{\partial x} + \dfrac{\partial \sigma_{xz}}{\partial z} = 0 \\ \dfrac{\partial \sigma_{xz}}{\partial x} + \dfrac{\partial \sigma_{zz}}{\partial z} = \rho_c g_0 \, ; \end{cases} \tag{216}$$

if we split the stress tensor into

$$\sigma = \sigma_0 + \tau \tag{217}$$

with σ_0, the reference stress, given by

$$\sigma_0 = \begin{pmatrix} 0 & 0 \\ 0 & \rho_c g_0 z \end{pmatrix} \tag{218}$$

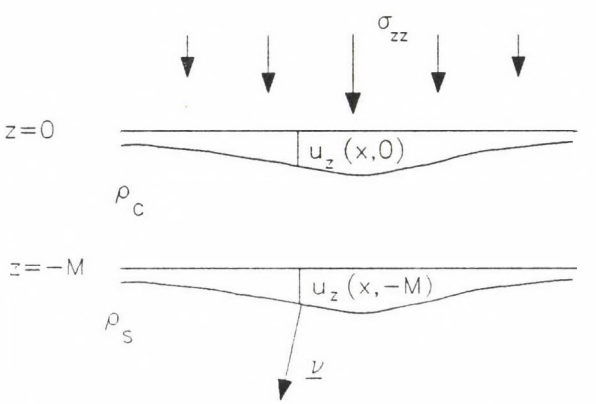

Fig. 16. The same two layers as in Fig. 15 with added load on the top surface

and τ infinitesimal of the first order, we see that τ satisfies simply the homogeneous equilibrium conditions (null body forces).

If $\mathbf{u}_0$ is the reference displacement vector

$$\mathbf{u}_0 = \begin{pmatrix} 0 \\ \frac{\rho_c g_0}{\Lambda}(z^2 - M^2) \end{pmatrix} \tag{219}$$

we put

$$\mathbf{u} = \mathbf{u}_0 + \mathbf{v} \,. \tag{220}$$

Since Hooke's law is linear, we have then

$$\begin{aligned} \tau &= 2\mu e(\mathbf{v}) + \lambda Tr e(\mathbf{v}) I \,, \\ e(\mathbf{v}) &= \begin{pmatrix} \partial_x v_x & \frac{1}{2}(\partial_x v_z + \partial_z v_x) \\ \frac{1}{2}(\partial_x v_z + \partial_z v_x) & \partial_z v_z \end{pmatrix} . \end{aligned} \tag{221}$$

A little more elaborate is the situation with the boundary conditions; let us treat them separately.

At $z = 0$; our condition can be written

$$\begin{aligned} \sigma(\boldsymbol{\nu}) &= \begin{pmatrix} 0 \\ P(x) \end{pmatrix} \\ (P(x) &= \text{given load}) \,. \end{aligned} \tag{222}$$

Since the upper surface is deformed, we have, to the first order

$$\boldsymbol{\nu} = \begin{pmatrix} -\partial_x u_z(x,0) \\ 1 \end{pmatrix} . \tag{223}$$

Moreover if we compute σ_0 on the deformed surface we get from Eq. (218)

$$\sigma_0 = \begin{pmatrix} 0 & 0 \\ 0 & \rho_c g_0 u_z(x,0) \end{pmatrix} . \tag{224}$$

By using Eqs (217), (223) and Eq. (224) in Eq. (222) and retaining only first order quantities, we find

$$\sigma_0 \boldsymbol{\nu} + \tau \boldsymbol{\nu} \cong \begin{pmatrix} \tau_{xz} \\ \rho_c g_0 u_z + \tau_{zz} \end{pmatrix} . \tag{225}$$

Since all quantities in Eq. (225) are already infinitesimals we can take this relation as holding at $z = 0$ and put

$$\begin{cases} \tau_{xz} = 0 \\ \rho_c g_0 u_z + \tau_{zz} = -P(x) \end{cases} \tag{226}$$

At $z = -M$; now the external force acting from the substratum on our layer is a pressure p in direction $-\boldsymbol{\nu}$.

Note that in this case the external normal is given by

$$\boldsymbol{\nu} = \begin{pmatrix} \partial_x u_z(x, -M) \\ -1 \end{pmatrix} . \tag{227}$$

Moreover the pressure will be

$$p = p_0 - \rho_s g_0 u_z$$

where p_0 is just the reference pressure at $z = -M$ while the other term is just the Archimedean force pushing up the deeper parts of the deformed lower boundary.

So, neglecting second order effects, the external forces at $z = -M$ are

$$-(p_0 - \rho_s g_0 u_z)\boldsymbol{\nu} = \begin{pmatrix} -p_0 \partial_x u_z \\ p_0 - \rho_s g_0 u_z \end{pmatrix} . \tag{228}$$

As for the internal stress vector one can write

$$\sigma\boldsymbol{\nu} = \sigma_0 \boldsymbol{\nu} + \tau\boldsymbol{\nu} , \tag{229}$$

where σ_0 on the deformed surface is

$$\sigma_0 = \begin{pmatrix} 0 & 0 \\ 0 & \rho_c g_0(-M + u_z) \end{pmatrix} . \tag{230}$$

Whence, again neglecting second order infinitesimals,

$$\sigma\boldsymbol{\nu} = \begin{pmatrix} -\tau_{xz} \\ -\rho_c g_0(-M + u_z) - \tau_{zz} \end{pmatrix} . \tag{231}$$

Equating Eqs (231) and (228), and noting that $\rho_c g_0 M = p_0$, we find

$$\begin{cases} \tau_{xz} = p_0 \partial_x u_z \\ \tau_{zz} = (\rho_s - \rho_c) g_0 u_z . \end{cases} \tag{232}$$

We note that in this linearized expression τ_{zz} depends nicely on the buoyancy force which in turn is proportional to the density contrast $\rho_s - \rho_c$.

We can now summarize the problem as:

$$\begin{cases} \dfrac{\partial \tau_{xx}}{\partial x} + \dfrac{\partial \tau_{xz}}{\partial z} = 0 \\ \dfrac{\partial \tau_{xz}}{\partial x} + \dfrac{\partial \tau_{zz}}{\partial z} = 0 \end{cases} \tag{233}$$

$$\begin{cases} \tau_{xx} = \Lambda \partial_x v_x + \lambda \partial_z v_z \\ \tau_{zz} = \Lambda \partial_z v_z + \lambda \partial_x v_x \quad (\Lambda = 2\,\mu + \lambda) \\ \tau_{xz} = \mu(\partial_x v_z + \partial_z v_x) \end{cases} \tag{234}$$

$$\begin{cases} \tau_{xz} = 0 \\ \qquad\qquad\qquad\qquad z = 0 \\ \tau_{zz} + \rho_c g_0 v_z = -P(x) \end{cases} \tag{235}$$

$$\begin{cases} \tau_{xz} - p_0 \partial_x v_z = 0 \\ \qquad\qquad\qquad\qquad z = -M \\ \tau_{zz} - (\rho_s - \rho_c) g_0 v_z = 0\,. \end{cases} \tag{236}$$

Now to solve the problem we can perform Fourier transformations of all quantities along the x axis, like in the Example 6.1; the only peculiarity is that instead of using the first of Eq. (234) as it is, we rather use a combination with the second of Eq. (234) giving

$$\Lambda \tau_{xx} - \lambda \tau_{zz} = (\Lambda^2 - \lambda^2) \partial_x v_x\,. \tag{237}$$

So performing the transformations, eliminating $\hat{\tau}_{xx}$ from Eq. (233) with the help of Eq. (237) and rearranging we get

$$\partial_z \cdot \begin{pmatrix} \hat{v}_x \\ \hat{v}_z \\ \hat{\tau}_{xz} \\ \hat{\tau}_{zz} \end{pmatrix} = \begin{pmatrix} 0 & ip & \mu^{-1} & 0 \\ i\lambda\Lambda^{-1}p & 0 & 0 & \Lambda^{-1} \\ p^2(\Lambda^2 - \lambda^2)\Lambda^{-1} & 0 & 0 & i\lambda\Lambda^{-1}p \\ 0 & 0 & ip & 0 \end{pmatrix} \cdot \begin{pmatrix} \hat{v}_x \\ \hat{v}_z \\ \hat{\tau}_{xz} \\ \hat{\tau}_{zz} \end{pmatrix} \tag{238}$$

which we synthetize, with obvious notation, in

$$\partial_z \xi(z,p) = A(p)\xi(z,p)\,. \tag{239}$$

Since the propagator $A(p)$ in Eq. (239) is independent from z, we can formally solve this system by

$$\xi(z,p) = e^{A(p)z}\xi(0,p)\,, \tag{240}$$

where $\xi(0,p)$ still contains 4 unknowns to be determined.

Using the transforms of Eqs (235), (236), namely

$$\begin{cases} \hat{\tau}_{xz} = 0 \\ \hat{\tau}_{zz} + \rho_c g_0 \hat{v}_z = -\hat{P} \end{cases}$$

$$\begin{cases} \hat{\tau}_{xz} + ip_0 p\hat{V}_z = 0 \\ \\ \hat{\tau}_{zz} - (\rho_s - \rho_c)g_0\hat{V}_z = 0 \end{cases}$$

one gets exactly the relations required to determine $\xi(0)$, and the problem is solved.

The only numerical problem left is the computation of e^{Az}, which however can be obtained by singular value decomposition or by series expansion.

It is interesting to observe that this approach in practice reduces our problem to an initial value problem for a system of first order differential equations. This is not so natural indeed when we have to solve a B.V.P. involving two surfaces at $z = 0$ and $z = -M$; yet this could be a natural approach to discuss a problem in which we really have data on one surface only. This is the case for example, in geodetic practice, if we perform a control survey on the upper boundary, coming to know the displacement field $\mathbf{u}$ at $z = 0$; on the same time if we don't have any load, we also can state that

$$\sigma_{xz} = \sigma_{zz} = 0 \quad (z = 0)\,.$$

This means that a deformation happens because there are forces acting on the other boundary of our layer and we might be interested in knowing such forces.

In this situation we could think of solving a true initial value problem and computing the stress tensor at $z = -M$.

Naturally to follow such a program one has to cautiously take care of the non well-posedness of such a problem. It is known in fact that a Cauchy problem for an elliptic operator gives rise to unstable solutions, which have then to be suitably regularized.

Example 6.4 (Kirchhof theory of thin plates). In view of the relative smallness of the width of "elastic plates" on the earth surface, with respect to their horizontal dimensions, it is often appropriate to treat them as thin plates.

The basic simplification here is obtained by assuming that after deformation the straight lines orthogonal to the plates become again almost straight lines orthogonal to the deformed plate (cf. Fig. 17).

Under this condition the displacement in the x direction is opposite at the lower and higher side of the plate, with respect to the mid-surface; in fact the higher side, in the example shown in Fig. 17, is contracted while the lower is extended.

Looking more closely at Fig. 18 one recognizes that

$$u_x = -z\frac{d\zeta}{dx}\,. \tag{241}$$

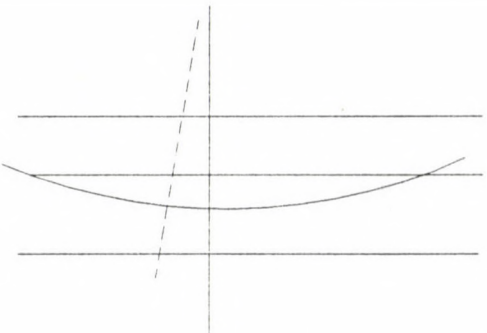

Fig. 17. Hypothesis of simple deformation of a thin plate

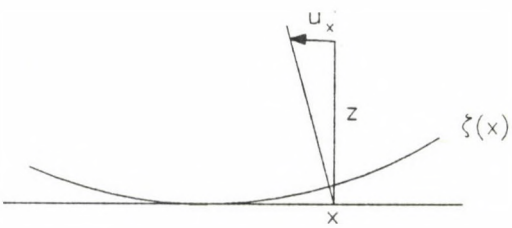

Fig. 18. Derivation of u_x as function of z

Generalizing to two dimensions one has

$$\begin{cases} u_x = -z\frac{\partial\zeta}{\partial x} \\ u_y = -z\frac{\partial\zeta}{\partial y}\,. \end{cases} \tag{242}$$

Moreover the planes parallel to the surface of the plate will deform almost parallely, so that we expect

$$u_z = \zeta(x, y) \tag{243}$$

with

$$\frac{\partial u_z}{\partial z} \cong 0\,.$$

Under these conditions we find

$$e_{xx} = -z\zeta_{xx} \quad e_{xy} = -z\zeta_{xy} \quad e_{yy} = -z\zeta_{yy}$$

$$e_{xz} = e_{yz} = 0 \qquad e_{zz} = \frac{\nu}{1-\nu} z\Delta_0\zeta$$

$$(\Delta_0\zeta = \zeta_{xx} + \zeta_{yy})\,.$$

We can then compute the internal energy

$$W = \tfrac{1}{2}\int Tr\sigma e\, dB = \tfrac{1}{2}\int \left[2\mu Tre^2 + \lambda(Tre)^2\right] dB =$$

$$= \tfrac{1}{2}\int_{-\frac{h}{2}}^{\frac{h}{2}} dz z^2 \left\{\int dx\, dy \left[\zeta_{xx}^2 + \zeta_{yy}^2 + 2\zeta_{xy}^2 + \tfrac{\nu^2}{(1-\nu)^2}(\Delta_0\zeta)^2\right] \cdot 2\mu + \right.$$

$$\left. + \left[-\zeta_{xx} - \zeta_{yy} + \tfrac{\nu}{1-\nu}\Delta_0\zeta\right]^2 \cdot \lambda \right\} .$$

As we see, we can write, with suitable constants, integrating in dz between $-\frac{h}{2}$ and $\frac{h}{2}$

$$W = \frac{h^3}{8}\int dx\, dy \left[\zeta_{xx}^2 + 2\zeta_{xy}^2 + \zeta_{yy}^2 + A(\Delta_0\zeta)^2\right] .$$

Then the total energy functional, in the presence of a body force F_z and external tractions T_+, T_- applied in the z direction too, on the two faces, can be written as (cf. Eq. (170))

$$\begin{aligned} E(\zeta) &= \tfrac{h^3}{8}\int dx\, dy \left[\zeta_{xx}^2 + 2\zeta_{xy}^2 + \zeta_{yy}^2 + A(\Delta_0\zeta)^2\right] + \\ &- \int dx\, dy \zeta \left[\int_{-\frac{h}{2}}^{\frac{h}{2}} dz F_z + T_+ - T_-\right] . \end{aligned} \tag{244}$$

The stationarity of Eq. (244) provides the sought equation of equilibrium.

Disregarding the boundary behaviour of ζ (i.e. setting $\zeta = \zeta_x = \zeta_y = 0$ on the boundary) we see that we obviously have

$$\delta \int (\Delta_0\zeta)^2 dS = -\int \delta\zeta(\Delta_0^2\zeta) dS \quad . \tag{245}$$

Moreover one has

$$\begin{aligned} \delta \int \left[\zeta_{xx}^2 + 2\zeta_{xy}^2 + \zeta_{yy}^2\right] dS &= -2\int \delta\zeta \left[\zeta_{xxxx} + 2\zeta_{xxyy} + \zeta_{yyyy}\right] dS = \\ &= -2\int \delta\zeta(\Delta_0^2\zeta) dS . \end{aligned} \tag{246}$$

Therefore if we put

$$q = \int_{-\frac{h}{2}}^{\frac{h}{2}} F_z dS + T_+ - T_-$$

we find, as variational equation, from Eq. (244),

$$D\Delta_0^2\zeta = q ; \tag{247}$$

a careful computation of the constant D shows that it is

$$D = \frac{Eh^3}{12(1-\nu^2)} . \tag{248}$$

We will see an application of this theory in the next exercise.

Exercise 6.3 Assume to have an infinite, elastic thin plate floating on an incompressible liquid of density ρ_s; the flexural rigidity of the plate, D, is given (cf. Eq. (248)).

On the upper face a load q^+ is applied; find the equation of the equilibrium configuration and solve it for the case that

$$q^+ = -P_0\delta(x) , \tag{249}$$

simulating the effect of a mountain chain concentrated on the line $x = 0$.

To solve the problem, note that due to buoyancy, the T_- force acting on the lower surface is essentially

$$T_- = \rho_s g_0 \cdot \zeta , \tag{250}$$

so that neglecting any body force we have

$$D\Delta_0^2\zeta + (\rho_s g_0)\zeta = q^+ . \tag{251}$$

To solve the case of the concentrated load Eq. (249) one can note that ζ is expected to be function of x only.

The sought solution has the form

$$\zeta = Ae^{-\lambda|x|}(\cos\lambda \mid x \mid + \sin\lambda \mid x \mid) ; \tag{252}$$

show that A and λ must be determined from the equations

$$\begin{cases} 4\lambda^4 D = \rho_s g_0 \\ 8\lambda^3 A = -\frac{P_0}{D} . \end{cases}$$

References

Artynshkov E V 1983: Geodynamics. In: Developments of Geotectonics, No. 18, Elsevier

Fung U C 1965: Foundations of solid mechanics. Prentice Hall

Lambeck K 1988: Geophysical geodesy. Clarendon Press, Oxford

Landau L D, Lifsits E M 1966: Theories de l'elasticité. In: Physique Theorique, Toure 7, Edition MIR, Moscow

Nečas J, Hlavacek I 1981: Mathematical theory of elasticity and elasto-plastic bodies. In: Studies in Applied Mechanics 3, Elsevier

Teisseyre R et al. 1984[7]: Constitution of the earth's interior. In: Physics and evolution of the earth's interior, Vol. 1, Elsevier

[7]These references are by far not complete; they have just been useful to read for the author when he was preparing these lecture notes.

Acta Geod. Geoph. Hung., Vol. 29(3–4), pp. 339–362 (1994)

LOCAL OBSERVATION AND INTERPRETATION OF GEODYNAMIC PHENOMENA

W ZÜRN[1]

Several examples of geodynamic phenomena which were observed locally by typical sensors (seismographs, gravimeters, strain- and tiltmeters) are described in order of increasing signal period. It will be stressed that local observations alone do usually not suffice for the interpretation and that local, regional or global networks of sensors provide significance to such observations. The examples are taken from the fields of seismic surface waves and free oscillations, oscillations in the earth's core, earth tides, near-field and secular deformations.

Keywords: core oscillation; gravimeter; seismic surface waves; tides; tiltmeter

1. Introduction

The following discussion of the local observation and interpretation of geodynamic phenomena is concerned with a very small subset of the possibilities available under the title. The first restriction concerns the type of observation: phenomena causing measurable displacements, tilts, strains and gravity variations will only be discussed. The second constraint concerns the frequency range, periods larger than 3 minutes will only be delt with. The third limitation is on examples, because it is nearly impossible to give a full review of the huge amount of work done even under these restrictions. In addition, all the examples except two come from work the author was somehow involved in, which is the fourth constraint. This subset does not contain examples of electromagnetic phenomena, of body wave seismology and of well tides, among many others fields. The instruments used in the examples are not described in this paper. An excellent and intensive account of strain and tilt instrumentation and associated problems is provided by Agnew (1986), including all such instruments mentioned in the following examples. Gravimeters are described in geodetical textbooks (*e.g.* Torge 1989) and one modern, widely used, broadband seismometer is described in detail by Wielandt and Streckeisen (1982).

Figure 1 depicts the frequency ranges of locally observable signals from geodynamic phenomena discussed in the paper and lists the sources producing such signals.

Space geodetic methods have become available during the last two decades and in this short time have become powerful tools to study geodynamic phenomena, as the contents of this volume document very well. Their potential certainly has not been fully exploited yet. These methods provide displacements of points on the earth's surface with accuracies approaching the mm-level and with respect to certain reference systems. The advantages of the local observation of earth deformation are

[1]Black Forest Observatory Schiltach, Heubach 206, D–77709 Wolfach, Germany

1217-8977/94/$ 4.00

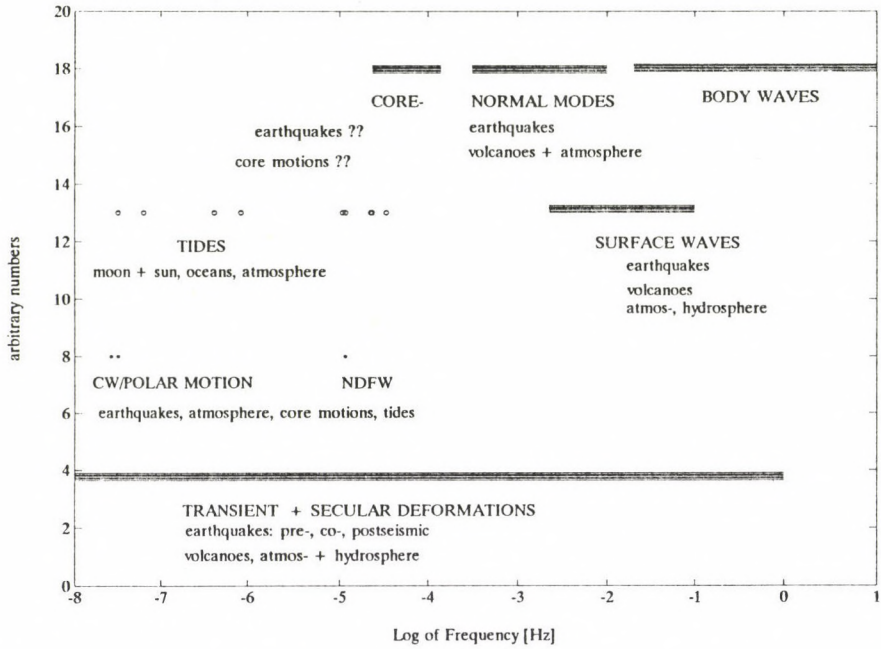

Fig. 1. Geodynamical signals, their sources and approximate frequency ranges. These signals are normally observed with seismometers, gravimeters, tilt- and strainmeters. Vertical scale does not indicate amplitude of signals

the high temporal resolution and precision which can be achieved with continuously recording seismographs, gravimeters, tilt- and strainmeters. Major problems are caused by the instrumental drift, the lack of an absolute reference and the coupling to the earth.

One problem common to both local and space geodetic techniques is their broad band nature. What is considered the signal in one investigation must be considered noise in another: one man's noise is another man's signal. For example, surface waves and free oscillations from large earthquakes severely disturb earth tide records, however in this case the two signals are well separated in the frequency domain. This is not always the case: the oceanic tides disturb the earth tides and there is no way to separate these two signals outside of the interpretation of the recorded effect in terms of both. Atmospheric loading causes signals (or noise) in the whole frequency band discussed here and since at present the understanding is at least incomplete, all signals are contaminated. It is understood that no noise is good and that therefore the instrumental and local noise level should be below the noise created by the earth. It is also understood that often it is impossible to minimize noise from the environment and then optimization must be the goal. For example, if a volcano to be studied is located on a small island in the 'roaring forties' of the Southern oceans, there is no way to get noise levels as low as in the middle of a continent under stable high pressure conditions. However, optimization

is always possible to a certain extent and the scientific value of such measurements critically depends on the efforts taken in this respect. Some of the examples below hopefully make this point very clear.

One important problem concerns the interpretation of locally observed signals from supposedly geodynamic phenomena. In one record one can never be sure, whether the unexpected signal or anomaly in a signal is caused by the instrument or its environment or by a true signal from the earth. All possible alternatives, especially of instrumental or man-made origin must be checked. There is abundant literature, especially in the field of the theoretically predicted core mode signals in gravity, where single records were used to put forward the claim of detection of new phenomena of geodynamical significance, without checking with other possibilities and without corroboration from other stations. A solid interpretation always needs support from records of other stations. The following examples almost without exception demonstrate this general problem. One very fine well-known example is the first observation of the earth's seismic free oscillations after the 1960 great Chilean earthquake, where several groups compared their observations at an international meeting (*e.g.* Bullen 1965, pp. 260–261). Anomalous (unexpected) signals are of high interest, of course, because they could present new information. However, such new signals often cannot easily be distinguished from the noise in the records and therefore the danger of erroneous interpretation of noise in the data is big (Slichter 1961: 'Why does the Lord always put the most interesting signals right at the noise level?'). Corroboration of a new observation is always mandatory, possibly with data from other stations. Again, the examples should demonstrate this important point. This does not necessarily mean that an observation is significant, if instruments at different stations show the same 'signal'. For instance, if all gravimeters in a region at different places show a drift rate of the order of a microgal/day in the same direction, this does not mean that the ground in the region rises or falls by 3 mm/day. In this case another interpretation, the well-known instrumental drift common to all these instruments with differing rates, is the correct one.

In the following the examples are arranged in order of decreasing signal frequency.

2. Examples of geodynamic phenomena

2.1 Globe-circling Rayleigh waves generated by the Plinian eruption of Mount Pinatubo, Philippines on June 15, 1991

Surface wave and free oscillation seismology are well developed fields of research both theoretically and observationally. The observations have been extensively exploited for studies of the interior of the earth and the source processes of earthquakes (especially very big ones). This is well documented in textbooks on these fields and seismology in general.

However, a completely new phenomenon (truly geodynamical in a wide sense) involving seismic surface waves and free oscillations was detected in local observations in 1991 by two research groups, but then corroborated by the study of seismograms

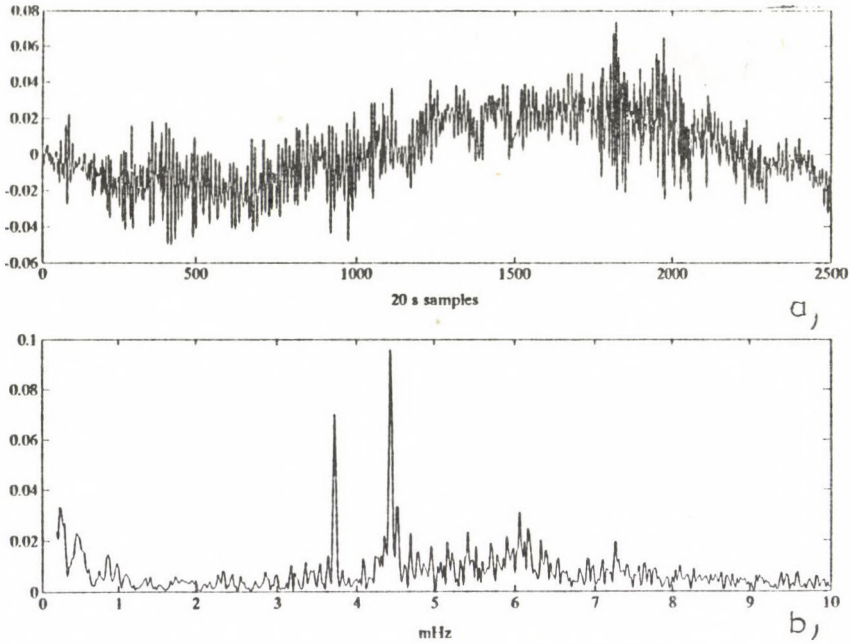

Fig. 2. a) Time series from the mode channel of LaCoste-Romberg earth tide gravimeter ET 19 equipped with electrostatic feed-back starting at 0600 UTC on June 15, 1991 and lasting nearly 14 hours. The instrument is located in the BFO mine observatory in SW Germany. The first 100 samples show the wave train R_4 (a Rayleigh wave which has traveled over the major arc to the station from the source and one full additional circumference of the globe) from an earthquake in the South Sandwich Islands. Then the signal from the Mount Pinatubo eruption starts. The peak-to-peak amplitude of the oscillations is slightly less than 1 microgal. b) Amplitude spectrum of above time series after application of Hanning window. Two strong resonances at 3.68 and 4.44 mHz are visible

from the global networks of digital long period seismographs (Kanamori and Mori 1992, Widmer and Zürn 1992, Kanamori et al. 1992). During the climactic phase of the June 15, 1991 Mount Pinatubo eruption an essentially bichromatic signal with frequencies of 3.68 and 4.44 mHz was recorded on gravimeters and very long-period (VLP) seismometers world wide. Figure 2 shows the record from a gravimeter at the Black Forest Observatory in SW-Germany and its spectrum as an example. Essentially all modern long period seismographs, including some superconducting gravimeters, observed these oscillations. The narrow band nature of this signal distinguishes these recordings from the usual broad-band signal recorded after large earthquakes. Group velocity estimates and particle motion show that the signals propagate as Rayleigh waves. The bichromatic spectra, which had not been observed during previous Plinian eruptions, can only be explained by source models which provide for harmonic forcing of the solid Earth. These oscillations last from about 0700 to 2000 UT on June 15, 1991. The source was then located using a cross-correlation method. From the lags between the individual traces a group velocity of 3.78 km/s was estimated which corresponds to the group velocity of Rayleigh waves

with the periods of the bichromatic signal (see Fig. 2 of Widmer and Zürn 1992). The Pinatubo source has excited the spheroidal fundamental modes ${}_0S_{28}$ and the triplet ${}_0S_{36}$, ${}_0S_{37}$ and ${}_0S_{38}$, which are established by globe-circling Rayleigh-waves. Thus the source spectrum encompasses three spheroidal modes near 4.44 mHz and only one at 3.7 mHz. This shows that the source spectrum is narrower at 3.7 mHz than at 4.44 mHz.

Zürn and Widmer (1992) studied the coherence of the signals using the Phasor-Walkout method revisited by Zürn and Rydelek (1994) and became convinced that the source of the Rayleigh waves radiated in phase for several hours.

Kanamori and Mori (1992) show in their Fig. 6 the microbarogram from Clark Air Force Base located about 21 km from the volcano. This record shows pressure pulses from individual explosions as well as continuous oscillations with peak amplitudes of about 300 Pa during the climactic phase of the eruption. Unfortunately the resolution in time is not good enough to determine the periods of these oscillations. Under the assumption that the periods of the atmospheric and seismic oscillations are identical, these authors have shown that the amplitudes of the atmospheric pressure oscillation are sufficient to generate Rayleigh waves with the observed amplitudes.

We searched the digital seismic data sets for other observations of the Pinatubo type, *i.e.* quasiharmonic wave trains of long duration. For recent eruptions of Bezymianny, Mount St. Helens, Galunggung, Colo, Mount Etna, Redoubt, Avachinsky and Mt. Spurr this kind of waves could not be observed in the seismic records. However, our search was successful in the case of El Chichón (Southern Mexico) on April 4, 1982 (Widmer and Zürn 1992). Again two peaks were observed with frequencies of 3.703 and 5.140 mHz, however, the oscillations lasted only for one hour.

No VLP seismograms are available for the great Krakatoa explosion in 1883 or the Tunguska meteorite impact in 1908. However, these events were recorded on microbarographs world-wide and lead to several theoretical investigations of the propagation of pulses in the atmosphere. From the published records one can clearly see that these signals were not of the Pinatubo or El Chichón type, but had a short duration, *i.e.* one dispersed wave train and a broad spectrum instead (Pekeris 1939, 1948). For the case of Krakatoa, however, Kanamori et al. (1992) claim to have observed atmospheric oscillations with periods of about 300 s in a barogram from Batavia (Java) about 200 km from Krakatoa.

Short dispersive wave trains were also typical for barograms and seismograms from atmospheric nuclear explosions, well documented in the literature (*e.g.* Gossard and Hooke 1975). Again the observations were made in the far-field of the events.

When Mount St. Helens exploded on May 18, 1980 modern digital long-period seismographs were observing seismic body waves and dispersive Rayleigh wave trains very similar to the ones observed after earthquakes. Also dispersive pressure wave-trains with speeds around 300 m/s and a broad spectrum were observed world-wide on barographs and seismographs or gravimeters (Müller and Zürn 1983, Bolt and Tanimoto 1981). Recently, Kanamori et al. (1992) claimed the observation

of an atmospheric oscillation with 300 s period on a digital long period seismograph at Longmire (LON, California) in a distance of 67 km from Mount St. Helens. As in the case of Krakatoa, these oscillations were only observed fairly close to the volcano.

Georges (1973) reports in a study of infrasound emitted from severe convective storms in USA that ionosounders which use radio waves to map the position of reflective horizons in the ionosphere find bichromatic waves above convective storms with frequencies of 3.7 and 4.8 mHz. Georges also finds that the line at 3.7 mHz remains unchanged from storm to storm whereas the higher frequency mode at 4.8 mHz varies by an appreciable (but unspecified) amount from event to event. These waves travel in vertical direction and only exist above severe convective storms. These spectral peaks are not observed on barograms, which on the other hand show rather broad spectral features. It appears, that these harmonic oscillations do not propagate very far away from their source.

Note that the lower frequency reported by Georges (1973) coincides nicely with the lower frequencies observed on seismographs for Pinatubo and El Chichón and that the higher frequency is between the two higher frequencies observed for the two volcanoes.

Kanamori and Mori (1992) suggested, that these are free oscillations of the atmosphere excited by the eruption. According to their analysis the pressure variations observed at Clark Air Force Base are large enough to explain the observed seismic amplitudes, when a source area of a radius of 37 km is assumed. However, in order to explain the phase coherence of the source one has to assume that the oscillator was either only excited once, or else, that reexcitation occurred perfectly timed. Single excitation can be ruled out because of the long duration of the climactic phases of the two eruptions. Zürn and Widmer (1994) therefore propose a source model with positive feed-back between the atmospheric free modes and the eruption and/or the plume, thus constituting a self-excited oscillator. This oscillator generates Rayleigh waves by periodically pushing on the surface of the earth, and Lamb waves (atmospheric acoustic-gravity waves propagating horizontally) by periodically pushing on the neighbouring atmosphere. We propose that the type of atmospheric oscillations which participate in the feed-back correspond to waves travelling vertically up and down. These obviously do not propagate away from the source region. The bichromatic spectra now lead to an interpretation as the cutoff-frequencies for the acoustic and gravity waves. Presently the evidence which Kanamori and Mori (1992) and Widmer and Zürn (1992) have collected is circumstantial and the mechanisms proposed need further verification.

2.2 First clear observation of the Twisting mode ${}_0T_2$ of the earth

The following is a unique example of a single-station observation, however, several pieces of evidence suggest its significance. Several authors claimed the observation of the fundamental torsional free oscillation of the earth after the great Chilean earthquake of 1960 on records by strainmeters and tiltmeters. The predicted period of oscillation of this mode ${}_0T_2$ for earth model 1066A (Gilbert and Dziewonski

1975) is 44.85 minutes (corresponding to 0.38008 mHz). According to Smith (1961) and Derr (1969) the best of these observations came from the Isabella (California) quartz extensometer. This record was reanalyzed by Widmer et al. (1992) with the multiple taper method by Park et al. (1987) and it was shown that the claimed observation of mode ${}_0T_2$ is spurious at best. The strong strike-slip earthquake on the Macquarie Ridge on May 23, 1989 offered a new opportunity to study low order fundamental modes. Widmer et al. (1992) studied many records from horizontal seismometers of the global digital seismic networks for this earthquakes and found the mode ${}_0T_2$ in the mode channel of the invar wire strainmeter (King and Bilham 1976) with azimuth $N2°E$ at the Black Forest Observatory (BFO). The corresponding spectral peak proved to be significantly above the noise in this digital record when analyzed with the same method which was used to show that the observation in the 1960 Isabella record was not significant. In this case records from other stations could not be used to confirm the observation. However, at BFO two more identical strainmeters with azimuths $N60°E$ and $N300°E$ recorded the free modes of the same earthquake. A simple analysis shows that the calibrated sum of the records from three strainmeters with azimuths 120° apart from each other results in 1.5 times the areal strain. Since torsional modes do not produce areal strains, the observed spectral peak (quite clear in two of the individual records) should disappear into the noise in the sum of the three records. This was shown to be the case (Fig. 3, top). In addition, from the three non-collinear strain records, the shear strain (which unlike areal strain is not an invariant of the strain tensor) in any coordinate system can be determined. This was also tested and Fig. 3 (bottom) shows the spectrum of the shear strain thus obtained for the coordinate system in which the amplitude of ${}_0T_2$ was a maximum. Clearly, the series of fundamental torsional modes is visible above the noise level, while here the spheroidal modes seen in the areal strain spectrum have dropped into the noise. Thus the sheer observation of a spectral peak at the right frequency (0.37660 mHz) is corroborated by the simultaneous additional observation of other properties of the mode used to interpret the spectral peak. Remember that the observation of a peak in the spectrum of a time series does not contain information about its physical cause. The estimated shear strain amplitude of the mode (about 2 hours after the quake) was $10 \cdot 10^{-12}$, while the shear strain computed for this mode at this station from the moment tensor of this quake is $6 \cdot 10^{-12}$. This shows that the observation does not need exotic physics to explain its observed amplitude.

2.3 Free modes in the Earth's cores

Abundant literature exists identifying spectral peaks in the period band between one and 24 hours in single gravity or vertical seismograf records as signals from the translational modes of the earth's inner core (Slichter modes) or oscillations in the liquid outer core (undertones). Such modes were predicted theoretically from standard earth models constrained by many seismological data to be properties of the earth (Slichter 1961, Busse 1974, Smith 1976, Crossley and Rochester 1980). So far none of these identifications was confirmed by other observations and for

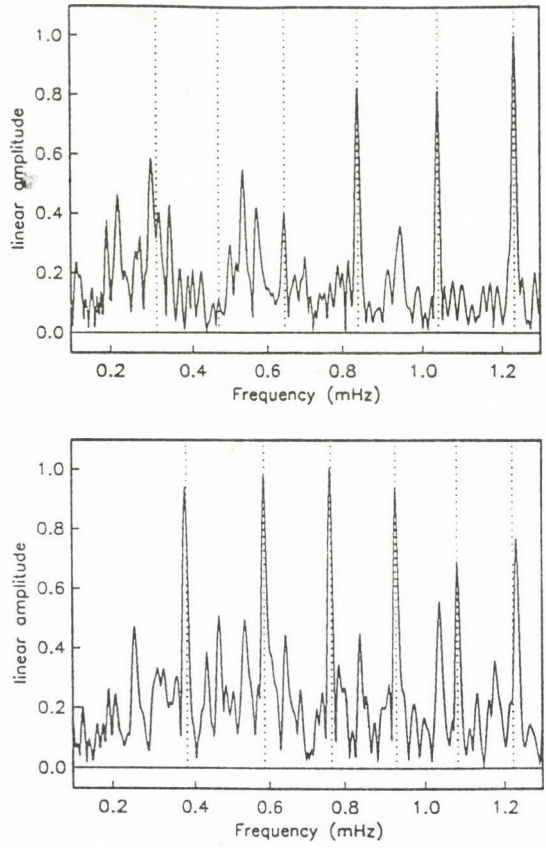

Fig. 3. Amplitude spectra of two different linear combinations of three BFO invar wire strainmeter records after the Macquarie Rise event, 1989. Time series of 40.0 hours were multiplied by a Hanning window before Fourier Transformation. Dotted vertical lines from left to right indicate theoretical degenerate eigenfrequencies of fundamental spheroidal modes ${}_oS_l$ in the upper panel and fundamental toroidal modes ${}_oT_l$ in the lower panel for $l = 2$ to 7. The areal strain-combination (simple sum, top panel) shows the fundamental spheroidal modes ${}_oS_7$, ${}_oS_6$, ${}_oS_5$ and, less certain, ${}_oS_4$ above the noise only. Toroidal modes, especially, cannot be identified here because they do not produce areal strain, while they are maximized in a certain shear combination (lower panel). ${}_oT_2$, ${}_oT_3$, ${}_oT_4$, ${}_oT_5$, ${}_oT_6$ and ${}_oT_7$ are clearly above the noise. Also some S-modes appear to be present at reduced amplitude with respect to the noise level

this and other reasons must be considered invalid. Of course, the identification and investigation of these modes and determination of their properties would shed new light on our understanding of the earth's cores and thus on geodynamic phenomena involving the cores, *e.g.* the geodynamo problem and the magnetic field of the earth. If these modes can get excited (*e.g.* by earthquakes or outer core turbulence) to observable amplitudes is an unsolved problem and thus the search for them may be in vain.

The commercial availability of superconducting gravimeters (Goodkind 1991) has created new hope for the detectability of these possible tiny signals from the

earth's cores and has led to new claims for positive identification of spectral features with the core modes (Melchior and Ducarme 1986, Smylie 1992). The first authors found peaks at 13.89 hours period in the record from the superconducting gravimeter at Brussels after deep earthquakes. The second author multiplied spectra from four records from superconducting gravimeters located in Central Europe and identified a triplet of peaks which he claims to be consistent with his computed triplet of frequencies for the Slichter mode. While the first authors really had a relative maximum of power at their 'mode', this was not so clear for the triplet of Smylie (1992). Observationally Smylie's evidence is extremely weak (not to say non-existent) and relies heavily on the frequencies he computed numerically. However, these computations have been questioned also for theoretical reasons by Crossley et al. (1992) and therefore the evidence becomes even more dubious. This debate about the correct theory for the Slichter modes is continuing (see several comments in *Geophys. J. Int.*, 115, 1993, 3). The 'core mode' interpretation by the first authors has been questioned by many papers (*e.g.* Zürn and Rydelek 1994). Cummins et al. (1991) in their analysis summarize the story of this 'detection' and demonstrate nicely, how such searches should be performed. Oscillatory modes have more

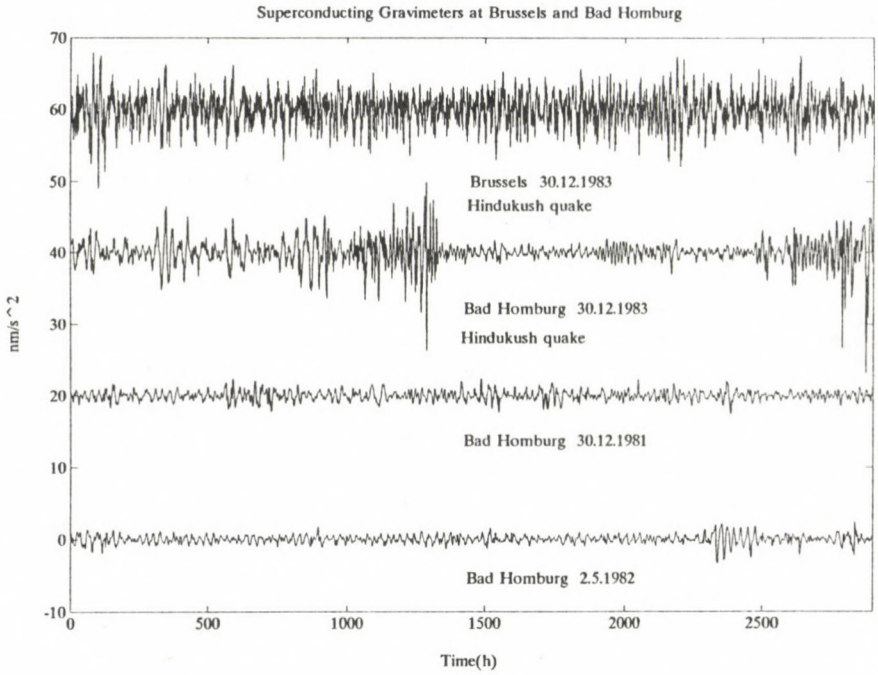

Fig. 4. Residual gravity (tides and barometric pressure effect model subtracted and high-pass filtered to remove periods longer than 1 day) at hourly intervals from two superconducting gravimeters at Brussels and Bad Homburg. Series start at 0 h UTC on day indicated and are 120 days long. The upper two series are simultaneous and start near an earthquake in Hindukush (depth 222 km, origin time 23:52:40.5 UTC on Dec. 30, 1983, magnitude $m_b = 6.7$). Note that the upper three series were shifted upward for clarity, but have the same vertical and horizontal scales

properties than just their eigenfrequencies: damping, amplitude distribution on the surface of the globe, phase coherency, compatibility of amplitude with excitation mechanisms. Detection of a spectral peak in one time series should be followed by scrutinous investigation of this time series.

However, the most important corroboration should come from other, distant stations. Figure 4 shows 4 residual gravity time series of 120 days each from the superconducting gravimeters at Brussels and Bad Homburg, Germany at the same scales of amplitude and time. All these residuals were obtained by least-squares fitting of tides and barometric pressure to the gravity data and subtraction of the resulting model from the data (Wenzel 1994). All 4 data sets were treated in the same manner to obtain the residuals. The first series from the top shows the 120 day record after the Hindukush quake of December 30, 1983, from the gravimeter at Brussels resulting in the clearest spectral peak at a period of 13.79 hours as published by Melchior and Ducarme (1986). The distance from Brussels to the site of the next superconducting gravimeter, Bad Homburg, is only 380 km, while the distance from the source of the proposed gravity signal to both stations is at least 2900 km. The second series is the simultaneous record from the gravimeter at Bad Homburg. For the less disturbed second half of this record Zürn et al. (1987) showed that the spectral peak reported for Brussels was not present, Zürn and Rydelek (1994) show that it is not present either in the total 120 hour record. It is important to realize that the larger disturbances in this series are due to identified instrumental problems (Richter pers. comm. 1987). This is verified by the third and fourth records, where rather quiet time series from the Bad Homburg record were selected. The conclusion from this plot can only be that when the peak at Brussels had its best signal-to-noise ratio, the record at Bad Homburg could produce smaller residuals by a factor of at least two, unless there were instrumental disturbances. Bad Homburg was the quieter site, but did not produce these peaks. Therefore they are only in the Brussels record and from the amplitude of the residuals at Brussels one must conclude that either instrumental, local or regional noise was much higher there, not a good condition for detecting something very tiny. One possibility remains: Bad Homburg was situated at a node of this signal, while the amplitude at Brussels was big (see Cummins et al. 1991 for the likelihood of this condition).

Smylie (1992) also made use of the data from both stations above and from the superconducting gravimeter at Strasbourg. Looking at Fig. 4 it appears to be unwise to study a product spectrum of a noisy and a quiet record, it appears to be more prudent to look in the most quiet records only for such tiny signals.

As far as this author is concerned, core modes have not been observed yet. The very strong earthquake in Northern Bolivia at a depth of 641 km on June 9, 1994 (moment magnitude 8.2) surely will initiate new searches for core modes and shed additional light on these dubious claims of excitation to observable amplitudes by quakes of much smaller magnitude.

2.4 Gravity tides and lateral heterogeneities

The tidal variations in gravity can be measured with very high signal-to-noise ratio by superconducting gravimeters and LaCoste-Romberg gravimeters with electrostatic feed-back (Zürn et al. 1991). The response of the solid earth to the tidal driving forces surely contains information about the internal structure of the planet. This response is usually expressed as the gravimetric factor and phase difference for specific tidal frequencies. However, detailed knowledge about this structure exists from other fields, especially from seismology. Considering the difference between tidal and seismic frequencies, the tidal response could shed much light especially on the problem of the rheology of the earth's mantle, if it could be determined with precision from the observations. It has long been known that differences in the spherically symmetric structure of the earth's mantle within the constraints posed by seismological results cause only very small changes in the tidal response of the earth, *i.e.* the gravimetric factor (*e.g.* Varga 1974, Wilhelm 1978, Varga and Denis 1988). The largest deviations from spherical symmetry of the earth are its rotation and flattening. At present the effects of this symmetry breaking on the tidal response are being reinvestigated theoretically (*e.g.* Wang 1994). While a latitude effect in the gravimetric factor of about 3 % (peak-to-peak) due to rotation and ellipticity seemed to be well established by previous theories (Wahr 1981, Dehant 1991), the new results by Wang (1991) indicate a negligible effect and there are indications of an error in the older computations (Dehant pers. comm. 1993). Even less clear from a theoretical point of view are the effects of other structural deviations from spherical symmetry. However, almost all theoretical estimations indicate that for reasonable models of such asymmetries, the effects on tidal gravity are appreciably below 1 % (*e.g.* Zürn et al. 1976, Molodenskii and Kramer 1980, Wang 1991).

Thus theoretically the expected effects from the solid earth are very small (well below 1 %). Two serious problems make it at present impossible to extract these signatures from the observed gravity tides: a lack of absolute calibration accuracy better than a few tenths of a percent of even the best gravimeters and the large size and uncertainties of the indirect effects of the tides in the oceans on the observed earth tides. These problems are documented by the lengthy and complicated argumentation of Melchior (1994) with respect to the necessity of a change of the adopted tidal parameters for Brussels. Several successful efforts have been made to improve the calibration accuracy of recording gravimeters using basic physical principles (Van Ruymbeke 1989, Richter 1991, Varga et al. 1994) and the models of the ocean tides will certainly be improved by the TOPEX/POSEIDON mission results. Figure 5 shows a plot of observed gravimetric factors for M_2 (the tide determined best) for some high quality tidal data recorded in Central Europe against geographical latitude. These observations were obtained by Baker et al. (1989), Wenzel et al. (1991), Rydelek et al. (1991), Timmen and Wenzel (1994), Dittfeld and Wenzel (1993) and and were corrected to the best of present knowledge with models for both the oceanic tidal maps and the loading response of the earth. The error bars contain uncertainties from the noise level and errors of calibration, but no estimate

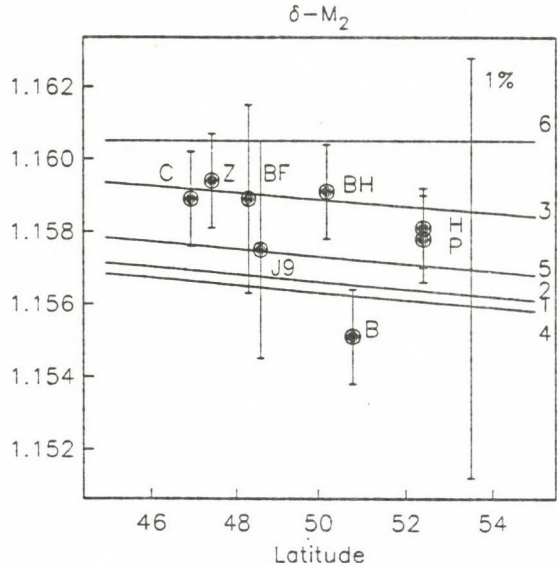

Fig. 5. Gravimetric factors for M_2, corrected for oceanic effects, *vs.* latitude in Central Europe (see Timmen and Wenzel (1994) for similar plots). Stations: C — Chur, Z — Zürich, BF — BFO — Schiltach, J9 — Strasbourg, BH — Bad Homburg, B — Brussels, H — Hannover, P — Potsdam. All stations except J9 (superconducting gravimeter) were occupied with LaCoste-Romberg gravimeters with electrostatic feed-back. Model calculations: 1/2 — Dehant and Zschau (1989), 3 — Dehant and Ducarme (1987), 4/5 — Dehant (1987), 6 — Wang (pers. comm. 1993). Uncertainty estimates contain calibration and noise-level, oceanic correction errors not included

of the uncertainty of the ocean correction is included. The data are compared with 6 different models for the gravimetric factor including latitude dependence of the gravimetric factor. Considering that appreciable additional uncertainties due to the oceanic corrections must exist, then the observations agree within their errors with all the model calculations and with each other. Similar statements can be made for other tides and for the observed phase differences as well. Note that the differences between the observations are well within 1 %. Given the errors a latitude effect cannot be discerned. Actually no other effect can be discerned either, such as an effect of heatflow or lateral heterogeneities. More than 300 tidal measurements have been accumulated in the gravity tide data bank by the International Center for Earth Tides (ICET) (Melchior 1994a). Yanshin et al. (1986) suggested from a correlation between M_2-residuals from this data bank and heat flow estimates at nearby locations that a physical effect exists connecting these two observables. While the statistical correlation exists in these data, Rydelek et al. (1991) questioned this conclusion. Their strongest argument is that the high quality data shown in Fig. 5, when plotted against heat flow show nothing of this sort for Central Europe, although the heat flow values obtained in the area of these sites range between 60 and 120 mW/m^2. In addition, the crustal structures for these stations differ dramatically as well, so there appears to be no discernible effect of local crustal structure

on gravity tide measurements at the present level of accuracy. The scatter of the observed, corrected gravimetric factors in the ICET data bank is much larger than in Fig. 5 (Baker et al. 1989). Rydelek et al. (1991) also point out that as long as ocean loading corrections are computed with a global model only and do not account for all the local seas, appreciable errors must be expected and with prudence no other effects can be expected to show up clearly. Another criticism by Rydelek et al. (1991) concerns the lack of appreciation of error estimates in all the work about correlation of tides with heat flow, notoriously error bars are not presented although estimates for the uncertainties due to noise are readily available from every program for tidal analysis. Of course, these would be minimum estimates because they do not include uncertainties in calibration and oceanic corrections, but they would give the reader an idea about the quality (or lack thereof) of the data. The controversy about this effect is continuing, since Melchior (1994b) responds to the criticism by Rydelek et al. (1991). Inspection of Fig. 3 in Rydelek et al. (1991) almost tells the whole story. Obviously noise levels, calibrations and oceanic corrections must be improved, before effects of lateral heterogeneities become clearly visible even in some of the best results.

2.5 Nearly diurnal free wobble

In the last section it was discussed that the spatial variation of gravimetric factors for earth tides at present does with high probability not contribute to our knowledge about the internal structure of the earth. In contrast, there is evidence that the variation of the gravimetric factors with frequency in the diurnal tidal band tells us something about the interior of our planet which seismology cannot provide. This variation with frequency is caused by resonant behaviour of the earth in the vicinity of 1 cycle per day (cpd) due to the existence of a normal mode of the rotating earth. Fortunately this resonance is sensitive to the tidal forcing for diurnal tides and the properties of the resonance are such that its eigenfrequency is located within this tidal band. In the literature this mode is called the 'Nearly Diurnal Free Wobble' (NDFW), the 'Free Core Nutation' (FCN) and the 'Core Resonance', among others. It is excited, when the rotation axes of mantle and outer core are slightly misaligned. In that case a restoring force is set up at the core-mantle boundary (CMB), which tries to realign the two axes. This restoring force only exists if the CMB is elliptical. Since the earth is spinning fast, the reaction is a damped wobble of the instantaneous rotation axis around the figure axis and a nutation in space of the rotation axis around the axis of total angular momentum. Both motions are two aspects of the same mode (Lambeck 1988). For hydrostatic flattening of the CMB the period of the nutation in space was computed for seismologically constrained earth models to be 466 (Wahr 1981) sidereal days. If anelasticity of the mantle is considered, this period would be lengthened by a few days (Wahr and Bergen 1986).

Measurements of the amplitudes of the forced nutations of the earth using the VLBI-techniques led Gwinn et al. (1986) to suggest a change in the parameters of the FCN from the theoretical values in order to account for anomalies. While the

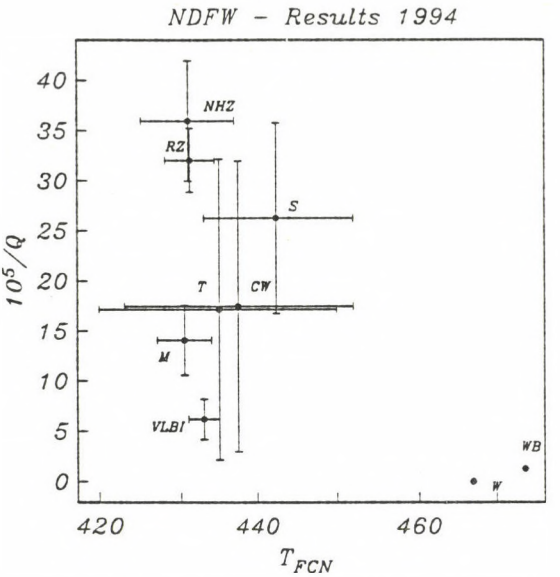

Fig. 6. Estimated parameters ($10^5/Q$ *vs.* T_{FCN} in sidereal days) of the NDFW/FCN eigenmode of the Earth from different authors: W — theory for hydrostatic seismologically constrained earth models (Wahr 1981), WB — effect of mantle anelasticity included (Wahr and Bergen 1986), VLBI — interpretation of VLBI nutation measurements (Gwinn et al. 1986), results from gravity tide measurements (Neuberg et al. 1987 (NHZ), Richter and Zürn 1988 (RZ), Sato et al. 1993 (T), Cummins and Wahr (1993) (CW), Merriam 1994 (M), S — result from tidal strain measurements (Sato 1991). Note that eigenperiod estimates overlap within their uncertainties, while damping estimates do not

quality factor was reasonably high, the free period had to be shifted down by about 30 days to fit the observations.

Independently Neuberg et al. (1987) and Richter and Zürn (1988) found a very similar frequency shift when the resonance parameters were retrieved from gravimetric factors from high quality tidal gravity data recorded in central Europe. Figure 6 shows the quality factors and free core nutation periods obtained by these authors and compares them to theoretical estimates. This is again a nice example for mutual corroboration of unexpected results. Levine et al. (1986) also noted in a gravity tide analysis that the resonance could be located at higher frequencies, without being as specific as the above authors. More corroboration for the frequency shift was provided by Sato (1991) studying strain tides in Japan, by Sato et al. (1993) through the study of gravity tides in Japan using records from three superconducting gravimeters, by Cummins and Wahr (1993) through a sparse global stack of data from the IDA-network (Agnew et al. 1986) and by Merriam (1994) in a study of the data from the Canadian superconducting gravimeter. These results are also shown in Fig. 6. The retrieved eigenperiods agree all nicely within their uncertainties, while the quality factors do not. However, there are several problems inherent in all these analyses. The uncertainties are certainly underestimated due to correlations between the parameters and due to systematic errors in the ocean

loading corrections and the modeling of the gravity effects of the atmosphere. Especially the latter problem is an extreme complication, because the effects of the atmosphere are complex right at those tides which must be used for the resonance analysis (*e.g.* Merriam 1994). For this reason the discrepancies between the quality factors Q of the individual estimates should not be taken too seriously. A global (sparse) network of superconducting gravimeters, maybe with some help of very few LaCoste-Romberg gravimeters with electrostatic feed-back (Zürn et al. 1991), is desperately needed to resolve this discrepancy and to solidify the results on the eigenperiod and resonance strength (two additional parameters obtained in this kind of analysis).

Gwinn et al. (1986) suggested that the significant shift of the eigenperiod to shorter values is caused by an increased ellipticity of the CMB by a couple of 100 meters. Several physical mechanisms to explain this shift and the observed Q value were discussed by Neuberg et al. (1990): their conclusion supports the conjecture of Gwinn et al. (1986). An increase of flattening between 250 and 350 meters suffices to explain the observation. Seismological methods probably can never reach this kind of resolution at the CMB, because the shortest observable frequencies have wavelengths there at least an order of magnitude larger. The study of the NDFW, however, provides only information on the ellipticity of the CMB and not about CMB-topography in general. Hinderer et al. (1991) verified that the observed real part of the resonance strength in tidal gravity is consistent with present seismological earth models. An interpretation of the Q factor has to await more precise estimates, the same is true for the imaginary part of the resonance strength.

2.6 Tidal tilts and strains

An excellent introduction into this field and several benchmark papers can be found in Harrison (1985). The rapidly developing field of tides in aquifers will not be discussed here (*e.g.* Rojstaczer and Agnew 1989). It was stated above that the distortions of the tidal gravity field by local, regional and even large scale deviations of the earth from spherical symmetry are very small and hardly observable at present. In contrast, the tidal deformation field, as measured by tilt- and strainmeters, is heavily distorted by these effects to the extent that the global response cannot be measured reliably with such instruments. This situation and its consequences for tidal research are well documented by Harrison (1985, part IV), who also did a good part of the pioneering work towards a better understanding of these problems. The distortions are called cavity, topographic and geological effects in the literature. Sato and Harrison (1990) made the most recent attempt of interpretation of anomalous tidal strains by taking carefully all the known effects into account. Actually, it is better to turn around and look for possibilities where the geological effects could tell us something new about the local or regional structure of the earth. This means that in an experiment the geological effects should be dominated by the object under study, and cavity and topographic effects should be minimized. For the interpretation models of the body tide (the global response of

the earth to the tidal forcing) and the ocean tide effect are assumed to be much better known than the other distortions and are therefore subtracted from the observations. Cavity effects cannot be avoided by measurements in mines, and they are mostly stronger in instruments with small baselines, therefore the experimental answer is to put small instruments into boreholes and/or to increase the baseline (Zürn et al. 1986, Emter et al. 1989). After this problem was understood, several attempts were undertaken to use arrays of borehole tiltmeters in geologically interesting situations in order to use the spatial (amplitude and phase) variation of observed tidal tilts for the geological studies. However, other researchers tried to find out if it is really possible to get consistent such tidal measurements in places where the rocks were assumed to be very homogeneous and no spatial variation of tidal tilts or strains should be observed. In the following some of this work is referenced. An important theoretical prediction was published by Beaumont and Berger (1976). These authors showed with the use of finite element models that rather large and easily observable temporal modification of tidal tilt- and strain amplitudes could occur in an earthquake zone, if it becomes dilatant (or its elastic parameters change for other reasons) during the time when crustal stresses built up before an earthquake. In essence this is also a positive outcome of the research on the distortion effects.

Three experiments using relatively small arrays of borehole tiltmeters were designed to test if consistent tidal signals within a supposedly homogeneous volume in the earth's crust would be recorded. Cabaniss (1978) had installed two arrays of three tiltmeters each in New England. For the shallow (18 m) array with borehole separations of 100 m he reports agreement of the tidal responses within 2 %, while for the deep array (100 m) with similar spacing instrumental problems obviously contaminated the tidal results. The tidal results reported by Peters and Beaumont (1987) were obtained from three boreholes in the Charlevoix earthquake zone in Quebec, Canada. The boreholes with depths of 47, 47 and 110 m were forming a triangle of approximately 80 m sides and were housing Askania tiltmeters. Unfortunately the results from the 110 - m borehole suffer from the fact that the orientation of the tiltmeter at the bottom of this hole was not known with sufficient precision, therefore the orientation had to be deduced from the tidal responses of the two components. One important but very disappointing result from these three boreholes was the dispersion of the observed tidal admittances (up to 20 % in the amplitude of the M_2 tide), which must be blamed on unexpected strain - tilt coupling by heterogeneities (fractures ?) in the vicinity of the boreholes.

At the Pinyon Flat Observatory (PFO) in Southern California an important set of experiments is carried out concerning the significance of observed tilts and strains for the state of crust not very far from two major faults (one of them the famous San Andreas). Agnew (1981) used the strain tides observed with three laser strainmeters (appr. 700–800 m long) to look for a nonlinear response of the rocks to the earth tides with a negative result. Wyatt and Berger (1980) and Wyatt et al. (1982, 1984, 1987, 1988) compared tilts (tides and other frequencies) from several installations (surface long-baseline and shallow and intermediate depth borehole) with each other and found significant discrepancies, which in several cases were

interpreted to be due to insufficiently well-known instrumental properties. Johnson et al. (1994) concentrate on the tidal results from one Askania borehole tiltmeter (this instrument is described by Agnew 1986) at PFO installed successively in two neighbouring boreholes with different depths. Here also it was found that the results differed significantly from each other. Kohl and Levine (1992 1994) studied the tidal tilts from another set of borehole tiltmeters in the neighbourhood of the Askania and observed rather large discrepancies between their installations. This disturbing result triggered their interesting theoretical investigation of cavity effects in borehole installations and they could correct some of their results to get better agreement. The result of all this research is that tidal amplitudes (and phases) obtained with the same techniques in supposedly homogeneous blocks of the earth's crust at close distance can differ appreciably due to effects from the installation itself or from the immediate neighbourhood of the borehole. The results from experiments designed to study a larger heterogeneity of geological interest must therefore be interpreted very cautiously.

Edge et al. (1981) used a borehole tiltmeter array to study the effect of a known batholith in the Lake District of England. They found remarkable coherency between tidal results from the different installations, at least within their uncertainties of about 3 %. A detailed analysis of the tidal results is not available. Gerstenecker et al. (1986) operated borehole tiltmeters along a profile across a deep-reaching (30 km) fault zone in the Rhenish Massif in Germany to see if this zone would produce anomalous tidal tilts in its vicinity. Figure 7 shows some of their results, which are the largest tidal tilt anomalies ever reported from borehole installations. In one case the amplitude of the anomaly reaches 100 % of the body tide amplitude. Several models of this fault zone were devised by the experimenters, but due to the large effect these models were rather exotic and somewhat improbable. Of course, and unfortunately, small scale heterogeneities in the immediate neighbourhood of the borehole can easily produce such large effects and therefore one must always worry about an interpretation. Meertens et al. (1989) used the same technique to study the tidal tilts in the Yellowstone, USA Caldera in order to find information on the structure beneath it. They found effects as large as 50 % of the expected signal from the body tide and ocean loading effects. Their models for a low velocity body underneath the caldera notoriously underpredict the effects. Since the modeling is much constrained by other geophysical data, the authors conclude that possibly a more complex or more rapid spatial variation of elastic parameters is sensed by the tidal tilts. This result is very similar to that of Gerstenecker et al. (1986) and basically not surprising considering the results obtained in mine installations and the many models of heterogeneities studied in that context. In summary, tilt- and strain tides are very sensitive to local structure, while for gravity tides large scale (global) structure dominates (Molodenskii and Kramer 1980).

Occasionally amplitude variations of tidal strains and tilts with time were reported in the literature in connection with earthquakes (*e.g.* Latynina and Rizaeva 1976). Some of these claims were later retracted, because instrumental problems were detected afterwards. Some other claims must be considered dubious because the long term stability of the tidal signals at times without quakes was not demon-

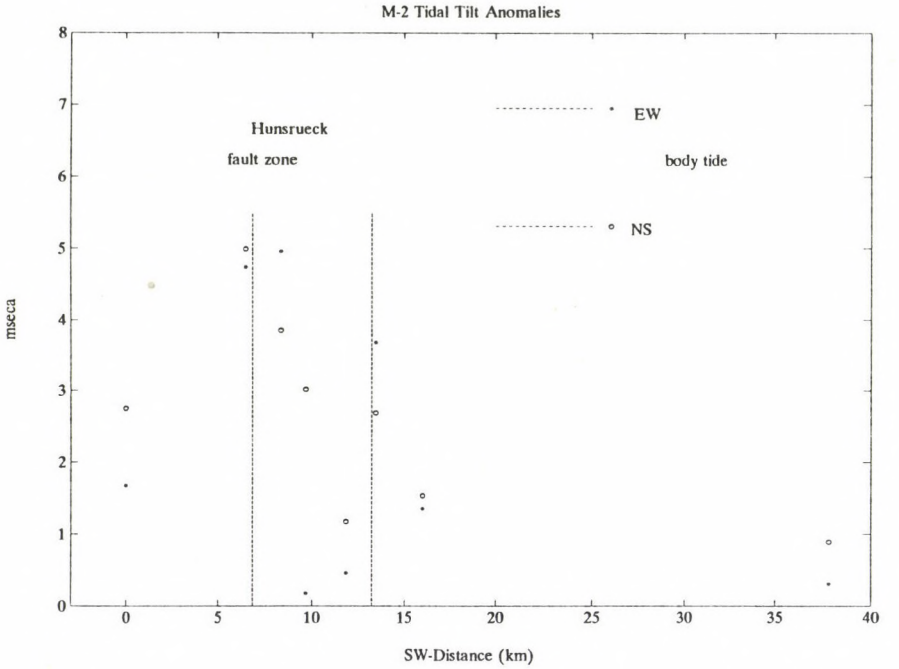

Fig. 7. Amplitude anomalies of tidal NS- and EW-tilts (in mseca) for M_2, observed along a NE–SW profile across the Hunsrück fault zones in Germany. The anomalies plotted are amplitudes of residual tidal vectors after subtraction of models for the body tide, the ocean loading effect and the topography. These amplitudes would be zero if no other physical mechanism is involved (and if corrections are perfect). Body tide amplitudes are given at top right for reference. Note that anomaly reaches almost 100 % of body tide amplitude for NS-tilt at the station with a distance of 7 km

strated. Often the effect is seen only on one instrument, which then could have a very local cause.

The most intensive search for temporal variations of tidal tilts is carried out with several Askania borehole tiltmeters installed in the North Anatolian fault zone. The Turkish-German earthquake prediction research program carries out multidisciplinary studies towards the western end of this very active fault. Westerhaus et al. (1991) report about correlations between microseismicity in this area and the tidal response at three stations. No big quake has occurred in the region under study, but the program is still going on. This application of earth tides as a probing signal for the rocks is rather interesting, but so far the experimental results are inconclusive. Hopefully the project in Turkey will settle this question when the next big quake in that area will strike.

2.7 Transients and steps in tilt and strain

In this section only a few remarks can be made about steps or pulses in tilt and/or strain with durations from seconds to several days. Here single-instrument

observations must be interpreted with great care and corroboration by other instruments and/or stations may be needed more urgently than with the signals discussed so far. The coupling of strain- and tiltmeters to the surrounding rocks is an important part of the installation (Agnew 1986, Wyatt et al. 1982) and can cause severe problems. The best high technology instrument cannot produce good results if it is not stably coupled to the earth's crust.

Meteorologically and/or hydrologically caused tilts and strains have been observed very frequently in the form of transient (but also secular) tilts and strains (*e.g.* Berger 1975, Edge et al. 1981, Müller and Zürn 1983, Evans and Wyatt 1984, Kümpel and Lohr 1985, Rabbel and Zschau 1985, Van Dam and Wahr 1987, Weise 1992) and will not be discussed here. In the widest sense these are also geodynamical signals, which sometimes can be exploited to study the solid earth.

Coseismic steps in tilt and strain are difficult to observe with reliability, because they occur during the passage of the seismic waves through the station. The violent shaking associated with strong quakes can cause instrumental (including the coupling to the rocks) disturbances which of course should not be interpreted in terms of the focal mechanism of the corresponding earthquake. Therefore it is very important here to have results consistent between different instruments and/or stations and consistent with models of the earthquake source derived from its seismic radiation. Wyatt (1988) and Agnew and Wyatt (1989) report about coseismic steps observed at PFO from seismic sources located at distances of about 90 km with magnitudes M_S of 6.2 and 6.6. These steps were observed in a consistent way by the laser strainmeters, a volumetric strainmeter and a long-baseline tiltmeter and were consistent in magnitude with models of the focal mechanisms. The steps were of the order of $3 \cdot 10^{-9}$ and $17 \cdot 10^{-9}$, respectively for the 6.2 and 6.6 quakes. There also was some indication of post-seismic creep a factor of 20 less in magnitude.

However, an even better opportunity for PFO arose, when the Landers quake ($M_S = 7.4$) struck on June 28, 1992 about 66 km from the observatory. Coseismic strain and tilt steps were clearly observed, of course ($1.7 \cdot 10^{-6}$ in strain). Rapid post-seismic changes were recorded on all of the long-baseline strain- and tiltmeters and on a number of short-base sensors (Agnew and Wyatt 1993). The NW-SE laser strainmeter showed the largest post-seismic creep ever observed of $120 \cdot 10^{-9}$. The signal looked roughly exponential with a time constant of about a week. The results in this case were not consistent between different sensors, though, for a large scale interpretation to hold. Therefore the authors favor a creep of near-surface local rocks in response to the stress changes caused by the earthquake. So again it appears that large scale deformation fields are locally distorted and affected, and tilt- and strainmeters are very sensitive to these distortions, just like in the case of the earth tides.

No preseismic transients were ever detected at PFO. On the other hand changes in tilt and strain before earthquakes were often reported (*e.g.* Wyss 1975) and belong to the more generally accepted possible precursors, however, they have not proven to be reliable enough for earthquake prediction (like every other precursory phenomenon so far). Again, one single observation alone should never be used for claiming detection of a precursor.

Some well monitored volcanoes, *e.g.* Kilauea (Hawaii) (Klein 1984) and Sakurajima (Japan) (Kamo and Ishihara 1989) among others are known to inflate before and deflate during an eruption, so among other phenomena, tilts can be observed and employed for eruption prediction on a few special volcanos. A very fine experiment was performed by Linde et al. (1993) at Hekla, Iceland. Five Sacks-Evertson borehole dilatometers (Agnew 1986) were installed at distances of 15 to 45 km from the main vent of the volcano and were recording continuously before and during the eruption in January 1991. The four most distant instruments recorded contractions only with magnitudes between 40 and $200 \cdot 10^{-9}$, while the closest device first showed a fast dilatation of $650 \cdot 10^{-9}$ and then returned to the pre-eruption level. The authors could quantitatively model the observations with a 6.5 km deep spherical magma reservoir with a radius of 2.5 km and a dike above it with a length of 4 km, a depth of 4km and a width of 0.85 m. Stage 1 consists of dike formation and deflation of the reservoir, while stage 2 consists, after the magma broke the surface at the end of stage 1, only of deflation of the reservoir. The closest station sees both the deep reservoir and the shallow dike, and thus experiences slow dilatation superimposed on quick compression caused by these two sources, respectively. The stations farther away only see the effects of the deep source. With a pressure change in the deep reservoir of 14 MPa at the end of stage 1 the model calculations fit the observations very netly. This analysis demonstrates the potential of arrays of tilt- and/or strainmeters for the study of geodynamic phenomena again.

2.8 Secular strains and tilts

It is well known that metal or quartz springs in gravity meters drift with rates corresponding to about 1 to 5 μgals/d. Many of them drift in the same direction, *i.e.* the springs become longer due to creep in the material. A careless interpreter (no expert) could conclude that this drift is due to the free air anomaly associated with changing elevation and could estimate that elevation changes of about 3 to 15 mm/d occur. However, in one year the total elevation change would be 1000 to 5000 mm, and it is clear immediately that this was a bad interpretation. Emter et al. (1989) observed a drift rate of 3 μrad/a with their long-baseline fluid tiltmeter. A regional interpretation of this number would mean that a baseline of 100 km would in one year be lifted at one end by 0.3 m. This is a result which can easily be checked by classical geodetic methods. Clearly, instrumental (or coupling) problems must be the reason for these drifts. There are many possible sources for such drifts in mechanical instruments. These two silly examples are meant to demonstrate that such secular signals or drifts can sometimes be judged by simple extrapolations in time and space and one always has to consider if the accumulated effect over a certain period of time can be verified by geodesists.

From the theoretical side other constraints can be found on the back of an envelope. Varga (1984) showed in a short, but enlightening paper that given a stress drop in the fault volume $p = 10^7$ N/m^2, a time interval of 100 a between large earthquakes and a shear modulus of the rocks of $3 \cdot 10^{10}$ N/m^2 the strain accumulation rate would be $3 \cdot 10^{-6}$/a or 3 mm in 1 km per year. This is an upper

bound on the strain rate to be expected (similar numbers can be found for tilt rates) because it would be inside the fault volume, so this rate would drop quickly with distance from the fault. Tilt and strain rates of this order of magnitude are therefore suspect. In addition, they can be detected by geodetic methods if one waits a while. Varga (1984) also compiled secular strain rates from local observations from the literature.

More effort than anywhere else to study this problem of how to determine crustal tilt and strain rates reliably is spent at the PFO in Southern California, at a place only 25 km west of the San Andreas fault and 12 km east of the San Jacinto fault. There were many different sensors deployed to measure tilt and strain and all these local measurements are constrained by geodetic surveys. The experimenters conclude from their concentrated effort with all these techniques, that tectonic strain and tilt rates at this special location are less than or equal to $0.05 \cdot 10^{-6}/a$ and $0.085 \cdot 10^{-6}/a$, respectively (Wyatt et al. 1990). The first result is derived from the 730 m long laser strainmeters (with optical anchors going down 26 m) from data of several years. Data were corrected for monument motion with the optical anchor records and changes of laser frequency measured independently. The resulting strain rate is consistent with geodetic measurements within their errors. The second result is derived from the 530 m long-baseline (half filled) water-tube tiltmeter of the University of California at San Diego (Agnew 1986). This instrument is installed at a depth of 1.2 m and fixed to the rocks underneath by optical anchors with lengths of 26 m. All other instruments at PFO show larger drift rates. The experimenters make some careful statements about the interpretation of these small rates: they are not sure if they have eliminated all instrumental (and coupling) contributions to these rates.

The work by Wyatt and colleagues are in the author's opinion the most careful and scrutinous efforts to determine tectonic tilt and strain accumulation. Every other research on this problem should be regarded objectively and honestly in the light of their efforts.

Acknowledgements

Special thanks go to P Varga for the invitation to lecture at the Sopron meeting in February 1994 and to him and the Hungarian colleagues for their hospitality. B Ducarme, P Melchior, B Richter and P Wolf kindly provided data from the superconducting gravimeters at Brussels and Bad Homburg. A copy of his ETERNA code for tidal analysis was provided by H-G Wenzel and D Emter critically read the manuscript. The data from BFO would not be what they are without the help of D Emter, W Grossmann, W Kaminski, H Otto and H-G Wenzel. Unpublished information was obtained from D C Agnew, H O Johnson and F Wyatt from PFO. This work was financially supported by *Deutsche Forschungsgemeinschaft* within the Sonderforschungsbereich 108: "Stress and Stress Release in the Lithosphere", Projekt D9. All this is gratefully acknowledged.

References

Agnew D C 1981: *J. Geophys. Res.*, 86, B5, 3969–3978.

Agnew D C 1986: *Rev. Geophysics*, 24, 579–624.

Agnew D C, Wyatt F K 1989: *Bull. seismol. Soc. Am.*, 79, 480–492.

Agnew D C, Wyatt F K 1993: Postseismic Strains and Tilts from the Landers Earthquakes: Fault Afterslip, Crustal Anelasticity or Local Hydrology. EOS (Trans. Am. Geophys. Union) 74, 43 (Suppl.), 184

Agnew D C, Berger J, Farrell W E, Gilbert J F, Masters G, Miller D 1986 : Project IDA: A Decade in Review. EOS (Trans. AGU) 67, 16, 203–211.

Baker T F, Edge R J, Jeffries G 1989: *Geophys. Res. Lett.*, 16, 1109–1112.

Beaumont C, Berger J 1976: *Geophys. J. R. astr. Soc.*, 39, 111–118.

Berger J 1975: *J. Geophys. Res.*, 80, 274–277.

Bolt B A, Tanimoto T 1981: Atmospheric Oscillations after the May 18, 1980 Eruption of Mount St. Helens. EOS (Trans. Am. Geophys. Union) 62, 529–530.

Bullen K E 1965: An Introduction to the Theory of Seismology, 3rd edition. Cambridge University Press, Cambridge

Busse F 1974: *J. Geophys. Res.*, 81, 3055–3065.

Cabaniss G H 1978: In: Applications of Geodesy to Geodynamics, I Mueller ed., Ohio State University Press, Columbus, Ohio, USA, 165–169.

Crossley D J, Rochester M G 1980: *Geophys. J. R. astr. Soc.*, 60, 129–161.

Crossley D J, Rochester M G, Peng Z R 1992: *Geophys. Res. Lett.*, 19, 1679–1682.

Cummins P R, Wahr J M 1993: *J. Geophys. Res.*, 98, 2091–2104.

Cummins P R, Wahr J M, Agnew D C, Tamura Y 1991: *Geophys. J. Int.*, 106, 189–198.

Dehant V 1987: *Phys. Earth Planet. Inter.*, 49, 242–258.

Dehant V 1991: *J. Geophys. Res.*, 96, 20235–20240.

Dehant V, Ducarme B, 1987: *Phys. Earth Planet. Inter.*, 49, 192–212.

Dehant V, Zschau J 1989: *Geophys. J. Int.*, 97, 549–556.

Derr J S 1969: *Bull. seism. Soc. Am.*, 59, 2079–2099.

Dittfeld H-J, Wenzel H-G 1993: Joint Gravity Tide Recording at Potsdam. IAG-Symp. 113: Geodesy Phys. Earth, Springer, Berlin, 186–189.

Edge R J, Baker T F, Jeffries G 1981: *Tectonophysics*, 71, 97–109.

Emter D, Zürn W, Mälzer H 1989: Underground Measurements at Tidal Sensitivity with a Long Baseline Differential Fluid Pressure Tiltmeter. Deutsche Geodät. Komm., B 288, Munich, 1–74.

Evans K, Wyatt F 1984: *Tectonophysics*, 108, 323–337.

Georges T M 1973: *Rev. Geophys. Space Phys.*, 11, 3, 571–594.

Gerstenecker C, Zschau J, Bonatz M 1986: In: Proc. 10th Int. Symp. Earth Tides, R Vieira ed., Cons. Sup. Invest. Cient., Madrid, 797–804.

Gilbert F, Dziewonski A M 1975: *Phil. Trans. R. Soc. Lond.*, A278, 187–269.

Gossard E E, Hooke W H 1975: Waves in the Atmosphere. Elsevier Scient. Publish. Comp., Amsterdam

Goodkind J M 1991: The Superconducting Gravimeters: Principle of Operation, Current Performance and Future Prospects. Cahiers Centre Europ. Geodyn. Seism., 3, 81–90.

Gwinn C, Herring T A, Shapiro I I 1986: *J. Geophys. Res.*, 91, 4755–4765.

Harrison J C ed. 1985: Earth Tides. Benchmark Papers in Geology Series, Van Nostrand Reinhold, New York

Hinderer J, Zürn W, Legros H 1991: In: Proc. 11th Int. Symp. Earth Tides, Helsinki 1989, J Kakkuri ed., Schweizerbart, Stuttgart, 549–555.

Johnson H O, Wyatt F, Agnew D C, Zürn W 1994: In: Proc. 12th Int. Symp. Earth Tides, Beijing 1993, H-T Hsu ed. (in press)
Kamo K, Ishihara K 1989: A Preliminary Experiment on Automatic Judgment of the Stages of Eruptive Activity Using Tiltmeter Records at Sakurajima, Japan. IAVCEI Proc. Volcanol., 1, Volcanic Hazards, J H Latter ed., Springer, Heidelberg, 585–598.
Kanamori H, Mori J 1992: *Geophys. Res. Lett.*, 19, 7, 721–724.
Kanamori H, Mori J, Harkrider D G 1992: Excitation of Atmospheric Oscillations by Volcanic Eruptions (Abstract). Trans. Am. Geophys. Union (EOS), 43, Suppl., 634.
King G C P, Bilham R 1976: *Bull. seism. Soc. Am.*, 66, 2039–2047.
Klein F W 1984: *J. Geophys. Res.*, 89, 3059–3073.
Kohl M, Levine J 1992: Using a Short Baseline Borehole Tiltmeter to Measure the Regional Tilt EOS (Trans. Am. Geophys. Union) 73, 43 (Suppl.), 121.
Kohl M, Levine J 1994: *J. Geophys. Res.* (in press)
Kümpel H-J, Lohr G 1985: *J. Geophysics*, 57, 184–190.
Lambeck K 1988: Geophysical Geodesy: The Slow Deformations of the Earth. Clarendon Press, Oxford
Latynina L A, Rizaeva S D 1976: *Tectonophysics*, 31, 121–127.
Levine J, Harrison J C, Dewhurst W 1986: *J. Geophys. Res.*, 91, 12835–12841.
Linde A T, Agustsson K, Sacks I S, Stefansson R 1993: *Nature*, 365, 737–740.
Melchior P 1994a: *Phys. Earth Planet. Inter.*, 82, 125–155.
Melchior P 1994b: *Phys. Earth Planet. Inter.* (in press)
Melchior P, Ducarme B 1986: *Phys. Earth Planet. Inter.*, 42, 129–134.
Merriam J 1994: *J. Geophys. Res.* (in press)
Meertens C, Levine J, Busby R 1989: *J. Geophys. Res.*, 94, B1, 587–601.
Müller T, Zürn W 1983: *J. Geophysics*, 53, 155–162.
Molodenskii S M, Kramer M V 1980: *Bull. Acad. Sci. U.S.S.R., Earth Phys.*, 16, 1–11.
Neuberg J, Hinderer J, Zürn W 1987: *Geophys. J. R. astr. Soc.*, 91, 853–868.
Neuberg J, Hinderer J, Zürn W 1990: In: Variations in Earth's Rotation, D C McCarthy, W E Carter eds, Geophys. Monogr. 59, IUGG Vol. 9, AGU and IUGG, Washington, D.C. 11–16.
Park J, Lindberg C R, Thomson D 1987: *Geophys. J. R. astr. Soc.*, 91, 755–794.
Pekeris C L 1939: *Proc. Roy. Soc. A.*, 171, 434–451.
Pekeris C L 1948: *Phys. Rev.*, 73, 2, 145–154.
Peters J, Beaumont C 1987: *Earth Planet. Sci. Lett.*, 84, 263–276.
Rabbel W, Zschau J 1985: *J. Geophysics*, 56, 81–99.
Richter B 1991: *Cahiers Centre Europ. Geodyn. Seism.*, 3, 99–107.
Richter B, Zürn W 1988: In: The Earth's Rotation and Reference Frames for Geodesy and Geodynamics, A Babcock, G Wilkins eds, Kluwer, Dordrecht, Holland, 309–315.
Rojstaczer S, Agnew D C 1989: *J. Geophys. Res.*, 94, 12403–12411.
Rydelek P A, Zürn W, Hinderer J 1991: *Phys. Earth Planet. Inter.*, 68, 215–229.
Sato T 1991: In: Proc. 11th Int. Symp. Earth Tides, Helsinki 1989, J Kakkuri ed., Schweizerbart, Stuttgart, 573–582.
Sato T, Harrison J C 1990: *Geophys. J. Int.*, 102, 3, 513–526.
Sato T, Tamura Y, Higashi T, Takemoto S, Nakagawa I, Morimoto N, Fukuda Y, Segawa J, Seama N 1993: *J. Geomagn. Geoelectr.* (in press)
Slichter L B 1961: *Proc. Natl. Ac. Sc. U. S.*, 47, 186–190.
Smith S W 1961: An Investigation of the Earth's Free Oscillations. PhD thesis, California Institute of Technology
Smith M L 1976: *J. Geophys. Res.*, 81, 3055–3065.

Smylie D E 1992: *Science*, 255, 1678–1682.
Timmen L, Wenzel H-G 1994: *Bull. Inf. Marees Terrestres*, 119, 8834–8846.
Torge W 1989: Gravimetry. De Gruyter, Berlin
Varga P 1974: *Pure Appl. Geophys.*, 112, 777–785.
Varga P 1984: *J. Geophysics*, 55, 68–70.
Varga P, Denis C 1988: *Geophys. Trans.*, 34, 263–282.
Varga P, Hajosy A, Csapó G 1994: *Geophys. J. Int.* (in press)
Van Dam T, Wahr J M 1987: *J. Geophys. Res.*, 92, 1281–1286.
Van Ruymbeke M 1989: *Bull. Geod.*, 63, 223–235.
Wahr J M 1981: *Geophys. J. R. astr. Soc.*, 64, 677–703.
Wahr J M, Bergen Z 1986: *Geophys. J. R. astr. Soc.*, 87, 633–668.
Wang R 1991: Tidal Deformations on a Rotating, Spherically Asymmetric, Visco-elastic and Laterally Heterogeneous Earth. Peter Lang, Frankfurt
Wang R 1994: *Geophys. J. Int.*, 117, 562–565.
Weise A 1992: Neigungsmessungen in der Geodynamik - Ergebnisse von der 3-Komponenten-Station Metsähovi. Dissertation, Techn. Univ. Clausthal, 181 pp.
Wenzel H-G 1994: *Bull. Inf. Marees Terrestres*, 118, 8719–8721.
Wenzel H-G, Zürn W, Baker T F 1991: *Bull. Inf. Marees Terrestres*, 109, 7849–7863.
Westerhaus M, Welle W, Büyükköse N, Zschau J, 1991: In: Proc. Int. Conf. Earthquake Predict., Strasbourg, 272–281.
Widmer R, Zürn W 1992: *Geophys. Res. Lett.*, 19, 8, 765–768.
Widmer R, Zürn W, Masters G 1992: *Geophys. J. Int.*, 111, 226–236.
Wielandt E, Streckeisen G 1982: *Bull. seism. Soc. Am.*, 72, 2349–2368.
Wilhelm H 1978: *J. Geophys.*, 44, 435–440.
Wyatt F 1988: *J. Geophys. Res.*, 93, 7923–7942.
Wyatt F, Berger J 1980: *J. Geophys. Res.*, 85, B8, 4351–4362.
Wyatt F, Beckstrom K, Berger J 1982a: *Bull. Seismol. Soc. Am.*, 72, 5, 1707–1716.
Wyatt F, Cabaniss G, Agnew D C 1982b: *Geophys. Res. Lett.*, 9, 7, 743–746.
Wyatt F, Bilham R, Beavan J, Sylvester A G, Owen T, Harvey A, Macdonald C, Jackson D D, Agnew D C 1984: *Geophys. Res. Lett.*, 11, 10, 963–966.
Wyatt F, Levine J, Agnew D C, Zürn W 1987: Side-by-Side Tidal Tilt Measurements - Some Disturbing Results. EOS (Trans. Am. Geophys. Union) 68, 44, 1247
Wyatt F K, Morrissey S-T, Agnew D C 1988: *J. Geophys. Res.*, 93, B8, 9127–9201.
Wyatt F, Agnew D C, Johnson H O 1990: Pinyon Flat Observatory: Comparative Studies and Geophysical Investigations. U. S. Geol. Survey, open-file report 90-380, 242–248.
Wyss M ed. 1975: Earthquake Prediction and Rock Mechanics. Birkhäuser, Basel
Zürn W, Rydelek P A 1994: *Surveys in Geophysics*, 15, 409–431.
Zürn W, Widmer R 1994: In: USGS Monograph on the June 15, 1991 Eruption of Mount Pinatubo, C Newhall and R Punongbayan eds (in press)
Zürn W, Beaumont C, Slichter L B 1976: *J. Geophys. Res.*, 81, 4923–4932.
Zürn W, Emter D, Heil E, Neuberg J, Grüninger W 1986: In: Proc. 10th Int. Symp. Earth Tides, Madrid (1985), R Vieira (ed.), Cons. Sup. Invest. Cient., Madrid, 61–69.
Zürn W, Richter B, Rydelek P A, Neuberg J 1987: *Phys. Earth Planet. Inter.*, 49, 176–178.
Zürn W, Wenzel H-G, Laske G 1991: *Bull. Inf. Marees Terrestres*, 110, 7940–7952.

Acta Geod. Geoph. Hung., Vol. 29(3–4), pp. 363–387 (1994)

GRAVITY FIELD RECOVERY FROM GPS

P N A M VISSER[1], E T HESPER[1], B A C AMBROSIUS[1]

In the first part of these lecture notes, the results of a study to the possibility of both a local and global gravity field recovery from GPS SST range measurements to a low Earth satellite equipped with a high-quality dual-frequency GPS receiver are described. In the second part some results of the TOPEX/Poseidon orbit determination from GPS SST tracking measurements are discussed.

It is shown in Part I by both global gravity field recovery analyses and a local gravity field recovery experiment that it will be possible to significantly increase our knowledge of the Earth's gravity field, and to improve the state-of-the-art in gravity field modeling if future low Earth satellites will be equipped with a high-quality dual-frequency GPS receiver.

Part II discusses the applicability of the Global Positioning System (GPS) for low earth-orbiting spacecraft positioning on basis of the processing of data obtained by the TOPEX/Poseidon satellite and a set of ground stations, distributed uniformly around the globe, during the period January 30 – February 8, 1993 (cycle 14). The results of basically two estimation strategies that have been investigated are described: (i) a dynamic technique, which relies on the accuracy of the dynamic models, characterized by the fact that the transition of the satellite state at different observation times is accomplished by the integration of the equations of motion, and (ii) a reduced-dynamic technique, which minimizes requirements for precision dynamic models and takes full advantage of the geometric information content available from the GPS measurements, characterized by the fact that the satellite state transition at different observation times is accomplished by both the integration of the equations of motion and the satellite positional change inferred from continuous GPS carrier phase measurements. In addition, the effect of the errors in the precise GPS ephemerides on the orbit accuracy of TOPEX/Poseidon was also examined. For this purpose, the GPS orbits of the International GPS Service for Geodynamics (IGS) were used with a claimed overall RMS orbit accuracy of 30 cm. Results show that the postfit residuals of the first-order ionospheric-free carrier phase and pseudorange measurements of TOPEX/Poseidon are less than 1/0.5 cm and 75/68 cm in an RMS sense for the dynamic/reduced-dynamic technique. With the dynamic technique, it is shown that the TOPEX/Poseidon RMS position difference over a 6-hour overlap between two 30-hour contiguous data arcs is 1.9 cm in radial, 4.5 cm in cross-track and 6.1 cm in along-track direction. With the reduced-dynamic technique, these values are 0.9 cm, 1.8 cm and 3.3 cm, respectively. Fixing the GPS orbits imposes an additional RMS orbit error of about 0.9 cm in radial, 3.0 cm in cross-track and 2.6 cm in along-track direction for both strategies.

Keywords: collocation; dynamic; geoid undulation; GPS; gravity anomaly; gravity field; least-squares; linear perturbation theory; orbit error; reduced-dynamic

[1]Delft University of Technology, Faculty of Aerospace Engineering, Section Space Research and Technology, Kluyverweg 1, 2629 HS Delft, The Netherlands

1217-8977/94/$ 4.00

Part I. Global and local gravity field recovery from GPS

Introduction

With the coming of the space age, technologies have been developed that enabled a systematic observation of the entire Earth. It is being realized that in order to be able to understand the complex nature of the dynamic Earth, a multidisciplinary approach is required in which the mapping of the Earth's gravity and magnetic fields with high resolution and accuracy is a prerequisite (Lambeck and Aristoteles 1990). High-resolution models of the gravity and magnetic fields of the Earth will help in modeling and understanding her structure and the driving forces behind plate tectonics, mantle convection, lithospheric motions, *etc.* (Gravity Workshop 1987). In addition, a high-resolution gravity field model will help to establish a physically meaningful reference surface for the oceans, in this case the geoid. With the extensive altimeter data sets of past missions like GEOS-3, SEASAT and GEOSAT, current missions, TOPEX and ERS-1, and future altimeter missions, ocean variations with respect to this surface can be studied on different geometrical and temporal scales. In addition, by combining altimetry and gravity, ocean currents can be deduced and possibly long-term effects like global sea level change can be studied.

In the first part of these lecture notes the recovery of gravity field information from GPS (Global Positioning System), SST (Satellite-to-Satellite Tracking) measurements to a low Earth satellite will be addressed. At least four areas in Earth sciences and applications will greatly benefit from a vast improvement in our knowledge of the gravity field: 1. geodynamics, 2. oceanography, 3. climate and sea level change studies, and 4. geodesy, orbit mechanics and navigation. In Fig. 1 (Gravity Workshop 1987) the state-of-the-art in gravity field modeling is represented by the JGM-2 gravity field model, which is a spherical harmonic expansion complete to degree and order 70. It can be seen that this model is insufficient to answer many open questions. The ability to accurately model the shorter-wavelength geoid undulations will be of major importance, as it will facilitate for the first time the detection of oceanic features with wavelengths shorter than the dynamic sea surface topography that can be recovered nowadays from satellite altimetry (Denker and Rapp 1990, Engelis and Knudsen 1989, Marsh et al. 1989b, Nerem et al. 1988, Visser 1992, Visser et al. 1993).

A high-quality dual-frequency receiver on board of a low Earth satellite will yield SST pseudo-range and carrier phase measurements. A pseudo-range measurement is obtained by measuring the transit time of coded radio-frequency signals transmitted by the GPS satellites and recorded by a GPS receiver, and by multiplying this transit time with the speed of light (Ambrosius et al. 1990). There are two types of pseudo-random noise codes modulated on carrier signals at the L-band frequencies ($L_1 = 1.575$ and $L_2 = 1.227$ GHz). The first code is the so-called "civilian access" (C/A) code, which is primarily intended to ease the acquisition of the second, more precise, P-code. The first code is the only code officially available to "civilian" users. This code has a "chip-rate" of about 1 MHz, and in conjunction with this, the highest ranging accuracy that can be achieved from these measurements is

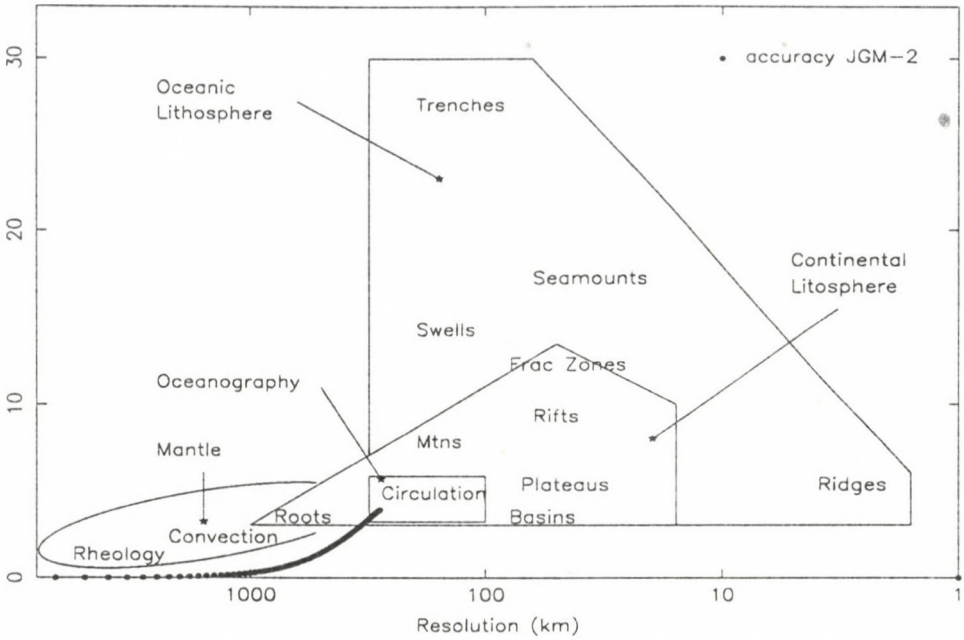

Fig. 1. Summary of requirements for gravity measurement accuracy as a function of the spatial resolution compared to gravity model accuracy (Gravity Workshop 1987) assuming 30 % of phenomena amplitude accuracy (mgal)

nowadays better than 5 m. The P-code has a "chip-rate" of about 10 MHz, and therefore the accuracy of these measurements is better than 1 m. Moreover, if P-code measurements are available in addition to C/A-code measurements, the so-called first-order ionospheric propagation delay can be recovered (Gurtner 1985). Apart from the C/A-code and P-code signals, the carrier itself may also be used for ranging. Although the system was not designed for this application, it was realized already early in the development of the system that the highly-stable oscillators on board of the GPS satellites would allow very precise range measurements on this signal. The precision of these measurements can be as high as a few mm.

The capability of GPS to significantly improve our knowledge of the long-wavelength part of the gravity field, *i.e.* for wavelengths longer than 1000 km, has been shown extensively in several studies, especially the studies related to the Gravity Probe B and TOPEX/Poseidon mission (Smith et al. 1988, Wu and Yunck 1986a and 1986b). In the studies described in the first part of these lecture notes, it will be shown that if a low Earth satellite flying at altitudes of 200 to 500 km, also gravity field information for smaller wavelengths can be extracted from the GPS measurements. The expected accuracy of carrier phase measurements by future space-borne GPS receivers seems to open the possibility to achieve this objective (Ambrosius et al. 1990).

This part of the lecture notes will start with a short description of the space segment of the GPS system. After this, a method will be described to estimate the

formal error of the global gravity field harmonic coefficients from GPS data acquired on board of the low Earth satellite, and presents some results of the application of this method. The GPS SST measurements were assumed to deliver precise information about the orbit of the low Earth satellite position perturbations in the radial, along-track and cross-track directions. These perturbations may be related to the (global) gravity field harmonic coefficients by a linear perturbation theory (Kaula 1966, Schrama 1989, Visser 1992, Visser et al. 1994).

In addition, a method will be described for a local gravity field recovery from GPS SST measurements to a low Earth satellite. A simulation has been performed for a 20° × 20° test area in Western Europe.

GPS space segment

The space segment of the NAVSTAR/GPS system will consist of 24 satellites after its completion (Fig. 2). Included are 3 spares, which will make it possible to maintain a near-optimal configuration, even after the failure of a few GPS satellites. The 24 GPS satellites are and will be located in 6 orbital planes, which have an inclination of 55°, at an altitude of about 20,000 km. The satellites complete 2 orbital revolutions per sidereal day. The space segment has almost been completed. At the writing of these lecture notes, 34 GPS satellites have been launched (Table I). These satellites can be divided into two groups. The first group consists of 11 so-called block I satellites, most of which are inactive and also the remaining will be decommissoned. The final configuration will consist of 24 block II satellites, and already 23 of these satellites have been launched.

In order to fully exploit the information content of GPS measurements between the high-flying GPS satellites and a low Earth satellite, a network of ground stations

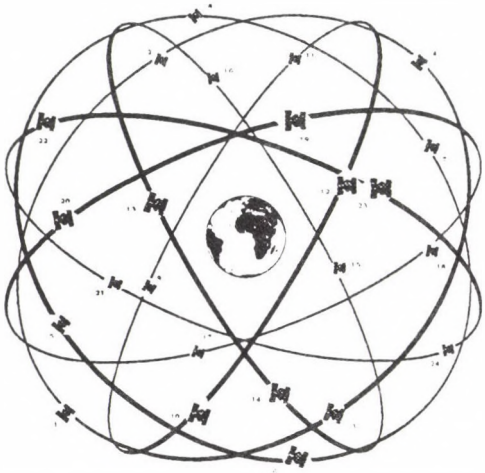

Fig. 2. The uniform-24 satellite constellation including three on-orbit active spares for the Global Positioning System

Table I. NAVSTAR/GPS constellation status

Block ID	*SVN*	*PRN Code*	*Orbital Plane*	*Launch Date (UT)*	*Clock*	*Active*	*Idle*
I							
	01	04		78-02-22		78-03-29	85-07-17
	02	07		78-05-13		78-07-14	81-07-16
	03	06	A-3	78-10-06	Rb[a]	78-11-13	92-05-18
	04	08		78-12-10		79-01-08	89-10-14
	05	05		80-02-09		80-02-27	83-11-28
	06	09		80-04-26		80-05-16	91-03-06
	07			81-12-18		Launch failure	
	08	11	C-3	83-07-14	Cs[b]	83-08-10	93-05-04
	09	13	C-1	84-06-13	Cs	84-07-19	
	10	12	A-1	84-09-08	Cs	84-10-03	
	11	03	C-4	85-10-09	Rb	85-10-30	
II							
II-1	14	14	E-1	89-02-14	Cs	89-04-15	
II-2	13	2	B-3	89-06-10	Cs	89-08-10	
II-3	16	16	E-3	89-08-18	Cs	89-10-14	
II-4	19	19	A-4	89-10-21	Cs	89-11-23	
II-5	17	17	D-3	89-12-11	Cs	90-01-06	
II-6	18	18	F-3	90-01-24	Cs	90-02-14	
II-7	20	20	B-2	90-03-26	Cs	90-04-18	
II-8	21	21	E-2	90-08-02	Cs	90-08-22	
II-9	15	15	D-2	90-10-01	Cs	90-10-15	
IIA							
II-10	23	23	E-4	90-11-26	Cs	90-12-10	
II-11	24	24	D-1	91-07-04	Cs	91-08-30	
II-12	25	25	A-2	92-02-23	Rb	92-03-24	
II-13	28	28	C-2	92-04-10	Cs	92-04-25	
II-14	26	26	F-2	92-07-07	Cs	92-07-23	
II-15	27	27	A-3	92-09-09	Cs	92-09-30	
II-16	32	01[c]	F-1	92-11-22	Cs	92-12-11	
II-17	29	29	F-4	92-12-18	Cs	93-01-05	
II-18	22	22	B-1	93-02-03	Cs	93-04-04	
II-19	31	31	C-3	93-03-30	Cs	93-04-13	
II-20	37	07	C-4	93-05-13	Cs	93-06-12	
II-21	39	09	A-1	93-06-26	Cs	93-07-20	
II-22	35	05	B-4	93-08-30	Cs	93-09-28	
II-23	34	04	D-4	93-10-26	Cs	93-11-22	
II-24							

[a]Rubidium clock.
[b]Cesium clock.
[c]PRN number changed to 1 on 93-01-28.

is necessary. This will enable to compare the GPS SST measurements with GPS measurements collected by these stations in order to eliminate *e.g.* clock errors. Several studies have shown that at least 6 globally distributed ground stations are required for a high-accuracy orbit determination of the low Earth satellite (Wu and Yunk, 1986a and 1986b). Figure 3 shows the tracking network that has been used in the orbit TOPEX/Poseidon computation for cycle 14 at the Section Space Research and Technology (SSR&T), Faculty of Aerospace Engineering, Delft University of Technology. This network consists of 10 globally distributed ground stations.

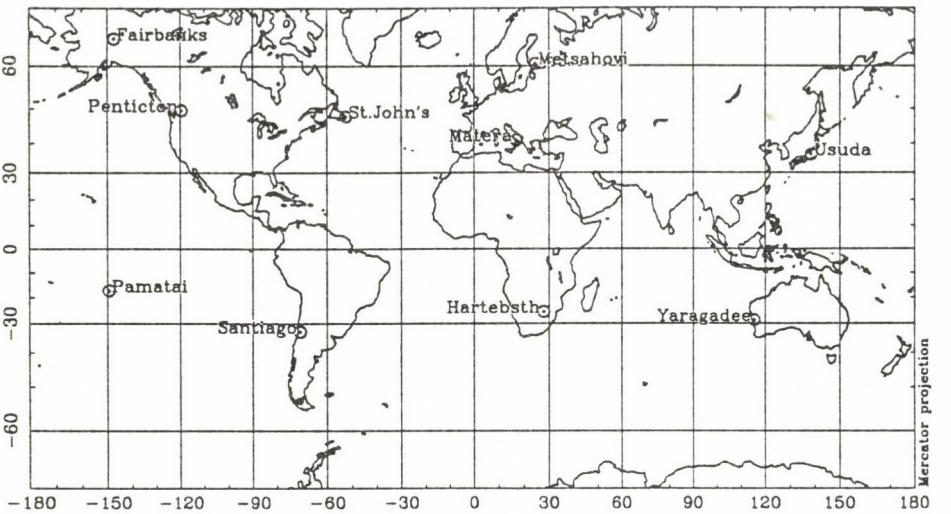

Fig. 3. TOPEX/Poseidon ground tracking network for cycle 14 (January 30 – February 8, 1993)

Global gravity field recovery from GPS

Covariance analyses have been performed to a global gravity field recovery from GPS SST tracking of a low Earth satellite. If the gravity field potential is described by a spherical harmonic expansion, the unknown part of the gravity field potential, denoted by T, may be expressed by (Kaula 1966):

$$T = \frac{\mu}{r} \sum_{l=2}^{\infty} \sum_{m=0}^{l} \left(\frac{a_e}{r}\right)^l \left(\Delta\bar{C}_{lm} \cos m\lambda + \Delta\bar{S}_{lm} \sin m\lambda\right) \bar{P}_{lm}(\sin\phi) \tag{1}$$

where μ denotes the gravity parameter of the Earth, a_e the mean equatorial radius of the Earth, $\bar{P}_{lm}$ the fully-normalized Legendre polynomial of degree l and order m, and $\Delta\bar{C}_{lm}$ and $\Delta\bar{S}_{lm}$ the (unknown) fully-normalized gravity field harmonic coefficients. The satellite position in the rotating geocentric coordinate frame is denoted by the radius r, the longitude λ and the geocentric latitude ϕ. From this equation, any other gravity field induced signal along a near-circular satellite orbit can be derived. In these lecture notes, these signals are orbit perturbations Δr_i, where the indices i and j denote directions in the satellite local-horizontal, local-vertical plane, *i.e.* in the so-called radial, along-track, and cross-track directions. It is assumed that these perturbations can be observed from GPS SST measurements. This will be discussed later in this Section. In order to find the relations for the orbit perturbations, Eq. (1) can be transformed into an equation that contains orbital elements (Kaula 1966). The low Earth satellite is expected to be in a circular repeat orbit, *i.e.* the eccentricity e is equal to zero, and its ground track will repeat after a certain time. For such an orbit, the unknown part of the gravity potential can be

written as:

$$T = \frac{\mu}{a_e} \sum_{l=2}^{nmax} \left(\frac{a_e}{r}\right)^{l+1} \sum_{m=0}^{l} \sum_{p=0}^{l} F_{lmp} S_{lmp} \left(\omega + M, \Omega - \theta\right) \tag{2}$$

where:

$$S_{lmp}\left(\omega + M, \Omega - \theta\right) = \begin{bmatrix} \Delta\bar{C}_{lm} \\ -\Delta\bar{S}_{lm} \end{bmatrix}_{l-m\ odd}^{l-m\ even} \cos((l-2p)(\omega+M) + m(\Omega-\theta)) + \begin{bmatrix} \Delta\bar{S}_{lm} \\ \Delta\bar{C}_{lm} \end{bmatrix}_{l-m\ odd}^{l-m\ even} \sin((l-2p)(\omega+M) + m(\Omega-\theta)) \ . \tag{3}$$

In these equations use has been made of the so-called Kepler elements. These are the satellite's orbital semi-major axis a, eccentricity e, argument of perigee ω, inclination i, right ascension of ascending node Ω and the mean anomaly M. The Greenwich hour angle is denoted by θ, while F_{lmp} is a function depending on the orbital inclination i only (Kaula 1966, Schrama 1989). The gravity field potential is truncated at a certain maximum degree $nmax$. The value for $nmax$ will be specified later.

The Hill equations represent a set of linearized equations for the motion relative to a circular reference orbit (Dunning 1973). If T represents the disturbing potential, the equations for the radial, along-track and cross-track orbit perturbations become for a general term T_{lmp} (Colombo 1989, Schrama 1989, Visser 1992, Zandbergen 1990):

$$\Delta r_{z,lmp} = a\left(\frac{a_e}{a}\right)^l F_{lmp} \left[\frac{2(l-2p)}{f_{lmp}} + \frac{4p-3l-1}{2f_{lmp}+1} + \frac{4p-l+1}{2(f_{lmp}-1)}\right] S_{lmp} \tag{4}$$

$$\Delta r_{x,lmp} = a\left(\tfrac{a_e}{a}\right)^l F_{lmp} \left[\frac{2(l+1)-3(l-2p)\frac{1}{f_{lmp}}}{f_{lmp}} + \frac{4p-3l-1}{f_{lmp}+1} + \frac{l-4p-1}{f_{lmp}-1}\right] S^*{}_{lmp} \tag{5}$$

$$\Delta r_{y,lmp} = \tfrac{1}{2} a \left(\tfrac{a_e}{a}\right)^l \frac{1}{f_{lmp}} \left[\left(\frac{F_{lmp}}{\sin i}((l-2p)\cos i - m) - F'_{lmp}\right) S^*{}_{(l+1)mp} - \left(\frac{F_{lmp}}{\sin i}((l-2p)\cos i - m) + F'_{lmp}\right) S^*{}_{(l-1)mp}\right] \tag{6}$$

where:

$$f_{lmp} = l - 2p - \frac{m}{n_r} \tag{7}$$

$$S^*{}_{lmp}(\omega + M, \Omega - \theta) = \begin{bmatrix} \Delta\bar{C}_{lm} \\ -\Delta\bar{S}_{lm} \end{bmatrix}_{l-m\ odd}^{l-m\ even} \sin((l-2p)(\omega+M) + m(\Omega-\theta)) - \begin{bmatrix} \Delta\bar{S}_{lm} \\ \Delta\bar{C}_{lm} \end{bmatrix}_{l-m\ odd}^{l-m\ even} \cos((l-2p)(\omega+M) + m(\Omega-\theta)) \tag{8}$$

and F'_{lmp} is the derivative of F_{lmp} with respect to the inclination i. The factor n_r is the number of nodal orbital revolutions per day. These equations establish the relation between the disturbing potential T and the orbit perturbations in the radial, along-track and cross-track directions. It can be shown that for an exact repeat orbit, these equations become true Fourier series (Colombo 1984). Several studies have shown the validity of these equations (Schrama 1989, Visser 1992).

The first step of an actual gravity field recovery and/or adjustment consists of a precise dynamic orbit determination of the low Earth satellite with the best state-of-the-art dynamic models. In this step, an orbit will be computed that fits best through the GPS SST measurements. This orbit will be subtracted from these measurements to obtain GPS SST measurement residuals. These residuals are assumed to be caused by measurement noise and deficiences in the dynamic models employed, and are used to determine corrections of the satellite position in the radial (Δr_z), along-track (Δr_x), and cross-track (Δr_y) directions. The measurement equations are obtained by summing the Eqs (4)–(6) over the degree l, order m, and the index p:

$$\Delta r_x = \sum_{l=2}^{nmax} \sum_{m=0}^{l} \sum_{p=0}^{l} \Delta r_{x,lmp} \tag{9}$$

$$\Delta r_y = \sum_{l=2}^{nmax} \sum_{m=0}^{l} \sum_{p=0}^{l} \Delta r_{y,lmp} \tag{10}$$

$$\Delta r_z = \sum_{l=2}^{nmax} \sum_{m=0}^{l} \sum_{p=0}^{l} \Delta r_{z,lmp} \tag{11}$$

Normal equations

It has been stated that for a circular repeat orbit, the linear relations connecting the unknown harmonic coefficients with the GPS SST measurements (or orbit perturbations) can be represented by Fourier series. If many observations are made during a complete repeat period and a constant sampling-rate is applied, it can be shown that the normal matrix becomes block-diagonal when organized per order (Colombo 1984, Rummel 1990, Schrama 1990 and 1991, Visser 1992, Visser et al. 1994). For different orders m of the gravity field coefficients, the frequencies of the orbit perturbations or gradiometer signal caused by such coefficients become decorrelated (provided a perfect coverage). The greatest dimension of these blocks is equal to approximately half the maximum degree of the gravity field harmonic expansion that is to be determined, because also $\Delta\bar{S}_{lm}$ and $\Delta\bar{C}_{lm}$ become uncorrelated. The inversion of the total normal matrix transforms to an inversion of block-matrices with dimensions much smaller than the total number of unknowns. This prevents the necessity of long and costly computer runs. An additional advantage is that only a small part of the total normal matrix (only the blocks on the diagonal) has to be stored. It is believed that the previous procedure can also be

applied in a real gravity field model recovery experiment using real GPS observations, and that the normal equations for the gravity field unknowns can be forced into a block-diagonal matrix and an iterative procedure can be applied.

The diagonal of the inverse of the normal equations contains the formal errors of the estimated gravity field parameters. These formal errors are properly scaled by taking into account the number of GPS SST measurements and the accuracy of these measurements. The scaling can be written as:

$$scale = \sigma(orbit)\sqrt{\Delta t} \tag{12}$$

where $\sigma(orbit)$ is the measurement accuracy, and Δt the measurement interval. The results to be discussed are based on these formal error estimates.

Simulation set-up and results

In the global gravity field recovery covariance analyses, it is assumed that the measurement interval is equal to 1 sec and the measurement accuracy of the observed orbit corrections 1 cm, nominally. This 1 cm accuracy estimate is based on the accuracy of ionospheric-free combinations of carrier-phase measurements. In fact, the GPS receiver on board of TOPEX/Poseidon collects carrier-phase SST measurements with an accuracy of the order of a few mm. The repeat period for orbit of the low Earth satellite is chosen equal to 90 days, and the orbit is assumed to be circular and to have an inclination of 90° (polar). The effect of the orbit height of the low Earth satellite on the gravity field recovery has been studied, and besides the nominal values for the measurement interval and measurement accuracy, some other values have been used.

It can be seen from Fig. 4 that for an orbit height of 200 km, a gravity field model can be estimated with a maximum significant degree equal to 115, *i.e.* a resolution of approximately 170 km, much better than the state-of-the-art in gravity field modeling (Fig. 1). The maximum significant degree is defined as the degree for which the formal errors are larger than the order of magnitude of gravity field parameters indicated by Kaula's rule of thumb:

$$O(\Delta\bar{S}_{lm}, \Delta\bar{C}_{lm}) = 10^{-5}/l^2 \tag{13}$$

For an orbit height of 1350 km, *i.e.* the orbit height of TOPEX/Poseidon, the resolution that can be achieved is about 500 km.

Figure 5 shows the effect of the measurement interval and accuracy for an orbit height of 200 km. It can be seen that if the measurement interval is equal to 10 sec, the maximum significant degree drops from 115 for the nominal case to about 97. If the measurement accuracy drops from 1 cm to 10 cm, the maximum significant degree becomes 79. These results show the importance of implementing a GPS receiver on board of the low Earth satellite that has the capability of measuring on both frequencies with the highest accuracy possible.

It is possible to summarize the previous results in the form of accuracy estimates for gravity anomalies and geoid undulations on the Earth's surface. This has been

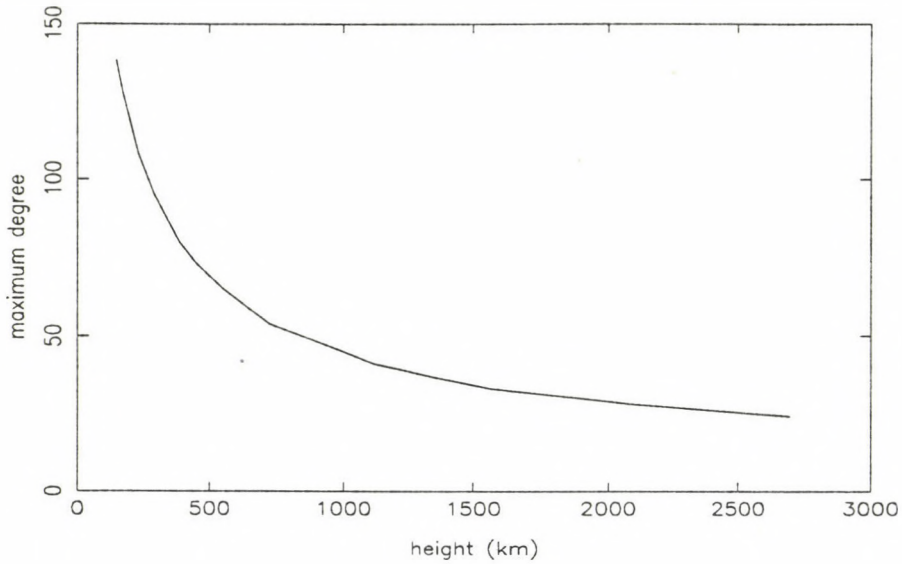

Fig. 4. Maximum significant degree of a gravity field model estimated from GPS SST measurements to a low Earth satellite as a function of the orbit height

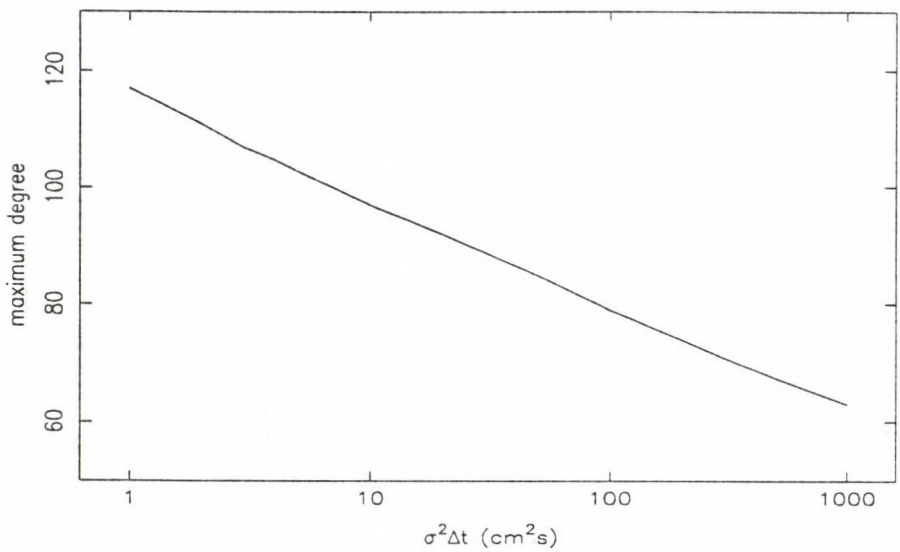

Fig. 5. Effect of measurement interval and measurement accuracy of the maximum significant degree of a gravity field model estimated from GPS SST measurements to a low Earth satellite

done for gravity anomalies and geoid undulations for a resolution of 1°, or about 100 km (in agreement with the resolution requirements for the ARISTOTELES mission (CIGAR 1989 and 1990)). By presenting results in this form, it will also be possible to compare the global gravity field recovery covariance analysis results with

the results of the local gravity field recovery experiment that will be discussed in the next Section. Results are listed in Table II. The values listed in this table were obtained by taking into account the formal errors up to the maximum significant degree. Omission errors caused by degrees larger than the maximum significant degree were added based on the Rapp'79 anomaly degree variance model (Rapp 1979). Again, the state-of-the-art in gravity field modeling is represented by the JGM-2 gravity field model. According to the calibrated covariance of this model and taking into account omission errors, the accuracy of $1° \times 1°$ gravity anomalies and geoid undulations is equal to 16.4 mgal and 104 cm, respectively. These numbers represent a global rms. It can be shown that the accuracy over the oceans is better and that the largest errors are made over the land areas.

Table II. Error estimates mean $1° \times 1°$ gravity anomalies (Δg) and geoid undulations (Δh)

Orbit height (km)	Δg *(mgal)*	Δh *(cm)*	*Orbit height (km)*	Δg *(mgal)*	Δh *(cm)*
145	11.1	41.5	450	16.0	90.8
169	11.9	46.6	551	16.6	102.2
193	12.5	51.6	727	17.4	119.3
229	13.3	58.5	1332	18.7	162.6
290	14.3	68.7	1564	19.0	174.9
385	15.4	82.6	2084	19.5	201.3
JGM-2	16.4	103.8			

sum of commission and omission errors
omission errors from Rapp'79 anomaly degree variance model

From Table II it can be seen that even if the low Earth satellite flies as high as 500 km, still an improvement in the state-of-the-art in gravity field modeling can be obtained. In fact, the results listed in Table II hold for a situation in which it is assumed that only 90 days of GPS SST data will be available, and in the computation of formal errors no *a priori* information, *e.g.* the information that led to the JGM-2 model, was taken into account. In reality, it is expected that much more GPS SST data will become available and that this data will be added to what we already know of the Earth's gravity field.

Local gravity field recovery from GPS

Besides the global gravity field recovery covariance analyses, a more thorough study to a local gravity field recovery from GPS SST measurements has been performed. The following procedure has been developed to extract local gravity field information from GPS SST measurements to a low Earth satellite. Just as in the global gravity field recovery from GPS SST measurements, it is assumed that first an *a priori* orbit of the low Earth satellite is computed with the best dynamic models available, and this orbit is used to compute GPS SST measurement residuals. These

residuals are used to estimate residual accelerations of the low Earth satellite for each GPS link in the direction of the relevant GPS satellite in view of the low Earth satellite. If more than 3 GPS satellites are in view, it is possible to transform these accelerations into accelerations in the radial, along-track and cross-track directions. These accelerations are used to compute a regular grid of satellite acceleration at a mean altitude of the low Earth satellite. Finally, this grid is continued downward to the Earth's surface in the form of gravity anomalies and geoid undulations. The preceding steps will be discussed in more detail in the following Sections.

Determination of residual satellite accelerations

GPS SST measurement residuals (Δr_{GPS}) are obtained by subtracting an *a priori* orbit from the GPS SST measurements. For each GPS link, a second-order polynomial is estimated from 30 sec of GPS SST residuals:

$$\Delta r_{GPS} = a_0 + a_1 t + \frac{1}{2} a_2 t^2 \tag{14}$$

The second derivative of this polynomial gives an estimate of the residual acceleration of the low Earth satellite in the direction of the line-of-sight between the relevant GPS satellite and the low Earth satellite, assuming that the GPS measurement residuals are caused by orbit errors of the low Earth *a priori* satellite orbit. It is assumed that the orbit errors of the high-altitude GPS satellites have a very long-wavelength and do not play a role in the estimation of a relative short-wavelength gravity field model. Furthermore, it is expected that GPS orbit errors are eliminated for the greater part for example by forming double differences of the GPS SST and ground station measurements. If more than 3 GPS satellites are in view of the low Earth satellite, it is possible to estimate residual accelerations in the radial, along-track and cross-track direction. This is done by the method of least-squares. The result is a huge number of residual accelerations along the orbit of the low Earth satellite.

Gridding of residual accelerations and downward continuation

The "observed" residual accelerations along the orbit of the low Earth satellite are used to estimate a regular grid of residual accelerations at a mean altitude of the low Earth satellite. This grid will have a constant resolution in both longitude and latitude. This process is referred to as gridding. The resolution will be referred to as the grid size. This gridding is achieved by collecting all residual satellite accelerations along the orbit of the low Earth satellite for cells that have a size equal to the grid size in longitude and latitude. For each cell, one acceleration in the radial, one in the along-track, and one in the cross-track direction is estimated by the method of least-squares collocation. Finally, this regular grid is continued downward to the Earth's surface in the form of gravity anomalies and geoid undulations, also by using the method of least-squares collocation. The result of this downward continuation will also be a regular grid, in this case of gravity anomalies and geoid

undulations. The latter step can be represented as (Moritz 1980):

$$\Delta g/\Delta h = C_{\Delta g/\Delta h,a}(C_{aa} + D)^{-1}a \tag{15}$$

where $C_{\Delta g/\Delta h,a}$ represents the correlations between the estimated gravity anomaly/geoid undulation ($\Delta g/\Delta h$) and the residual satellite accelerations of the regular grid (a), C_{aa} is the covariance matrix of the residual accelerations, and D represents the covariance matrix of measurement uncertainties. In the computation of the (co)variances use is made of the previously mentioned Rapp'79 anomaly degree variance model.

Simulation set-up

The procedure described above has been used to simulate a local gravity field recovery from GPS SST measurements to the ARISTOTELES satellite, which was supposed to fly in a 200 km altitude orbit (CIGAR 1989 and 1990). The objective of the ARISTOTELES mission is to map the Earth's gravity and magnetic fields with high accuracy and with a resolution of 100 km. Although the mission has been canceled at the moment of this writing, it is believed that the results of this study will be valid for future satellites that will be equipped with a GPS receiver, possibly Gravity Probe B and STEP. An area in Western Europe was selected, extending from 0° to 20° longitude, and from 40° to 60° latitude (Fig. 6). Relatively high frequency gravity signals can be found in this area and therefore, this area is expected to be suitable to show the possibilities for a local gravity field recovery from GPS SST measurements.

In the experiment, it was assumed that the current state-of-the-art in gravity field modeling can be represented by the GEM-T1 gravity field model, complete to degree and order 36 (Marsh et al. 1986). The "real-world" was simulated by the OSU86f gravity field model, complete to degree and order 360 (Rapp and Cruz 1986). The latter model has a resolution of about 50 km, smaller than the ARISTOTELES mission goal. With these models, ARISTOTELES and GPS orbits, GPS SST measurements, and gravity anomalies and geoid undulations were simulated.

A data period of 30 days was simulated in the local gravity field recovery experiment. In the computation of the simulated *a priori* and "real-world" ARISTOTELES and GPS orbits use was made of an 11th-order Adams-Moulton integrator with a step size of 2.5 sec. The measurement interval of GPS SST measurements was taken equal to the integration step size. A small step-size was necessary because of the high resolution of the OSU86f gravity field model (approximately 50 km). In the simulation of the *a priori* ARISTOTELES and GPS orbits the full GEM-T1 model was used. The full OSU86f model was used in the simulation of the "real-world" ARISTOTELES orbits. However, for the simulation of the "real-world" GPS orbits, the OSU86f model was truncated at degree and order 36, because the high-altitude GPS satellites are almost insensitive to gravity terms with a degree larger than 36. ARISTOTELES was assumed to be in a near-circular orbit with an eccentricity of 0.001, a mean semi-major axis of 6577.05 km, an inclination of 96.3° (near-polar). The orbit repeats itself every 91 days after completing 1479 revolutions (CIGAR

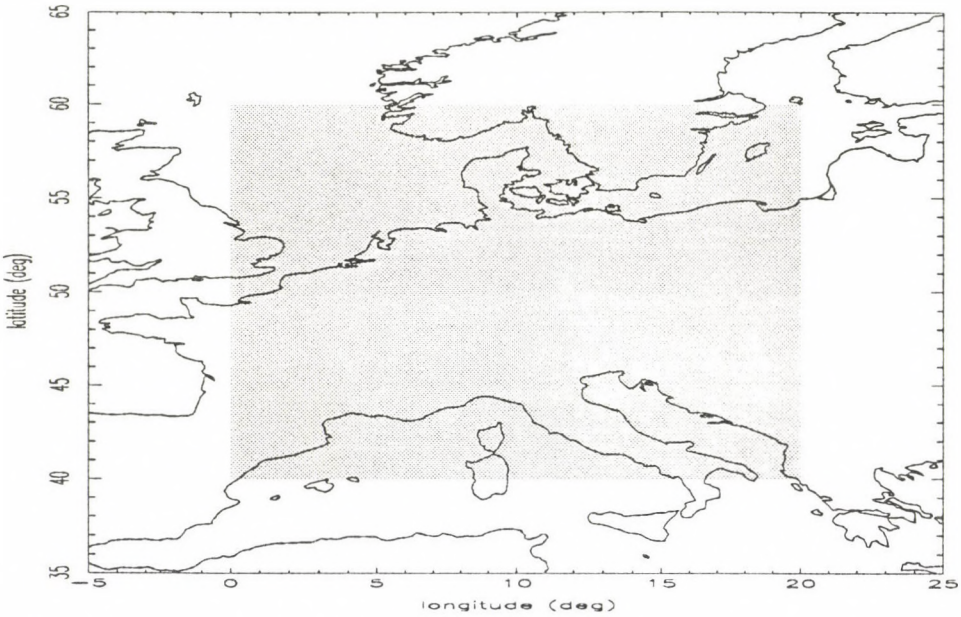

Fig. 6. Selected test area in Western Europe for the local gravity field recovery experiment

1989). The GPS space segment was simulated by 18 GPS satellites uniformly distributed around the Earth in 6 orbital planes with an inclination of 63.45° at an altitude of about 20,000 km (CIGAR 1989). The GPS orbits were supposed to be near-circular and to complete two orbital revolutions per sidereal day.

The simulated *a priori* and "real-world" orbits were used to simulated GPS measurement residuals. GPS SST range measurements were simulated by computing the range between ARISTOTELES and the GPS satellites. This was done using both *a priori* and "real-world" orbits, and GPS SST measurement residuals were obtained by subtracting the *a priori* from the "real-world" measurements. Two cases were investigated, one for which the data were assumed to be perfect (zero noise) and one for which it were assumed that the measurement uncertainty can be modeled by uncorrelated measurement errors with a normal distribution and a standard deviation of 2 mm (this is of the same order as the measurement accuracy of the GPS receiver on board of TOPEX/Poseidon).

The GPS SST measuremement residuals are caused by the differences between the GEM-T1 and OSU86f gravity field models, defined as the residual gravity field. This residual gravity field is used to compute residual gravity field anomalies and geoid undula tions in the selected test area in Western Europe for a regular grid with a resolution of 1° in longitude and latitude. The rms of point gravity anomalies and geoid undulations for this test area is equal to 19.0 mgal and 2.39 m, respectively. These are the signals that were to be recovered from the simulated GPS SST measurement residuals. In other words, the GPS SST measurement residuals will be used to estimate gravity anomalies and geoid undulations, which will be compared with the gravity anomalies and geoid undulations computed directly from the

residual gravity field. This enables to give an accuracy estimate of the procedure described in the beginning of the part about local gravity field recovery from GPS.

The OSU86f gravity field model is only complete to degree and order 360. This means that an omission error will be made due to terms with a degree larger than 360. For mean $1° \times 1°$ gravity anomalies, this error is about 5 mgal according to the Rapp'79 anomaly degree variance model. It was found that the rms of mean $1° \times 1°$ gravity anomalies computed from the residual gravity field (as opposed to point gravity anomalies) is equal to 16.6 mgal. If the omission errors are added to this rms, a value is obtained comparable to the rms of the point gravity anomalies (19.0 mgal). Therefore, the results to be presented hold for point gravity field anomalies and point geoid undulations, but are assumed to be representative for a mean $1° \times 1°$ gravity anomaly/geoid undulation recovery.

Results

From the GPS SST measurement residuals, it was possible to recover residual satellite accelerations with an accuracy of the order of 0.3 and 1.0 mgal (compared to an rms of the residual accelerations of the order of 2–3 mgal) if the data noise was equal to 0 and 2 mm, respectively. From these accelerations, it was possible to recover gravity anomalies and geoid undulations with accuracies of 8.9 and 12 mgal, and 45 and 63 cm, if the data noise was equal to 0 and 2 mm, respectively. As noted above, these accuracy estimates were obtained by comparing the recovered gravity anomalies/geoid undulations with those computed directly from the residual gravity field. These results show that even for perfect GPS SST measurements (0 mm data noise), errors in the recovered gravity anomalies and geoid undulations of about 9 mgal and 45 cm are made. This is caused by the fact that the downward continuation of the residual satellite accelerations to the Earth's surface is an instable process, in which small errors at the satellite altitude are magnified (Tikhonov 1963). In addition, small errors are made in the estimation of residual satellite acceleration for each GPS link (Eq. (14)). These accelerations were obtained by taking the second derivative of a second-order polynomial, fitted through a batch of 30 sec of GPS SST measurement residuals. In fact, the second-order polynomial is an approximation of the real residuals. Finally, in a local gravity field recovery, long-wavelength (the test area has a size of only 20°) gravity field phenomena are only partly visible.

The results of the local gravity field recovery were found to be in close agreement with those of the global gravity field recovery covariance analyses. For example, for a data noise of 2 mm, the accuracy of recovered gravity anomalies in the local gravity field recovery experiment is equal to about 12 mgal. From Table II it can be seen that the global gravity field recovery covariance analyses predict that for an orbit height of 200 km (ARISTOTELES orbit height) it will be possible to recover gravity anomalies with a comparable accuracy. It must be noted that the results of the local gravity field recovery experiment hold for a 30-day period, with a measurement accuracy of 2 mm and a measurement interval of 2.5 sec. For the global gravity field recovery covariance analyses, the results hold for a 90-day period,

with a measurement accuracy of 1 cm and a measurement interval of 1 sec. If in the global analyses, also a 30-day period, a measurement interval of 2.5 sec and a measurement accuracy of 2 mm was applied, the results are almost the same (see also Eq. (12)).

Conclusions

Both the global gravity field recovery covariance analyses and the local gravity field recovery experiments indicate that the current state-of-the-art in gravity field modeling can be improved significantly if future low Earth satellites will be equipped with a high-quality dual-frequency GPS receiver. Even if the satellite will fly in an orbit as high as 500 km, still a significant improvement in gravity field modeling can be obtained.

Part II. TOPEX/Poseidon orbit determination from GPS tracking

Introduction

Covariance and actual analyses have already demonstrated that, using observations of the NAVigation Satellite by Timing And Ranging (NAVSTAR) Global Positioning System (GPS), the orbit of the TOPEX/Poseidon satellite can be determined to the decimeter or even sub-decimeter accuracy level (Wu and Yunk 1986a and 1986b). However, the results obtained from these error analyses are based on the observed data quality and the expected GPS observing geometry using hypothetical GPS measurements acquired by the TOPEX/Poseidon flight receiver and a global network of ground receivers. In the results to be discussed in this part, real GPS observations have been processed at Delft University of Technology, Section Space Research & Technology (DUT/SSR&T). These observations were obtained from the GPS Demonstration Receiver (GPSDR) on TOPEX/Poseidon and about 10 globally distributed ground receivers for the period January 30 to February 8, 1993 (cycle 14). All observations were processed with the GIPSY-OASIS II (GPS Inferred Positioning SYstem – Orbit Analysis and SImulation Software II) software system, developed at the Jet Propulsion Laboratory (JPL) in Pasadena, California.

To assess the GPS performance on TOPEX/Poseidon, two estimation strategies were investigated, *viz.* (i) a fully dynamic technique and (ii) a reduced-dynamic technique. The general characteristics of the dynamic approach are:

- it reduces measurement noise
- it yields greater information concerning the satellite state at a single epoch
- it links the satellite state at different observation times by integration of the equations of motion
- it preserves maximum data strength

- it yields the lowest formal orbit errors
- it causes systematic errors in the satellite state, which tend to grow as the data arc length increases, when the satellite forces (dynamics) are not adequately modeled
- it is favoured above an altitude of about 2000 km when dynamic model errors are small and GPS observability diminishes.

Those of the reduced-dynamic approach are:

- it exploits the geometric information available from GPS measurements while minimizing requirements for precision dynamic models
- it uses the dynamic and non-dynamic methods of state transition with each given appropriate weighting
- it is of greatest value when the dynamic and non-dynamic solutions are comparable
- it yields the lowest overall orbit errors
- it is favoured between altitudes of about 400 and 2000 km when GPS observing geometry is reasonably good.

The TOPEX/Poseidon and GPS orbit accuracies were determined by examination of the observation residuals and by comparison of the ephemeris differences over 6-hour overlaps of ten 30-hour contiguous solution arcs. Furthermore, the solved-for GPS orbits were compared with the *a priori* precise GPS ephemerides, which were determined from the orbits of the International GPS Service for Geodynamics (IGS) having an overall RMS orbit accuracy of about 30 cm.

TOPEX/Poseidon

TOPEX/Poseidon, a US/French oceanographic mission, was launched on August 10, 1992 aboard the Ariane Launch vehicle of the European Space Agency (ESA), into a near-circular orbit with an eccentricity of 0.000455, a mean semi-major axis of 7720 km (1340 km altitude), and an inclination of 66°. The ground track repeats every 10 days after TOPEX/Poseidon completes 127 orbital revolutions. It carries a GPS Demonstration Receiver (GPSDR), which is the first highly sophisticated GPS receiver that can actually record both carrier phase and pseudorange measurements at an extremely high data rate. The pseudo-ranges are recorded every second, while the carrier phases are smoothed against the carrier and recorded every 10 seconds. Since the GPSDR can take measurements from up to six GPS satellites simultaneously, it, accordingly, provides a continuous, global and dense coverage of observations. At any instant of time this coverage is also in many directions. This is obviously an advantage of the GPS tracking system compared to both the SLR (Satellite Laser Ranging) and DORIS doppler systems, which do not have

continuous coverage of observations in many directions at one time (Bertiger et al. 1993). The TOPEX/Poseidon ground tracking network for cycle 14 was already shown in Fig. 3, showing a global distribution of the 10 ground stations.

Dynamic models and estimation scenario

As mentioned before, the GPS observations were processed in a dynamic and reduced-dynamic estimation scheme with the GIPSY-OASIS II analysis software. The dynamic models used for TOPEX/Poseidon have been improved considerably. For example, the Joint (NASA/GSFC (National Aeronautics and Space Administration/Goddard Space Flight Center), Greenbelt, Maryland and CSR/UTA (Center for Space Research/the University of Texas at Austin), Austin, Texas) Gravity field Model 2 (JGM-2), complete to degree and order 70, was used for TOPEX/Poseidon. In comparison, the GEM-T3 model, truncated at degree and order 8, was used for the GPS satellites. A second noteworthy model element is the solar pressure model. TOPEX/Poseidon was modeled as a combination of 8 components, the so-called box-wing model, with each component having its own properties regarding area, specularity and diffusivity. This macro model was also adopted for the earth radiation and atmospheric drag model of TOPEX/Poseidon. Furthermore, three tide models were applied, such as solid Earth tide, ocean tide and pole tide. Finally, and this is really very important, it is possible to specify initial values for the empirical accelerations of the custom force model of TOPEX/Poseidon. Information about the constant accelerations and the sine and cosine 1- and 2-cpr (cycle-per-revolution) terms of the empirical force on TOPEX/Poseidon can be obtained by first performing a dynamic filtering and solve for these parameters. Then, these solved-for empirical force terms can be kept fixed in a subsequent reduced-dynamic filtering. The initial state vector of TOPEX/Poseidon was obtained from the SLR data processing.

Furthermore, the GPS orbits of IGS (International GPS Service for Geodynamics) were used as *a priori* information for the epoch states of all GPS satellites. These orbits are a weighted mean of the orbits computed by several different processing centers, *viz.*

(i) European Space Operation Centre (ESOC) in Darmstadt, Germany,
(ii) GeoForschungs Zentrum (GFZ) in Potsdam, Germany,
(iii) Centre of Orbit Determination Europe (CODE) in Bern, Switzerland,
(iv) National Geodetic Survey (NGS) in Greenbelt, Maryland,
(v) Jet Propulsion Laboratory (JPL) in Pasadena, California,
(vi) Scripps Institute of Oceanography (SIO) in La Jolla, California,
(vii) Energy Mines Resource (EMR) in Ottawa, Canada,

and are believed to have an average 3-D RMS orbit accuracy of about 30 cm. As observations, we used first-order ionospheric-free pseudo-ranges and carrier phase observations with a data noise of 1 m and 1 cm for the 10 ground receivers, and 3 m and 2 cm for TOPEX/Poseidon. The observations were compressed to a data rate of 1 observation per 5 minutes using a cut-off elevation angle of about zero degrees for TOPEX/Poseidon and 15 degrees for all ground stations. For cycle 14,

ten 30-hour data arcs were selected with a 6-hour overlap. The estimated parameters included the TOPEX/Poseidon state vector at epoch, the constant and 1- and 2-cpr cross-track and along-track terms of the TOPEX/Poseidon empirical force, the radial antenna phase center offset of TOPEX/Poseidon, the GPS state vectors at epoch, the constant GPS solar pressure scale factor, the constant and stochastic GPS solar radiation Y-biases, the stochastic GPS solar radiation scaling factors X and Z, the non-fiducial station locations, the random walk modeled troposheric dry zenith residual delays, the X and Y pole position and corresponding rate, the UT1-UTC rate, the carrier phase biases, the TOPEX/Poseidon receiver clock (modeled as white noise), the GPS transmitter clocks (modeled as white noise) and the ground receiver clocks (modeled as white noise) except for one reference clock. As said before, the observations are processed with GIPSY-OASIS II, making use of a SRIF (Square Root Information Filter) filter. This filter is a so-called epoch state filter in which all measurements are processed sequentially. Furthermore, all measurements are divided into finite discrete time intervals known as batches. The time update of the filter propagates the satellite state estimates and covariances from one batch to the next using a state transition model, while the measurement update of the filter incorporates a new batch of measurements. Now, to obtain the optimal parameter estimates and covariances at all batch observation times from 100% of the measurements, a backward smoothing in time is performed. At the same time, it is possible to perform a forward mapping in time, since the filtered state estimates and covariances apply to the epoch state.

Results

Figure 7 (top) shows the TOPEX/Poseidon orbit differences in all three position components between reduced-dynamic and dynamic orbits (with GPS orbits adjusted). Since the major dynamic modeling errors are absorbed by the empirical force in a dynamic filtering, the difference presented actually shows the effect of the higher-order dynamic modeling errors, which still give an rms error in the altitude of about 3 cm, 3.5 cm cross-track and 8 cm along-track. It should be noted that about the same results are obtained when the GPS orbits are not adjusted. The bottom part of Fig. 7 shows the effect of errors in the IGS orbits on the orbit accuracy of TOPEX/Poseidon. Errors in the IGS orbits, which are believed to be accurate to the 30-cm level, still cause errors in the TOPEX/Poseidon orbit of about 1 cm in altitude, 3 cm cross-track and 2.5 cm along-track.

Fixing the GPS orbits reduces the number of estimated parameters considerably and, as a consequence, requires also less CPU time for the filter. This is obviously a major advantage from a computational point of view. However, since GPS orbit errors do not translate directly into errors in the TOPEX/Poseidon orbit but are reduced by roughly a factor of ten (Bertiger et al. 1993), overall RMS orbit errors of about 3 cm remain for TOPEX/Poseidon when the IGS orbits are not adjusted, which can be seen from Fig. 7. Hence, to achieve the highest orbit accuracies for TOPEX/Poseidon, the IGS orbits should be estimated as well, the fact notwithstanding that these orbits appear to be accurate enough for TOPEX/Poseidon orbit determination to the sub-decimeter accuracy level.

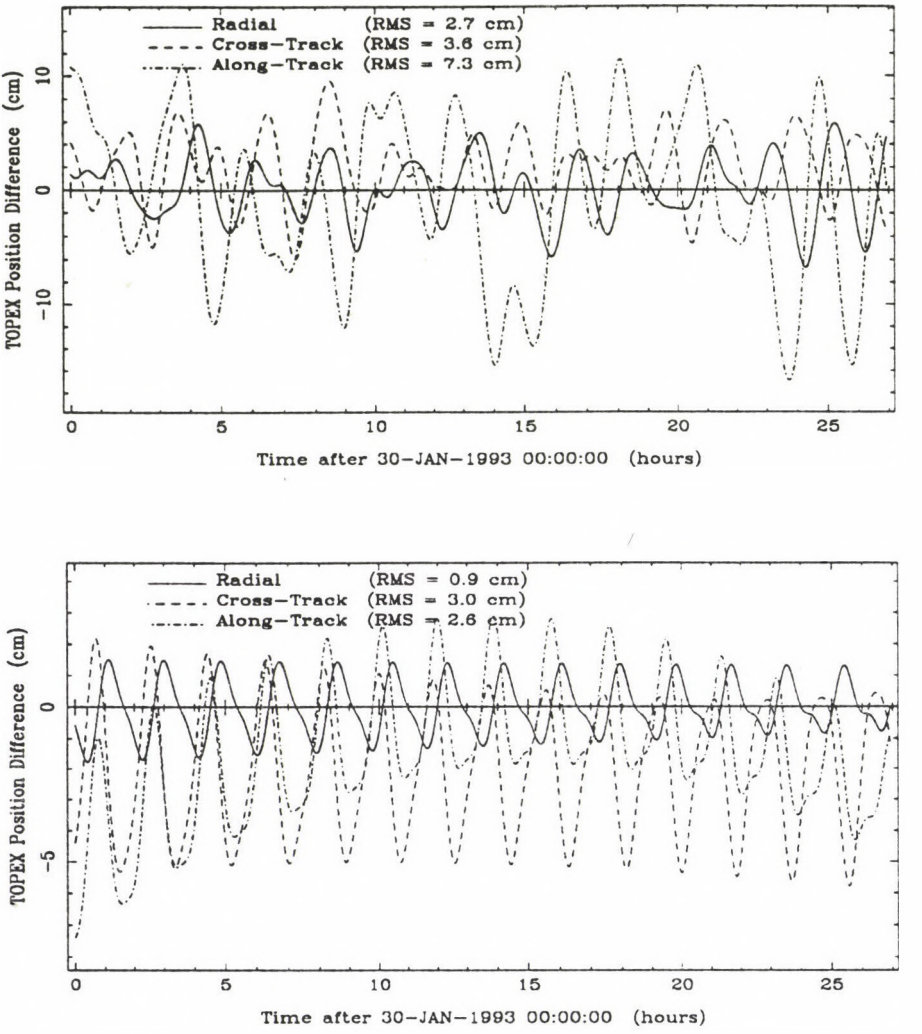

Fig. 7. Differences in TOPEX/Poseidon position components due to the higher-order dynamic modeling errors (top) and due to errors in the precise GPS orbits (IGS, bottom)

Starting with the results of dynamic filtering, the top part of Fig. 8 shows the TOPEX/Poseidon difference in all position components over a 6-hour overlap between two 30-hour data arcs with GPS orbits adjusted. In this case, sub-decimeter orbit accuracy can also be achieved with dynamic filtering, since the RMS difference is about 2 cm in altitude, 4.5 cm cross-track and 6 cm along-track. Up to now, it can be concluded that errors in the TOPEX/Poseidon orbit remain when dynamic filtering is used and when the GPS orbits are not adjusted. The effects of both have been shown. Therefore, to absorb the higher-order dynamic modeling errors and to eliminate the errors introduced by not estimating the IGS orbits, a second

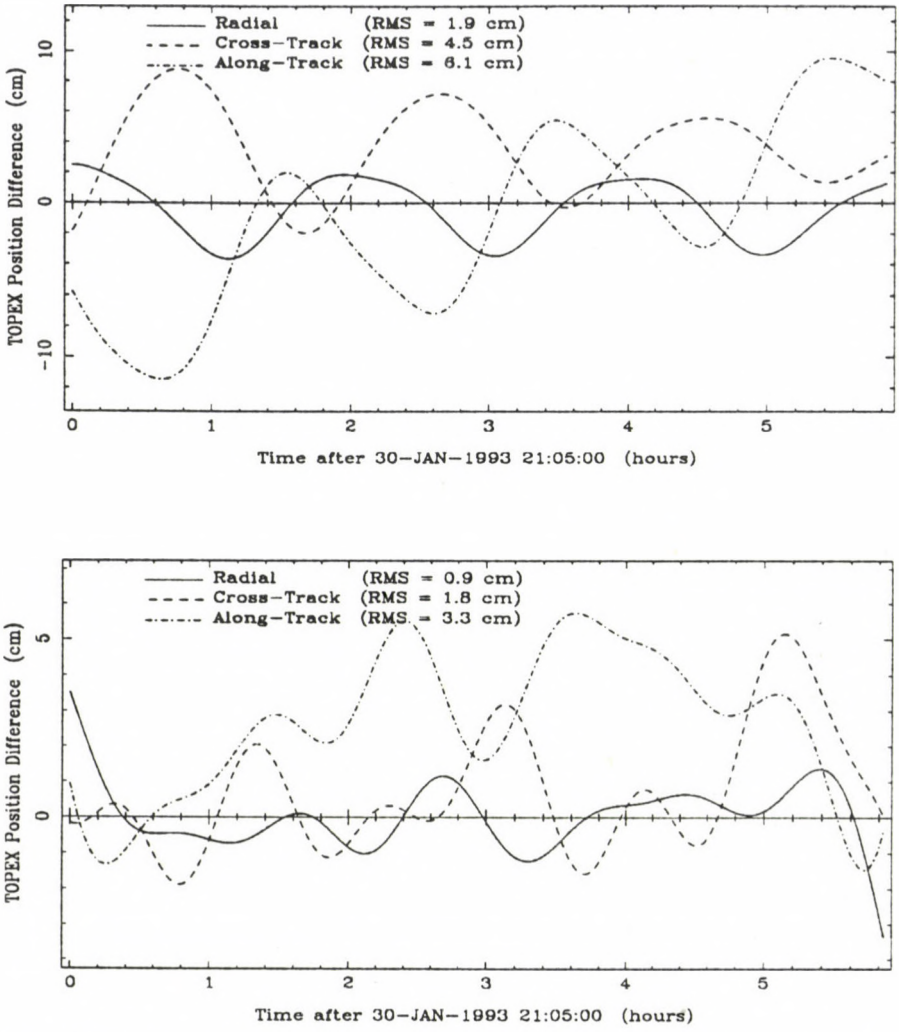

Fig. 8. Differences in TOPEX/Poseidon position components over a 6-hour overlap between two 30-hour data arcs using a dynamic (top) / reduced-dynamic (bottom) strategy with precise GPS orbits (IGS) adjusted

filtering was performed, called the reduced-dynamic filtering with GPS orbits adjusted. Using this approach, a three-dimensional process-noise force was estimated to absorb the remaining small higher-order dynamic modeling errors. The bottom part of Fig. 8 shows the overlap comparison using reduced-dynamic filtering with GPS orbits adjusted. Obviously, the best results are now obtained, with an RMS difference in the TOPEX/Poseidon orbit of about 1 cm in altitude, 2 cm cross-track and 3 cm along-track.

The postfit residual summary for cycle 14 using a dynamic technique is listed in

Table III. TOPEX/Poseidon postfit residual summary for cycle 14 using a dynamic technique

Arc Id	Start Time	Number of Observations	Overall RMS of fit (cm)	TOPEX/Poseidon Observations	TOPEX/Poseidon RMS of fit (cm)
1	29-JAN-1993 21:00	17013/17013	45.59/0.45	1589/1589	69.71/0.92
2	30-JAN-1993 21:00	18886/18882	42.98/0.48	1754/1754	67.72/0.88
3	31-JAN-1993 21:00	16299/16302	51.08/0.47	1755/1755	68.14/0.90
4	01-FEB-1993 21:00	18124/18122	32.49/0.46	1736/1736	70.14/0.86
5	02-FEB-1993 21:00	16439/16438	29.51/0.40	1706/1706	66.53/0.75
6	03-FEB-1993 21:00	17682/17682	31.36/0.46	1743/1743	70.90/0.88
7	04-FEB-1993 21:00	19360/19360	31.97/0.52	1761/1761	71.24/1.05
8	05-FEB-1993 21:00	19513/19513	34.53/0.53	1740/1740	72.38/1.05
9	06-FEB-1993 21:00	19207/19205	41.34/0.50	1717/1717	68.74/0.96
10	07-FEB-1993 21:00	14250/14250	49.09/0.50	1575/1575	75.11/1.01

Table IV. TOPEX/Poseidon postfit residual summary for cycle 14 using a reduced-dynamic technique

Arc Id	Start Time	Number of Observations	Overall RMS of fit (cm)	TOPEX/Poseidon Observations	TOPEX/Poseidon RMS of fit (cm)
1	29-JAN-1993 21:00	17013/17013	45.57/0.39	1589/1589	68.35/0.47
2	30-JAN-1993 21:00	18886/18882	42.90/0.43	1754/1754	66.71/0.44
3	31-JAN-1993 21:00	16300/16302	51.16/0.41	1755/1755	64.17/0.45
4	01-FEB-1993 21:00	18124/18122	31.80/0.41	1736/1736	66.55/0.44
5	02-FEB-1993 21:00	16439/16438	29.40/0.36	1706/1706	66.00/0.40
6	03-FEB-1993 21:00	17682/17682	30.39/0.41	1743/1743	66.60/0.44
7	04-FEB-1993 21:00	19360/19360	30.74/0.45	1761/1761	64.88/0.48
8	05-FEB-1993 21:00	19513/19513	33.55/0.46	1740/1740	66.39/0.49
9	06-FEB-1993 21:00	19207/19205	40.79/0.44	1717/1717	64.18/0.48
10	07-FEB-1993 21:00	14250/14250	48.27/0.42	1575/1575	67.20/0.51

Table III for ten 30-hour data arcs. The left part of relevant columns refers to first-order ionospheric-free pseudo-ranges and the right part to first-order ionospheric-free carrier phase measurements. The mean for both data types is about zero and is accordingly not included. The rms of fit is below 50 cm for the pseudo-ranges and about 0.5 cm for the carrier phases. The average number of observations is almost 18000 for each data arc and each data type. Only the last data arc contains less observations, since less ground station data are used. Considering the TOPEX/Poseidon part of the postfit residual summary, it can be concluded that the rms of fit of the pseudo-ranges and carrier phases is less than 75 cm and 1 cm, respectively.

The corresponding postfit residual summary for cycle 14 using a reduced-dynamic strategy is shown in Table IV. The overall rms of fit has only been changed significantly for the carrier phase observations and is about 0.4 cm. The same observation can be made for the TOPEX/Poseidon part of the summary, showing a dramatic improvement of the rms of fit values down to the 0.4 cm level, which is due to the

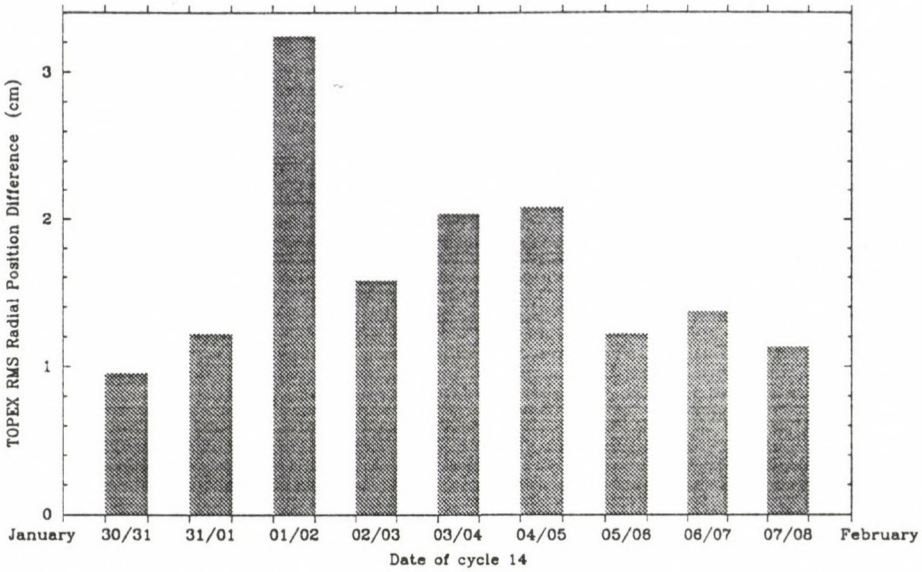

Fig. 9. RMS radial position differences of TOPEX/Poseidon over 6 hours for ten 30-hour consecutive data arcs using a reduced-dynamic technique

fact that the remaining higher-order dynamic modeling errors have been absorbed by a three-dimensional process-noise force. As was said before, the purpose of the TOPEX/Poseidon orbit determination is to determine the radial component of the orbit to the sub-decimeter accuracy level. Figure 9 shows the radial position overlap results for a 6-hour overlapping time period between 30-hour consecutive data arcs using reduced-dynamic filtering. Overall, it is obvious that the RMS radial position overlap is about 1.5 cm, an extremely good result regarding the fact that the precision of the altimeter measurements is about 2 to 3 cm.

Conclusions

The conclusions that can be drawn are that the IGS orbits seem to be accurate enough for TOPEX/Poseidon orbit determination to the sub-decimeter accuracy level. However, for the highest orbit accuracies, the GPS IGS orbits should be estimated as well. The second conclusion is that a preliminary dynamic filtering must be performed to determine the constant and sine and cosine 1- and 2-cpr terms of the two-dimensional (cross-track and along-track) empirical forces in order to absorb the gross effects of mismodeled dynamics for TOPEX/Poseidon. The values of these constant parameters are kept fixed in a subsequent reduced-dynamic filtering. Then, reduced-dynamic filtering must be performed to better approach the true orbit through absorption of the remaining higher-order dynamic modeling errors by a three-dimensional process-noise force. The best results are obtained with reduced-dynamic filtering and GPS orbits adjusted, showing an RMS position difference of TOPEX/Poseidon of about 1 cm in altitude, 2 cm cross-track and 3 cm along-track

over a 6-hour overlap between two 30-hour consecutive data arcs. Furthermore, the rms values of the postfit residuals are about 0.4 cm for the ionospherically calibrated carrier phase observables and about 40 cm for the corresponding pseudo-ranges.

References

Ambrosius B A C, Visser P N A M, Wakker K F 1990: On the use of GPS for the ARISTOTELES mission, contribution to the Dornier/ESA ARISTOTELES Additional Study

Bertiger W, Wu S, Yunck T, Muellerschoen R, Willis P, Bar-Sever Y, Davis A, Haines B, Munson T, Lichten S, Sunseri R 1993: *Spaceflight Mechanics*, 82, 863–876.

Consortium for the Investigation of Gravity Anomaly Recovery (CIGAR) 1989: Study on precise gravity field determination and mission requirements, final report CIGAR Phase I prepared by CISI, ESA Contract 7521/87/F/FL

Consortium for the Investigation of Gravity Anomaly Recovery (CIGAR) 1990: Study on precise gravity field determination and mission requirements, final report CIGAR Phase II prepared by CISI, ESA Contract 8153/88/F/FL

Colombo O L 1984: The global mapping of gravity with two satellites, Netherlands Geodetic Commission, vol. 7, No. 3, Delft, The Netherlands

Colombo O L 1989: Mapping the gravity field with orbiting GPS receivers, Proceedings IUGG GPS Symposium, Edinbourgh

Denker H, Rapp R H 1990: *Journal of Geophysical Research*, 95, 13,151–13,168.

Dunning R S 1973: The orbital mechanics of flight mechanics, NASA, Langley Research Center, and Scientific and Technical Information Office, NTIS, Springfield

Engelis T, Knudsen P 1989: Orbit improvement and determination of the ocean geoid and topography from 17 days of SEASAT data, Department of Geodetic Science and Surveying, Ohio State University, Columbus, Ohio

European Space Agency (ESA) 1991: The solid-Earth mission ARISTOTELES, Proceedings of the Anacapri Workshop, ESA SP-329, Paris

Gravity Workshop 1987: Geophysical and geodetic requirements for global gravity field measurements 1987-2000, Report of a Gravity Workshop Colorado Springs, Geodynamics Branch, Division of Earth Science and Applications, NASA

Gurtner W 1985: GPS-papers presented by the astronomical institute of the university of Berne in the year 1985, Mitteilungen der Satelliten-Beobachtungsstation Zimmerwald, No. 18.

Kaula W M 1966: Theory of satellite geodesy, Blaisdell Publishing Co.

Lambeck K, Aristoteles 1990: *ESA Journal*, 14, No. 1.

Marsh J G, Lerch F J, Putney B H, Christodoulidis D C, Felsentreger T L, Sanchez B V, Smith D E, Klosko S M, Martin T V, Pavlis E C, Robbins J W, Williamson R G, Colombo O L, Chandler N L, Rachlin K E, Patel G B, Bhati S, Chinn D 1986: An improved model of the Earth's gravitational field: GEM-T1, Geodynamics Branch GSFC, EG&G/Washington Analytical Services Center, and Science Applications and Research Corp.

Marsh J G, Lerch F J, Putney B H, Felsentreger T L, Sanchez B V, Klosko S M, Patel G B, Robbins J W, Williamson R G, Engelis T E, Eddy W F, Chandler N L, Chinn D S, Kapoor S, Rachlin K E, Braatz L E, Pavlis E C 1989a: The GEM-T2 gravitational model, NASA Technical Memorandum 100746, Greenbelt, Maryland

Marsh J G, Lerch F J, Koblinsky C J, Klosko S M, Robbins J W, Williamson R G, Patel G B 1989b: Dynamic sea surface topography, gravity, and improved orbit

accuracies from the direct evaluation of SEASAT altimeter data, NASA Technical Memorandum 1000735, Greenbelt, Maryland

Moritz H 1980: Advanced physical geodesy. Herbert Wichmann Verlag, Abacus Press

Nerem R S, Tapley B D, Shum C K, Yuan D N 1988: A model for the general ocean circulation determined from a joint solution for the Earth's gravity field, Proceedings of the Chapman Conference on Progress in the Determination of the Earth's Gravity Field, Ft. Lauderdale, Florida

Rapp R H 1979: Potential coefficient and anomaly degree variance modeling revisited, Report No. 292, Department of Geodetic Science, Ohio State University, Columbus, Ohio

Rapp R H, Cruz J Y 1986: Spherical harmonic expansions of the Earth's gravitational potential to degree 360 using 30 minute mean anomalies. Department of Geodetic Science and Surveying, Ohio State University, Columbus, Ohio, Report No. 376.

Rapp R H, Pavlis N K 1990: *Journal of Geophysical Research*, 95, 21,885–21,911.

Rummel R 1990: A method of global recovery of harmonic coefficient from SGG data with peculiar geometry, CIGAR Phase II: Work Package 510 A, Contribution to final report of "Study on Precise Gravity Field Determination Methods and Mission Requirements (Phase 2)", ESA Contract No. 8153/88/F/FL

Schuyer M, Silvestrin P, Aguirre M 1992: Probing the Earth from space - The ARISTOTELES mission, ESA Bulletin, No. 72, November 1992

Schrama E J O 1989: The role of orbit errors in processing of satellite altimeter data, Ph.D. Thesis, Delft University of Technology

Schrama E J O 1990: Gravity field error analysis: applications of GPS receivers and gradiometers on low orbiting platforms, NASA Technical Memorandum 100679, Goddard Space Flight Center, Greenbelt, Maryland

Schrama E J O 1991: *Journal of Geophysical Research*, 96, B12, 20041–20051.

Smith D E, Lerch F J, Colombo O L, Everitt C W F 1988: In: Proceedings of the Chapman conference on progress in the determination of the Earth's gravity field. Ft. Lauderdale, Florida, 159-163.

Tikhonov A N 1963: *Soviet Math. Dokl.*, 4, 1624–1627.

Visser P N A M 1992: The use of satellites in gravity field determination and model adjustment, Ph.D. Thesis, Delft University Press, Delft

Visser P N A M, Wakker K F, Ambrosius B A C 1993: *Marine Geodesy*, 16, 215–239.

Visser P N A M, Wakker K F, Ambrosius B A C 1994: *Journal of Geophysical Research*, 99, 2841–2851.

Wu J-T, Yunck T P 1986a: Topex orbit determination and gravity recovery using GPS data from repeat orbits, AIAA-86-0575, Jet Propulsion Laboratory, California

Wu S C, Yunck T P 1986b: Non-dynamic decimeter tracking of Earth satellites using the global positioning system, AIAA-86-0404, Jet Propulsion Laboratory, California

Zandbergen R C A 1990: Satellite altimeter data processing, from theory to practice. Ph.D. Thesis, Delft University Press, Delft, The Netherlands

Acta Geod. Geoph. Hung., Vol. 29(3–4), pp. 389–395 (1994)

SOME REMARKS ON THE USE OF GPS FOR GEODYNAMICAL RESEARCH

H LICHTENEGGER[1]

Positioning by space techniques is an adequate mean for geodynamical research where GPS is favoured from economic reasons for baselines up to some 100 km. The paper discusses the model for the measured carrier phases in order to estimate effects of unmodeled biases and noise. Also, some remarks in the context of modeling regional or local deformations by strain analysis are given.

Keywords: GPS; phase model; strain analysis

1. Introduction

Nowadays, positioning by satellite techniques yields global accuracies in the cm-level. Geodesy is, therefore, capable to contribute substantially to geodynamical research, in particular to monitor crustal deformation caused by neotectonics.

Comparing modern techniques for positioning, the highest accuracy for very long baselines is still achieved by Very Long Baseline Interferometry (VLBI) and Satellite Laser Ranging (SLR). For regional or local geodynamics with baselines in the range of 100 km the Global Positioning System (GPS) gives comparable precision, see Fig. 1, but GPS is favoured because of its mobility. Note that only for short baselines up to 3 km conventional terrestrial techniques may be superior in precision.

In principle, VLBI and SLR are also applicable in mobile mode but they require rather expensive instrumentation. Also, SLR has the disadvantage of being weather-dependent. Compared to VLBI and SLR, GPS is highly mobile, very economical, and weather-independent.

So far the advantages of GPS for monitoring crustal movements on a regional or local scale have been mentioned. In the following, however, some drawbacks in the application of GPS for this purpose are outlined, see also Hofmann-Wellenhof et al. (1992).

For the sake of completeness some comments on modeling the observed displacements for geodynamical interpretation are added. However, these remarks are generally valid and do not concern only GPS.

2. Relative positioning with GPS

2.1 General remarks

Having in mind applications of GPS for the determination of crustal movements, then highest accuracy is required. Thus, only the (static) relative positioning technique with observed phases is appropriate. This means that at least two stations

[1]University of Technology, Rechbauerstrasse 12, A–8010 Graz, Austria

1217-8977/94/$ 4.00

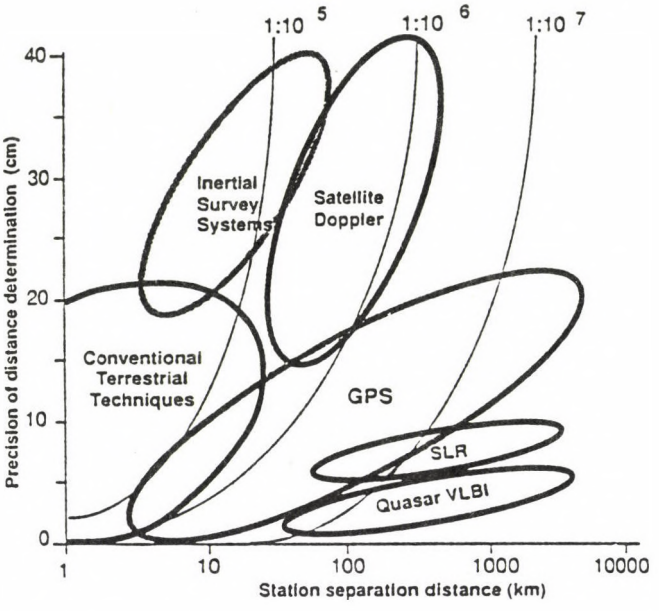

Fig. 1. Precision of relative positioning techniques after Mueller (1990)

(forming a baseline) must track simultaneously the same satellites where the coordinates of one station must be known and the other station is to be determined relatively to the known station.

To monitor geodynamical displacements only, the observation sites must be tied to monuments in solid rock or stable buildings to assure a precise reference when reoccupied.

At present the achievable accuracies are between 0.1 ppm and 1 ppm for baselines up to some 100 km and become even better for longer baselines. Assuming a baseline length of 50 km, an accuracy of 5 mm to 50 mm would result. Note that this level of accuracy is not sufficient to detect small crustal movements within a short time. To get statistically significant statements one has either to extend the timeinterval between remeasurements or to improve the precison of GPS. In the sequel some error sources are discussed with the aid of the model for observed carrier phases.

2.2 Phase model

The model for the phase is given by

$$\begin{aligned}\Phi_i^j(t) = \quad & \varrho_i^j(t) + d\varrho_i^j(t) + c\,(\tau^j(t) - \tau_i(t)) + \lambda\, N_i^j - \\ & \Delta_i^j(t)_{iono} + \Delta_i^j(t)_{tropo} + \Delta_i^j(t)_{mpath} + \text{bias} + \text{noise}\end{aligned} \tag{1}$$

where $\Phi_i^j(t)$ is the measured phase of the carrier wave scaled to range. The term $\varrho_i^j(t)$ is the distance between the (fixed) terrestrial station i and the satellite j (for

which broadcast or precise ephemerides are available), and $d\varrho_i^j(t)$ is the correction due to orbital errors. As usual, c is the speed of light, $\tau_i(t)$ and $\tau^j(t)$ are the receiver and satellite clock errors, λ is the wavelength of the carrier signal which is multiplied by N_i^j, the ambiguity inherent to phase measurements. The effect of ionospheric and tropospheric refraction is denoted by $\Delta_i^j(t)_{iono}$ and $\Delta_i^j(t)_{tropo}$, multipath is taken into account by the term $\Delta_i^j(t)_{mpath}$. Finally, some minor biases (*e.g.*, relativistic effects) and a noise term are added to the phase model.

In the case of a baseline, the phase model must be set up for the two stations involved and usually differences between corresponding observations are formed to reduce or eliminate some error sources.

2.3 Discussion

The highest resolution (*i.e.*, some tenths of a millimeter) for the phases $\Phi_i^j(t)$ is provided by receivers using code correlation for the reconstruction of the two carriers $L1$ and $L2$. However, this technique permits non-authorized users to reconstruct only $L1$ when Anti-spoofing (A-S) is on. Thus, in dual frequency receivers the carrier $L2$ is reconstructed by codeless (or quasi-codeless) techniques. But these methods decrease the signal-to-noise ratio by at least 15 dB, cf. Hofmann-Wellenhof et al. (1994).

The geometric distance $\varrho_i^j(t)$ is a function of the coordinates of station i and the coordinates of the satellite j at epoch t. In order to determine the components of a baselinevector, the coordinates of one endpoint must be given. The baseline solution depends on these coordinates; according to a rule of thumb, a change in position of 20 m changes the baseline solution in the order of 1 ppm. Thus, when processing remeasured baselines one has to apply the same reference coordinates.

The orbital error $d\varrho_i^j(t)$ contains a nominal part and the effect of Selective Availability (SA) which is implemented in all Block II satellites. The nominal part can be reduced by using precise ephemerides and improved models (*e.g.*, for the solar radiation pressure) for the orbits. The effect of SA is reduced (but generally not eliminated) by differentiating the measured phases. According to a rule of thumb, the resulting relative baseline error equals the relative orbital error. Consequently, for geodynamical research one should use only precise ephemerides (which prevents real-time solutions) or one should process the data with software providing orbit relaxation (which fails for small areas).

The satellite clock error $\tau^j(t)$ can be modeled, for example, by a polynomial with coefficients transmitted in the navigation message. However, this approach is degraded by SA. Thus, it is advantageous to eliminate this error by differentiating the measured phases to the satellite j between the baseline stations.

Generally, the internal receiver clock is not very stable and the error $\tau_i(t)$ cannot be modelled. To overcome this problem one could use an external frequency standard with better stability. Another possibility is to eliminate the receiver clock error by differentiating the measured phases from station i to all observed satellites. Thus, forming double differences the receiver and satellite clock errors are eliminated. But note that single differentiating increases the noise level by a factor $\sqrt{2}$ and produces

correlations. In practice, physical correlations are always neglected, mathematical correlations must be taken into account in case of double (and triple) differences.

The ambiguity N_i^j is by definition an integer value and independent of time t as long as no loss-of-lock occurs. The integer feature of the ambiguities is also preserved in case of differentiating. However, when adjusting for example double differences the ambiguity combination is estimated as a real value. Various strategies are applied to find the most probable integer number. In a second adjustment the fixed ambiguities are then introduced as known quantities.

The effect of ionospheric refraction $\Delta_i^j(t)_{iono}$ is one of the most annoying terms. It is true that most of the effect cancels by forming the "ionospheric-free" phase combination but the $L3$-observable is more noisy than the original phases. Also, the corresponding combination of ambiguities is no longer an integer value. The remaining part of the ionospheric refraction depends on the local ionospheric activity which is a function of time, latitude and the number of sunspots. The effect is minimized during the night and when the 11-year solar cycle has its minimum.

Numerous models exist to correct the effect of tropospheric refraction $\Delta_i^j(t)_{tropo}$. Common to all these models is the sensitivity with respect to the partial pressure of water vapor. Its influence is minimum at low temperatures. Good practice is also to set up a water vapor radiometer at least in a central station of the geodynamical network. Most of the software packages introduce the vertical delay as an additional unknown which is estimated during adjustment.

Critical is the influence of multipath denoted by $\Delta_i^j(t)_{mpath}$. This effect cannot yet be modelled although a lot of research is performed on this topic, cf. Cannon and Lachapelle (1993). The effect is receiver and satellite dependent. This means that repeated measurements for the derivation of displacements should be performed with similar satellite configuration and with unchanged station environment. Moreover, multipath can be reduced or even eliminated by supplying the antennas with groundplanes or by using chokering antennas.

It is beyond the scope of this paper to discuss the problems related to baseline processing. But it should be clear that only multipoint solutions using dual frequency phases and precise ephemerides provide adequate data for geophysical interpretation.

In the case of repeated measurements the results must be transformed to a unique geodetic datum. That means that during each epoch-measurement at least three common points (outside the geodynamic area under investigation) must be occupied.

In some cases the GPS net is supplemented with terrestrial observations like precise ranges or levelled height differences between selected sites. So the problem of a combined ajustment of GPS- and terrestrial observations arises. Finally, the transformation of the Cartesian coordinates into local coordinates requires the knowledge of the local geoid.

3. Geophysical models for crustal movement

3.1 Displacement vectors

Positioning with GPS ends up with the three-dimensional position vector $\mathbf{X}$ and the corresponding covariance matrix $\mathbf{Q}_X$ for each observation site related to a global and geocentric coordinate system. Each remeasurement yields another set of position vectors which, as previously mentioned, must be related to a unique geodetic datum.

Due to crustal movement the position vectors are dependent on time t_i and remeasurements at later epochs lead to the shift vector $d\mathbf{X} = \mathbf{X}(t_2) - \mathbf{X}(t_1)$. The corresponding covariance matrix $\mathbf{Q}_{dX}$ is given by the sum of the covariance matrices at epochs t_1 and t_2. The shift vector can be transformed into a local coordinate system where it may be denoted as $d\mathbf{x}$. Considering the smallness of the shifts, spherical approximation may be applied and the transformation reads

$$d\mathbf{x} = \mathbf{R}\, d\mathbf{X} \tag{2}$$

where the rotation matrix $\mathbf{R}$ depends on the latitude φ and the longitude λ of the observation site. In detail the matrix $\mathbf{R}$ is given by:

$$\mathbf{R} = \begin{bmatrix} -\sin\varphi\cos\lambda & -\sin\varphi\sin\lambda & \cos\varphi \\ -\sin\lambda & \cos\lambda & 0 \\ \cos\varphi\cos\lambda & \cos\varphi\sin\lambda & \sin\varphi \end{bmatrix}. \tag{3}$$

The components of the shift vector $d\mathbf{x}$

$$d\mathbf{x} = \mathbf{x}(t_2) - \mathbf{x}(t_1) = (dx, dy, dz)^T = (a\, d\varphi,\ a\cos\varphi\, d\lambda,\ dh)^T \tag{4}$$

may be interpreted as shifts of the local plane coordinates (dx, dy) and as ellipsoidal height change dh where a denotes the mean radius of the earth. The covariance matrix $\mathbf{Q}_{dx}$ corresponding to the shift vector $d\mathbf{x}$ follows after applying the law of covariance propagation to Eq. (2) and is given by

$$\mathbf{Q}_{dx} = \mathbf{R}\,\mathbf{Q}_{dX}\,\mathbf{R}^T. \tag{5}$$

3.2 Strain analysis

For modeling regional or local deformations the method of strain analysis is adequate since this approach is almost independent of possible changes in the geodetic datum of data sets for different epochs. The shift $d\mathbf{x}$ of an arbitrary point $\mathbf{x}$ is set identical to the displacement vector $\mathbf{u}(\mathbf{x})$ at this point. For a neighbouring point $\mathbf{x}' = \mathbf{x} + \Delta\mathbf{x}$ the corresponding shift $d\mathbf{x}'$ is expanded into a Taylor series. Considering only the linear term, the relative displacement $\Delta(d\mathbf{x})$ between the two sites is obtained by

$$\Delta(d\mathbf{x}) = d\mathbf{x}' - d\mathbf{x} = \frac{\partial\mathbf{u}(\mathbf{x})}{\partial\mathbf{x}}\Delta\mathbf{x} = \mathbf{u}_x\,\Delta\mathbf{x}. \tag{6}$$

The system is poorly conditioned with respect to the vertical gradients. Thus, strain analysis is restricted to the horizontal components for practical applications. Decomposing the deformation matrix $\mathbf{u}_x$ for the two-dimensional case yields the strain tensor $\mathbf{U}$ and the skew-symmetric axiator $\mathbf{A}$:

$$\mathbf{u}_x = \mathbf{U} + \mathbf{A} = \begin{bmatrix} u_{11} & u_{12} \\ u_{12} & u_{22} \end{bmatrix} + \begin{bmatrix} 0 & a_{12} \\ -a_{12} & 0 \end{bmatrix} . \tag{7}$$

The single element a_{12} of the axiator can be interpreted as the parameter of a differential rotation which may result from datum changes but does not affect the strain.

Defining a vector $\mathbf{p}$ containing the unknown parameters by

$$\mathbf{p} = (u_{11},\, u_{12},\, u_{22},\, a_{12})^T \tag{8}$$

and introducing an auxiliary matrix $\mathbf{H}$ composed by the elements of the position-difference vector $\Delta\mathbf{x} = \mathbf{x}' - \mathbf{x} = (\Delta x, \Delta y)^T$:

$$\mathbf{H} = \begin{bmatrix} \Delta x & \Delta y & 0 & -\Delta y \\ 0 & \Delta x & \Delta y & \Delta x \end{bmatrix} , \tag{9}$$

Eq. (6) for the relative displacement $\Delta(d\mathbf{x})$ can be written as

$$\Delta(d\mathbf{x}) = d\mathbf{x}' - d\mathbf{x} = \mathbf{H}\,\mathbf{p} . \tag{10}$$

3.3 Discussion

Constant parts in the displacements caused by a shift in the geodetic datum are eliminated by differentiating them in Eq. (6). A possible rotation in the (two-dimensional) datum is described by the element a_{12} in the axiator of Eq. (7).

Rigorously, the strain model, cf. Eq. (6), holds only in case of homogeneous deformation (*i.e.*, constant elements of the matrix $\mathbf{u}_x$) or for small areas. But what is a small area?

Considering the relative displacements as pseudo-observations, the desired components of the strain tensor can be derived. A minimum of $n = 3$ points is required and for $n > 3$ the solution follows from least squares adjustment. Consequently, the area under consideration should be subdivided into triangles. In order to avoid misinterpretation, it may be useful to check some of the strain components with the aid of extensometers.

A more precise geodynamic interpretation would result from physical quantities like the stress tensor instead of geometric quantities like the strain tensor. But how to relate the two tensors or, in other words, which rheological model should be used? Direct measurements of some stress components may help to find the correct answer.

So far only the horizontal components of the displacement field have been taken into account. The vertical components reflect ellipsoidal height changes which are

purely geometric quantities and are, therefore, not directly suitable for geodynamical interpretation. For this purpose, GPS must be supplemented by the observation of gravity or by conventional leveling. Here, further details are omitted.

4. Conclusions

In spite of the previous critical remarks, GPS can substantially contribute to geodynamical research, particularly, if some principal issues are taken into account. Each repeated measurement, for example, should be performed in the same configuration and using the same hardware and software as during the zero-epoch measurement.

The accuracy level, presently achievable with GPS, is too low to detect crustal movements after a short time. However, the earlier one starts a project, the sooner statistically significant results can be obtained. As an example for this strategy the project AGEDEN (Austrian Geodynamic Densification Network) is mentioned, cf. Lichtenegger (1990). The first-epoch measurements were performed in the fall of 1987 using the state-of-the-art receivers of that time. These measurements were repeated in the fall of 1990 during the AGREF (Austrian Geodynamic Reference Network) project. A relative accuracy of about 0.1 ppm has been achieved in this campaign which yields accuracies of ± 5 mm for the 50 km baselines along some tectonic active zones in Austria. Thus, significant crustal movements will be detected after 15 years if movement rates of 1 mm per year are expected.

The future capability of GPS will be improved by new technologies in the space segment (*e.g.*, third carrier frequency), by the combination of GPS with other satellite based positioning techniques (*e.g.*, GLONASS), and through innovations in hardware and software development.

Acknowledgment

The author is grateful to Miss Jodi Zeise for proof-reading the manuscript and for valuable comments.

References

Cannon M E, Lachapelle G 1993: GPS – Theory and Applications. Lecture Notes. University of Technology at Graz

Hofmann-Wellenhof B, Lichtenegger H, Collins J 1992: In: Proceedings of the Sixth International Geodetic Symposium on Satellite Positioning, Columbus, Ohio, 1, 336–344.

Hofmann-Wellenhof B, Lichtenegger H, Collins J 1994: GPS – Theory and Practice. Third edition. Springer, Wien, New York

Lichtenegger H 1990: In: Proceedings of the First International Symposium on Gravity Field Determination and GPS-Positioning in the Alps-Adria Area. K Čolić, P Pešec, H Sünkel eds, Mitteilungen der geodätischen Institute der Technischen Universität Graz, 67, 392–403.

Mueller I I 1990: In: Proceedings of the Second International Symposium on Precise Positioning with the Global Positioning System, Ottawa, Canada, 48–66.

Acta Geod. Geoph. Hung., Vol. 29(3–4), pp. 397–420 (1994)

VLBI IN GEODYNAMICAL INVESTIGATIONS

H SCHUH[1], J CAMPBELL[2]

The Very Long Baseline Interferometry (VLBI) is one of the most effective methods in the geodynamics. Firstly the fundamentals of this method are described. Then the applications in geophysics and geodesy are shown together with the most important international geodynamical projects.

The primary objectives of the geodetic VLBI are:

- creation of a quasi-inertial extragalactic reference system
- realisation of a global terrestrial system
- monitoring the Earth rotation
- improvement of coefficients for precession and nutation
- estimation of the elastic parameters of the Earth
- determination of regional and local motions of the Earth's crust and verification of the plate tectonic models
- study of the processes at the plate boundaries with a particular interest to seismological research.

Keywords: crustal motions; geometric model; Earth rotation; nutation; plate tectonics; precession; quasi-inertial extragalactic reference system; terrestrial reference system; VLBI; VLB interferometry

1. Introduction

Very Long Baseline Interferometry was developed as a new observation method for radio astronomy aimed at increasing the limited resolution power of the single dish antennas. The step from cable connected local interferometers to 'very long baseline interferometers' was made in 1967, when the first interference fringes were produced by correlating tapes containing the precisely timed radio signals recorded simultaneously at two distant radio telescopes equipped with independent atomic oscillators. The enormous potential of the newly born VLBI technique for geodetic baseline measurements was soon recognised. In the following two and a half decades the precision of an individual group delay measurement, the prime geodetic VLBI observable improved to the present level of 10 picoseconds, *i.e.* 0.3 centimeter.

The geodetic and geophysical interest in VLBI is based on the use of an inertial reference frame formed by a given set of extremely compact extragalactic radio sources. VLBI measures very accurately the angles between the Earth-fixed baseline vectors and the space-fixed radio sources. Thus, even the most subtle changes in the baseline lengths and in the angles between the reference systems can be detected. The main geodynamical phenomena such as polar motion, UT1 variations, nutation

[1]DLR, Linder Höhe, D–51147 Köln, Germany
[2]Geodetic Institute, University of Bonn, Nussallee 17, D–53115 Bonn, Germany

1217-8977/94/$ 4.00

and precession, Earth tides, ocean tidal response, and tectonic plate motions can be monitored with unprecedented accuracy.

At present geodetic VLBI may be seen to have matured and regular observing campaigns such as the IRIS Earth rotation monitoring program and the NASA Crustal Dynamics Project have been providing data that show significant evidence of present-day plate motions. The degree of agreement with models derived from geological data is surprising. This suggests that our present understanding of plate tectonics cannot be far away from reality. The Earth rotation data, on the other hand, are clearly showing that the main factors causing the short period variations in the length-of-day have to be related to the solid Earth tides, the ocean tides and to atmospherical processes, primarily the zonal wind pattern.

The future of VLBI looks promising if we consider the growing number of new telescopes having been built in the last years. Soon, the remaining gaps in the global VLBI network will be closed, allowing man's inquisitive mind to probe even deeper into the mysteries of the Earth's interior.

2. Fundamentals of the VLBI technique

The most striking difference between Very Long Baseline Interferometry and other space techniques is that the interferometric observables are obtained a posteriori by the alignment in a processor of the two identical signal streams received at different times at the two telescopes. Thus, the telescopes plus the processor can be seen to embody one instrument and the baseline separating the telescope sites may be called an instrumental calibration constant.

The main elements of a VLBI system are shown in Fig. 1: the radio signals coming from a given source are observed simultaneously at two or more stations at a preselected frequency in the GHz-band, converted to baseband (video frequency range) and recorded (usually in a digitized form) on high data-rate tape. Before recording, the signal streams are provided with precise time information derived from the local frequency standards (atomic clocks, H-maser oscillators). Later, when the tapes are brought together in the playback processor, this permits a phase coherent correlation of the approximately aligned signal streams for a certain interval of time T (coherent integration time). Because the true time lag is still unknown, the correlation is done at a number of different delays (lags) separated by $\Delta\tau = 1/2B_s$ where B_s is the total spanned bandwidth of the observed recording channels. During correlation the data stream from one station is delayed quasi-continuously in such a way that the changing geometric delay τ_g is almost completely compensated for. This gives rise to a rather low residual fringe frequency, the phase of which varies slowly on the scale of a few turns per minute.

The output of the correlator is usually described by the complex crosscorrelation function $R(\tau, t)$, which translates the response of an interferometer system to a radio source, see *e.g.* Moran 1976. In Fig. 2 the aspect of a typical fringe pattern as produced by a single 2 MHz channel interferometer looking at a point source is shown. Here the delay spacing is $1/(2 \cdot 2\ \mathrm{MHz}) = 0.25\ \mu\mathrm{sec}$.

The analysis of the interference fringe pattern yields a wealth of astronomical and

geodetic information. As many as four — albeit partially dependent — observables can be derived from the crosscorrelation function $R(\tau, t)$:

$$\begin{aligned} \Phi(t) &= \text{fringe phase,} \\ A(t) &= \text{fringe amplitude,} \\ f(t) &= \text{fringe rate,} \\ \tau(t) &= \text{group delay.} \end{aligned}$$

Now let us have a closer look at each one of these observables: The actual travelling time τ of a particular wavefront between the antennas at the two sites can be expressed in two ways: either as the delay of a wave group formed by the wide-band signals (group delay) or as the phase difference of a given monochromatic constituent of the signal stream (phase delay). The *group delay* $\tau(t)$ is defined as the derivative of phase versus frequency in the band and can be estimated unambiguously by finding the peak of the crosscorrelation function, which for an ideal square bandpass assumes the form of a $\sin x/x$ function. The group delay observable yields a full baseline solution and therefore plays the most important role in geodetic VLBI.

The *fringe phase* $\Phi(t)$ and the *fringe rate* (fringe frequency) $f(t)$ are obtained from the sine and cosine parts of the crosscorrelation function. Due to the close relationship of Φ and f, these observables are determined simultaneously, either from an ordinary sine wave adjustment (only possible in the case of strong fringes) or using the Fourier transform into the frequency domain. These methods which allow to establish the function $\Phi(t)$ over a certain interval of time (usually the duration of an uninterrupted source scan) are often referred to as "phase tracking" (Thomas 1972). One of the reasons why the phase observable cannot be used in the same way as in the Global Positioning System (GPS) resides in the fact that radio sources are too weak to be detected with an omnidirectional antenna, and therefore a continuous phase tracking of more than one source at a time is impossible (except for very small angular separations).

The *fringe rate* or the delay rate $\dot{\tau}(t)$ ($\dot{\tau}(t)$ is $f(t)$ divided by the observing frequency) describes the relative phase drift between the signals recorded at the two stations induced by the differential Doppler shift due to Earth rotation. This observable is unambiguous, but it is insensitive to the z-component of the baseline vector. Compared to the group delay, the delay rate plays a less important role in geodetic baseline determinations.

The *fringe amplitude* $A(t)$ is essential only for the mapping of source structure, but in terms of the SNR it is used to estimate the instrumental delay noise. However, this should not be mistaken as group delay precision.

The precision of the group delay t as it is estimated from the crosscorrelation function depends on the SNR (height of the main peak above the noise), but it depends to a much higher degree on the halfwidth of the main peak, which is given by $1/B_s$. Therefore efforts have been concentrated on reducing the width of the main peak by increasing the total spanned bandwidth B_s. In view of the limited recording bandwidth, the synthesis method developed by A E E Rogers, Haystack Observatory, Massachusetts, USA in the late sixties constitutes a major breakthrough for

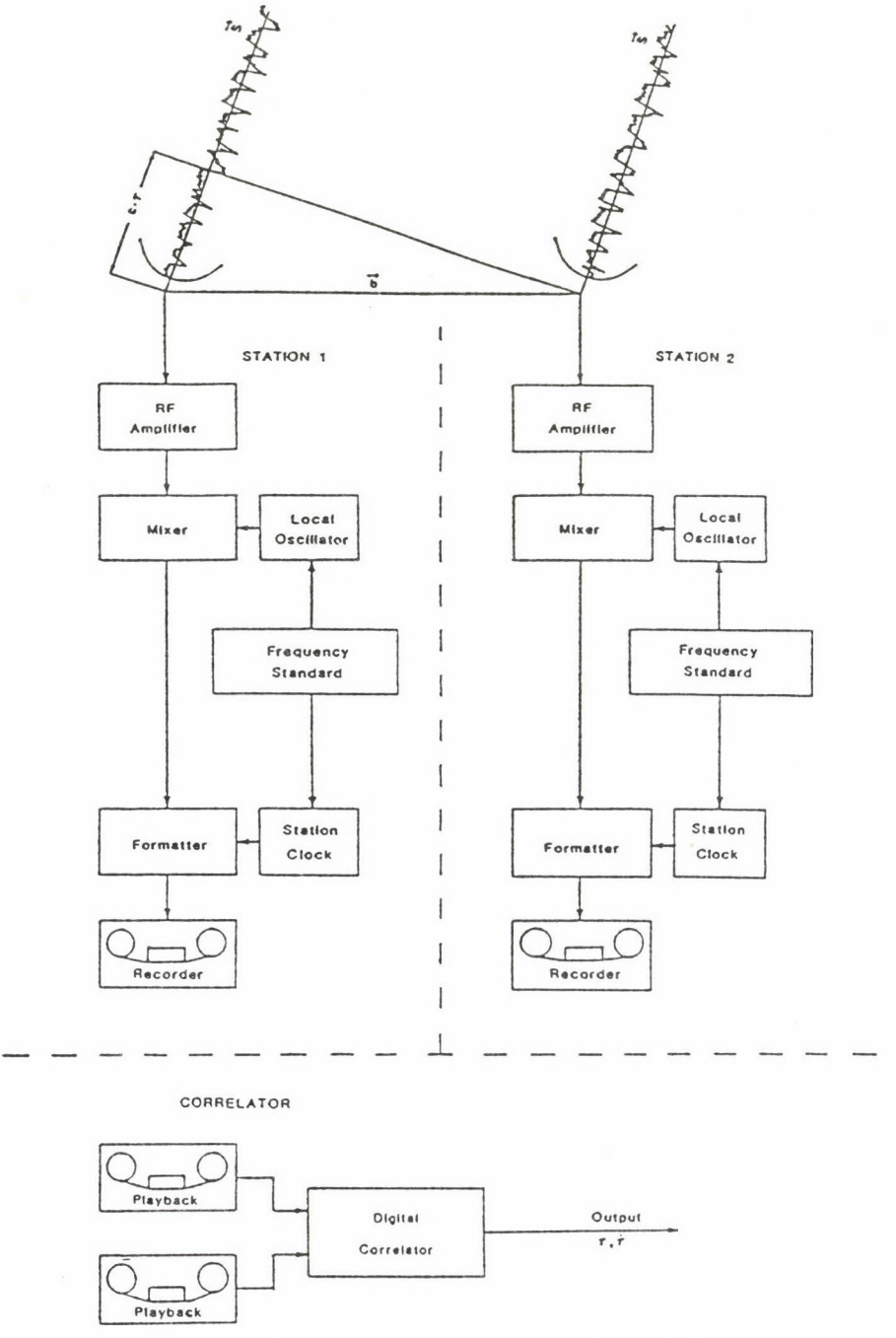

Fig. 1. Functional diagram of a VLBI system

Table I. MkIII frequency setup for geodetic VLBI

X-Band	S-Band
8210.99 MHz	2217.99 MHz
8220.99 MHz	2222.99 MHz
8250.99 MHz	2237.99 MHz
8310.99 MHz	2267.99 MHz
8420.99 MHz	2292.99 MHz
8500.99 MHz	2302.99 MHz
8550.99 MHz	
8570.99 MHz	

the realisation of high precision geodetic VLBI (Rogers 1970). *Bandwidth synthesis* is achieved by splitting the total recorded bandwidth into smaller units and distributing them over the much wider receiver bandwidth window. With the third generation recording system, the MkIII developed at the Haystack Observatory, it is possible to record a data stream of 112 Megabit per second on 28 parallel tracks of tape (Clark et al. 1985). In the standard MkIII setup for geodetic VLBI (see Tab. I) the X-band frequency of 8.4 GHz has a spanned bandwidth of 360 MHz with 8 channels of 2 MHz spread over the entire window. In this case, the width of the crosscorrelation peak becomes as small as 1/360 MHz, *i.e.* 2.8 nsec or about 1 meter. A typical example of a multichannel delay resolution function with sidelobes is shown in Fig. 3.

The fringe analysis, which includes a fine delay estimation algorithm (Whitney 1976), allows to determine the group delay to better than 1 % of the halfwidth of the main peak, corresponding to 0.02 to 0.03 ns (20 to 30 picoseconds) or 0.7 to 1.0 cm. This figure depends on the signal-to-noise ratio (SNR) achieved with a scan of a given duration on one particular source. After averaging the correlated signal samples over a total of $2 \cdot B_s \cdot T$, *i.e.* the number of registered bits within T, the SNR can be expressed by

$$SNR = \eta S/2k \cdot \sqrt{A_1 \cdot A_2}/T_s \cdot \sqrt{2B_s T}\,, \tag{1}$$

where k is Boltzmann's constant.

This expression shows that, apart from the fundamental relationship of $\Delta\tau = 1/B_s$, the delay precision is proportional to the flux density S of a point source (in Jansky), the loss factor η due to digitization, the geometric mean of the antenna apertures A_1 and A_2, the square root of the recorded bandwidth B_s and integration time T, and inversely proportional to the geometric mean of the system (receiver) noise temperatures T_s at both stations.

Efforts to improve the sensitivity of a Very Long Baseline Interferometer have been concentrated on the sampling rate $2B_sT$. By doubling the X-band spanned bandwidth and by further technical improvements, the group delay precision could be increased to $\sim$ 13 psec or 0.4 cm (Ray and Corey 1991) at the beginning of the 1990s. The high density recording system MkIIIA allows a more than tenfold higher data rate. It will be succeeded by the next generation VLBI system MkIV in the near future (Whitney et al. 1991).

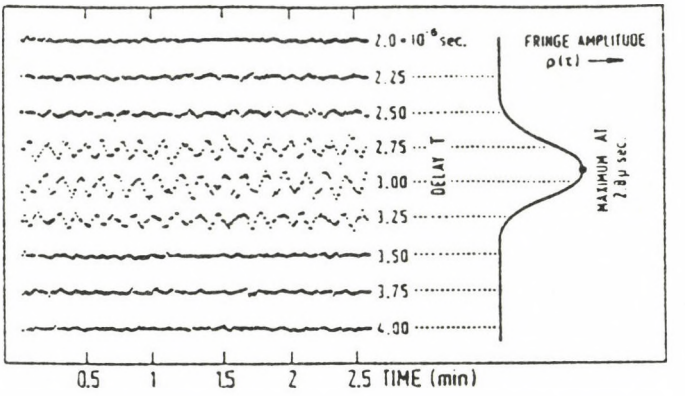

Fig. 2. Response of the MkII VLBI system to a point source

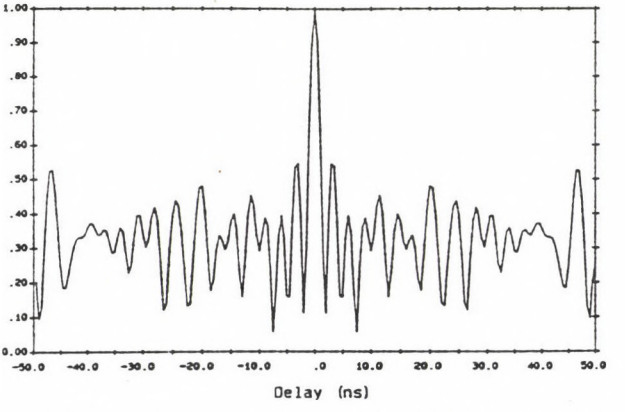

Fig. 3. Multichannel delay resolution function (MkIII system, X-band)

The available coherent integration time T depends on the stability of the frequency standards and on the state of the atmosphere. Due to the latter the ultimate achievable phase stability is limited to around 10^{-15} over time scales of $10^2 - 10^4$ seconds. Hydrogen maser frequency standards reach stabilities in the order of 10^{-15} over the same periods of time, which is acceptable for most applications.

Using expression (1) for the SNR, the instrumental phase error

$$\sigma_\Phi = (SNR)^{-1} \tag{2}$$

and the group delay error

$$\sigma_t = (2\pi SNRB_s)^{-1} \tag{3}$$

can be computed. To illustrate these expressions, a typical example for a VLBI system consisting of two 20 m antennas with 50 % efficiency each and equipped with MkIII data acquisition terminals is given:

observing frequency: 8.4 GHz
8 channels, 2 MHz each
correlated flux density of observed radio source:
1 Jansky (1 Jy = $1 \cdot 10^{-26}$ W $\cdot$ sec $\cdot$ m^{-2})
system temperature: 160° K
coherent integration time T: 300 sec
$\Rightarrow$ signal to noise ratio SNR = 18.2 and $\sigma_\Phi = \pm 3°$

with bandwidth synthesis method and a spanned bandwidth of 360 MHz
an 'effective bandwidth' of $B_{\text{eff}} = 140.2$ MHz is obtained and
$\Rightarrow \sigma_t = \pm 0.062$ nsec (= ± 2.1 cm).

By using cooled receivers the average SNR is usually around 50 which yields a σ_t of ± 0.020 nsec or ± 0.7 cm for the standard MkIII frequency setup. As already mentioned above these numbers can be improved by further technical refinements.

3. High precision interferometry

The term high precision interferometry is used here to include all those applications of VLBI that rely on the exploitation of the group delay observable. It is this quantity which allows to determine the 'macroscopic' geometry of the interferometer, *i.e.* the baseline-source geometry that relates the location of the radio telescopes on the revolving Earth to the infinitely distant compact radio sources. These pointlike emitters without proper motion are ideally suited to serve as fixed beacons in the heavens, allowing to monitor even the smallest departures from the computed motions of the receiving stations.

The research fields that profit most from the geometric potential of VLBI are those dealing with the motions of the celestial bodies, in particular the Earth-Moon system, and the orientation and the size of the Earth itself: astrometry and geodesy. These fields are usually meant to imply a much broader area, namely fundamental astronomy and geosciences, such as geodesy, geophysics, oceanography *etc.* The topic of gravitational light deflection is also intimately related to all of these fields, because it forms part of the fundamental model describing the physical reality of VLBI.

The VLBI data analysis model is developed using the knowledge presently available to mathematically recreate, as closely as possible, the situation at the time of observation. Then, a least squares parameter estimation algorithm is used to determine the best values of the quantities to be solved for. Before this process starts, the raw observations have to be cleaned from several systematic effects, which in fact limit the final accuracy of the results. The flow diagram of a typical geodetic VLBI data analysis software package is shown in Fig. 4. The system can be seen to have two main streams, one containing the actual observations which undergo instrumental and environmental corrections, and the other to produce the so-called theoreticals, beginning with the "*a priories*", a set of starting values for the parameters to be estimated. Both streams converge at the entrance to the least squares algorithm, where the "observed minus computed" are formed.

The systematic *instrumental effects* include clock instabilities, electronic delays

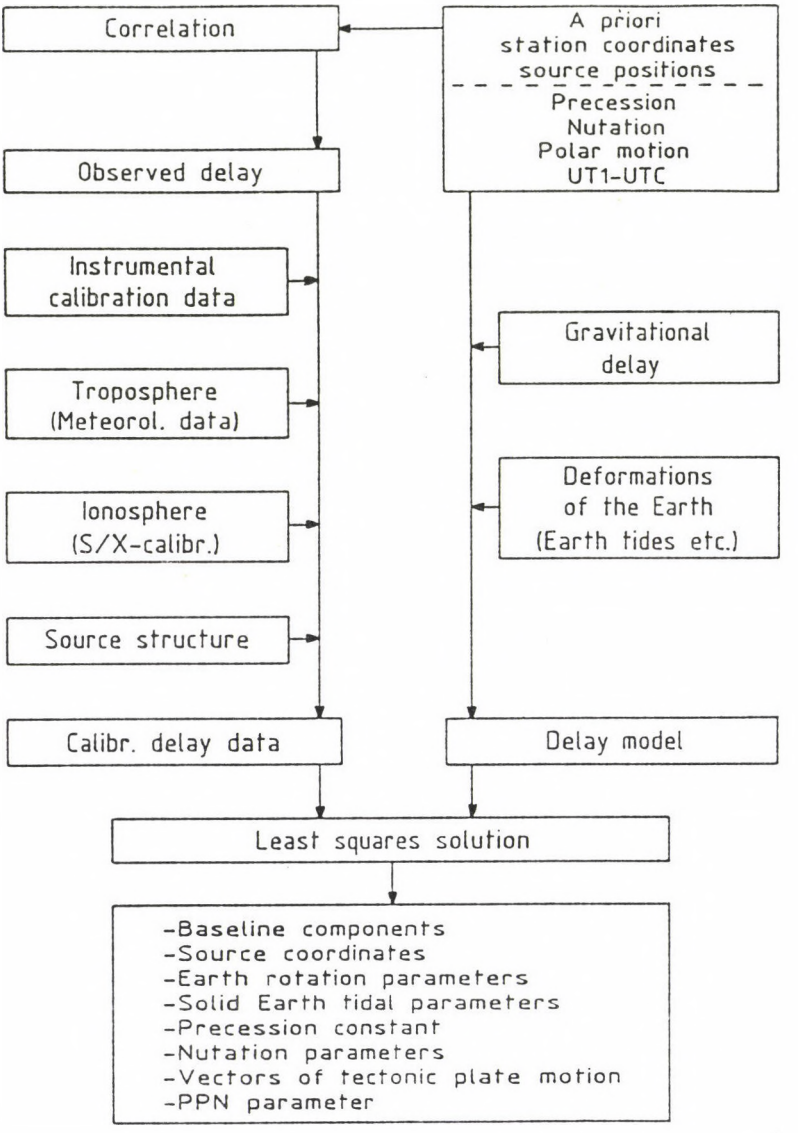

Fig. 4. Flow diagram of a geodetic VLBI software system

in cables and circuitry and deformations of the telescope structure. As a clock model usually second order polynomials are used and occasional breaks have to be introduced. Clock modelling is still very much an interactive procedure and belongs to the editing session. In present standard VLBI solutions the clock estimation algorithm is designed to model short-term, random clock variations while enforcing realistic physical constraints on continuity and rates of change. When all clocks are 'well behaved' a typical algorithm which is applied is as follows: the clock at

one site is designated the reference clock and the differences between this clock and the other site clocks are modeled. These differences are modeled as the sum of two functions: a second order polynomial and a continous, piecewise-linear function with an initial value of zero. The three coefficients of the polynomial correspond to clock epoch offset, clock frequency offset, and clock frequency drift. They are unconstrained in the solution because these parameters can be of any size for real hydrogen masers. In the piecewise-linear function, the offset at the end of each linear segment is estimated. In the solutions done by the NASA/GSFC VLBI group, the linear segments are only one hour long each (Ryan et al. 1993) whereas European VLBI experts usually chose longer segments of for instance six hours (Nothnagel and Campbell 1993).

The instrumental delay changes are monitored by the phase and delay calibration system which is part of the MkIII system. In the telescope the distance between the feed horn and the axis intersection which constitutes the baseline reference point is assumed to be constant at the mm-level. In this case it becomes part of the clock offset parameter. An axis offset model is applied to each antenna where the pointing axes do not intersect. Large telescopes such as the Effelsberg 100 m antenna exhibit elevation dependent changes in the focal distance which can however be modeled to a level of a few millimeters (Rius et al. 1987).

The effect of the atmosphere on VLBI observations is still considered to be the most serious problem, because at widely separated stations the elevation angles of the telescopes pointing to the same source differ greatly as well as the meteorological conditions themselves. The *ionosphere*, which is a highly dispersive medium in the radio frequency band, can be dealt with to first order by using two different observing frequencies. In geodetic VLBI the frequency pair of $f_s = 2.3$ GHz (S-band) and $f_x = 8.4$ GHz (X-band) is used throughout. The ionospheric group delay corrections for the X-band observations are computed from the differences of group delay measurements on X-band and S-band:

$$\Delta\tau_x^{\mathrm{ion}} = (\tau_x - \tau_s) f_s^2 / (f_x^2 - f_s^2) \,. \tag{4}$$

In contrast to GPS, where a very close frequency pair has been chosen, in VLBI the factor to convert the difference into a correction for the higher band is very small: 0.081, so that there is no appreciable error contribution from the S-band observations.

The neutral atmosphere, essentially the *troposphere*, presents the same problems in VLBI as in GPS observations. Its influence on radio signals adds up to an extra zenithal path of 1.8 to 2.5 meters. The contribution of the dry part is rather stable, although care has to be taken to choose a proper mapping function for the lower elevation angles (Davis et al. 1985, Niell 1993). The wet component, although the smaller part of the total tropospheric effect, changes rapidly and can also be monitored by some means. The most promising — albeit costly — method appears to be the water vapor radiometer (WVR) technique, which consists of measuring the microwave thermal emission from water vapor near 22 GHz in the line-of-sight (Elgered et al. 1982).

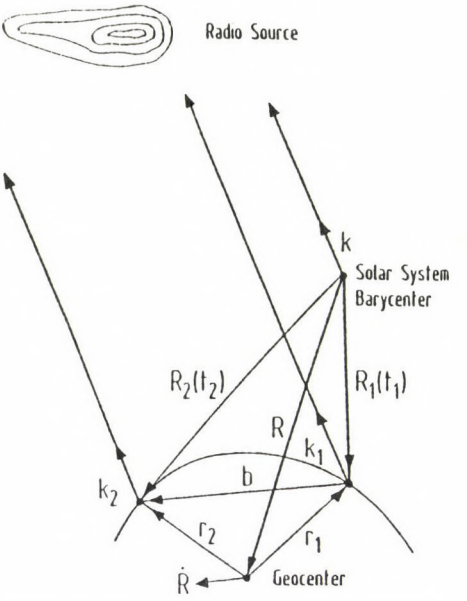

Fig. 5. Geometric VLBI Model

Now let us turn to the model side of the geodetic VLBI analysis: The fundamental observation equation relating the group delay τ to the baseline vector $\mathbf{b}$ and the source vector $\mathbf{k}$ may be written in its simplest form:

$$\tau(t) = -\mathbf{b} \cdot \mathbf{k}(t)/c \tag{5}$$

where $\mathbf{b} \cdot \mathbf{k}(t) = b_x \cos\delta \cos h(t) + b_y \cos\delta \sin h(t) + b_z \sin\delta$
with the geocentric baseline components b_x, b_y, b_z
the radio source positions α, δ
and the Greenwich hour angle $h(t)$ of the source $h(t) = \text{GST} - \alpha$
GST - Greenwich sidereal time.

The baseline vector components b_x, b_y, b_z are referred to the instantaneous Earth rotation axis. The unit vector of the source $\mathbf{k}$ points to the apparent position at the time of observation (see Fig. 5). The negative sign in Eq. (5) accounts for the fact that the motion of the incoming wave front is opposite to the direction of the unit vector $\mathbf{k}$ and the time delay τ is defined $\tau = t_2 - t_1$.

At this stage there are $3 + 2 \cdot n$ fundamental parameters to be determined in a least squares fit: the three baseline components b_x, b_y, b_z and the coordinates α, δ of n observed radio sources. Due to the fact that the offset between the clocks at both stations is not known to better than around 100 ns, clock parameters (minimum one clock offset and one clock rate) have to be added to the solution as explained in the previous section. Therefore, minimal solutions are possible only with observations on at least three epochs and to at least two different sources. Usually a set of 12 to 18 sources spread over the sky as evenly as possible is used in a schedule of 24 hours

during which these sources are observed in an interleaved mode in order to obtain stable geometric conditions of the solution. On new strategies for an optimised VLBI observing schedule see Steufmehl (1993).

In view of the extremely high precision inherent to VLBI, the modeling accuracy has to be brought down to better than a few millimeters on the global scale. Great efforts have been made to develop comprehensive geodetic VLBI data analysis software systems, which include all aspects of the multi-faceted reality of VLBI. Now we will take a summary look at the most important model components.

The *fundamental geometric model* of the time delay τ_g forms the heart of the system. This model has evolved from its basic form in a geocentric system to the fairly complex relativistic formulation in the solar system barycenter (SSB). Initially the basic model of Eq. (5) was corrected for the so-called retarded baseline effect (Cohen and Shaffer 1971), which accounts for the finite travel time of the signals between reception at the two telescopes on the revolving Earth. In spherical astronomy this effect is known as diurnal aberration and in fact it turns out that to first order the application of the retarded baseline effect is equivalent to correcting the source vector $\mathbf{k}$ at station 2 for diurnal aberration. This latter approach is evident in a more rigorous elementary formulation given by Thomas (1972).

The relativistic formulation includes both the effects of special relativity (SRT) and of general relativity (GRT), but for reasons of practicality these are treated separately and added together on the level of the time delay (see Preuss and Campbell 1992):

$$\tau_g = \tau(SRT) + \tau(GRT) \tag{6}$$

with $\tau(SRT)$

$$\begin{aligned}\tau_g = {} & \tau_0[1 - (\dot{\mathbf{R}} + \dot{\mathbf{r}}_2)\cdot\mathbf{k}]/c + \tau_0[(\dot{\mathbf{R}}\cdot\mathbf{k})^2 + 2(\dot{\mathbf{R}}\cdot\mathbf{k})(\dot{\mathbf{r}}_2\cdot\mathbf{k})]/c^2 + \\ & +(\mathbf{b}\cdot\dot{\mathbf{R}})[(\dot{\mathbf{R}}\cdot\mathbf{k})/2 + (\dot{\mathbf{r}}_2\cdot\mathbf{k})]/c^3 - \tau_0(U + \dot{R}^2/2 + \dot{\mathbf{R}}\cdot\dot{\mathbf{r}}_2)/c^2 - \\ & -(\mathbf{b}\cdot\dot{\mathbf{R}})/c^2\end{aligned} \tag{7}$$

and $\tau(GRT)$ which is computed as a sum of the influences of all gravitating bodies which are close to the signal path in particular the Sun and the Earth itself

$$\tau^{\odot}_{\text{grav}} = \frac{(1+\gamma)r^{\odot}}{c}\ln\left[\frac{R_1 + \mathbf{R}_1\cdot\mathbf{k}}{R_2 + \mathbf{R}_2\cdot\mathbf{k}}\right] \tag{8}$$

$$\tau^{\oplus}_{\text{grav}} = \frac{(1+\gamma)r^{\oplus}}{c}\ln\left[\frac{r_1 + \mathbf{r}_1\cdot\mathbf{k}}{r_2 + \mathbf{R}_2\cdot\mathbf{k}}\right] \tag{9}$$

where $r^{\odot}$ and $r^{\oplus}$ are the Schwarzschild radii of the Sun and the Earth. For the other bodies of the solar system the corresponding Schwarzschild radii and the vectors from these bodies to the VLBI antennas have to be used in the above formula (8) if necessary, *i.e.* if the direction to the observed radio source is close to that body.

The effects of *special relativity* arise from the fact that quantities defined in coordinate frames moving relative to each other have to be related by transformations of the Lorentz type with v/c^2 as the characteristic quantity in the time delay correction terms. The choice of two particular coordinate systems (the SSB and the

geocentric system) used to describe the VLBI model arises from practical considerations: the motions of bodies in the solar system and the positions of the radio sources are most readily defined in the SSB, while the actual baselines between the telescopes are usually required in a geocentric system. The vector $\dot{\mathbf{R}}$ describes the velocity of the geocenter with respect to the SSB ($\sim$ 30 km/sec) and $\dot{\mathbf{r}}^2$ is the velocity of the station 2 with respect to the geocenter ($<$ 0.46 km/sec) (Fig. 5). In the geometric VLBI model these velocities have to be computed with an accuracy of 10^{-6}. The last two terms in Eq. (7) account for the difference in SSB coordinate time and the geocentric proper time as well as for the fact that the station clocks are located at fixed points on the Earth's crust. Here, U is the magnitude of the gravitational potential of the solar system at the geocenter.

The *effect of gravity* on the propagation of electromagnetic waves (GRT) is no less important. According to GRT, space-time is deformed by the presence of masses. The most massive object in our vicinity is of course the Sun, which accounts for more than 99 % of the total effect. Even at an angle of 90 deg away from the Sun the differential delay effect for a 6000 km baseline is still 0.56 nsec (Table II). However, at the present level of accuracy of VLBI the major planets also contribute a bending effect which cannot to be entirely neglected. If Jupiter arrives within a few degrees of an observed source, its influence on ray bending has to be taken into account as can be seen from Table II. Another small but significant contribution ($<$ 20 psec) comes from the gravity field of the Earth itself according to Eq. (9).

Since the early 80's VLBI observations have been used extensively to verify Einstein's theory in its Parameterized Post Newtonian formulation (PPN). Two approaches have been used, one designing special experiments to observe sources such as 3C279 and 3C273 during their close approach to the Sun and the other using all available data from routine geodetic experiments to achieve the accuracy by the sheer number of the observations. The γ-factor, which in the Einstein theory should be equal to unity, has been found to show no significant departure from this value to the level of 0.1 % (Carter et al. 1985). Recently this accuracy level has been further improved to 0.02 % (Robertson et al. 1991). Attempts have also been made to verify the gravitational bending near Jupiter (Schuh et al. 1988), but the effect is only marginally significant ($<$ 100 psec) at close encounters, *i.e.* less than a few arcmin (Campbell 1989, Treuhaft and Lowe 1991).

In October 1990 a workshop was held at U.S. Naval Observatory to bring the VLBI model builders (mostly theoretical relativists) and the model users (mostly geodesists with little experience in relativity) together. As an output of this workshop a so-called *consensus model* to guarantee picosecond delay accuracy was obtained for the geodetic VLBI observables (Eubanks 1991).

The description of the Earth's orientation with respect to the celestial system (*precession, nutation*), the motion of the Earth's axis with respect to the crust (polar motion) and the phase angle of the Earth's rotation (expressed by UT1-UTC) have to reach the same level of accuracy as all the other model components, which means roughly 0.0002 arcsec (0.2 milliarcsecond). Although models are used to calculate *a priori* the periodic variations of the Earth rotation parameters (ERPs) due to the Earth tides (Yoder et al. 1981, Tamura 1993) and due to the ocean tides (Brosche

Table II. Gravitational path delay as a function of spherical distance Θ from the Sun and Jupiter. Given are maximum values for a 6000 km baseline (Schuh 1987)

Θ (Sun) [°]	τ^S_{grav} [ns]	Θ (Jupiter) [°]	τ^J_{grav} [ns]
0.267	169.52	Rim	1.582
1	45.30	0.017 ($\widehat{=}1'$)	0.605
5	9.06	0.167 ($\widehat{=}10'$)	0.062
10	4.54	0.5	0.021
30	1.53	1	0.010
60	0.79	5	0.002
90	0.56	10	0.001
120	0.46		
150	0.41		
180	0.40		

et al. 1989, Wünsch and Seiler 1992), usually the required accuracies cannot be met without parametrisation. Therefore, with longer series of VLBI experiments, the nutation parameters in longitude and in obliquity and the components of polar motion x_p, y_p plus UT1 are included as parameters in the least squares solution. Also, precession can be solved for if longer time spans of data are analysed.

Periodic and aperiodic deformations of the Earth's crust have to be taken into account as well. Solid *Earth tides* show diurnal and semidiurnal oscillations which cause vertical deformations in a range of ± 20 cm and horizontal displacements of about 30 % of the vertical effect. Good models are available, but the relevant parameters (the Love numbers) can also be estimated from larger sets of data (Mitrovica et al. 1994). More difficult to model are the tidal loading effects of the oceans ('*ocean loading*'), which amount to as much as a decimeter on some coastal or island sites (Scherneck 1991). By detailed analyses of VLBI measurements even these effects could be revealed in the data (Schuh and Mühlmann 1989, Sovers 1994). The loading effects due to air pressure variations ('*atmospheric loading*') also reach the level of significance in VLBI modeling (Rabbel and Schuh 1986, Macmillan and Gipson 1994).

The theory of *plate tectonics*, which stipulates that the Earth's crust is formed by a mosaic of separate plates that are in motion relative to each other, has now been universally accepted. Predictions derived from geophysical evidence yield motions of a few centimeters per year (Minster and Jordan 1978, DeMets et al. 1990, DeMets et al. 1994). With the large global VLBI data sets including now well over a decade of regular observations, parameterised plate models have been successfully determined in recent VLBI solutions (Ryan et al. 1993, Ray et al. 1994).

A major problem is constituted by the fact that most of the observed 'compact' radio sources tend to show structure at the level of a few mas. These effects, in particular the changes in the structure, pose a limit on the accuracy of the radio reference system. Permanent monitoring of the structure, which is also accomplished by analysing VLBI data, can be done in parallel to the geodetic analysis, thus providing a means to correct for the structure effects (Schalinski et al. 1988, Campbell et al. 1988, Charlot 1993, Zeppenfeld 1993).

In geodetic VLBI data processing there are two levels of *least squares solutions*, one in which only the "local" unknowns are estimated (such as clock and atmospheric parameters for the participating stations, Earth rotation parameters and nutation offsets for each observing session, *etc.*) thus creating a first data base version of each particular experiment, and another which collects all available experiments for a combined solution including the 'global' unknowns such as station and source positions, geodynamical parameters *etc.*

Among the various VLBI software systems the MkIII Data Analysis System has been mentioned already. It is built around the CALC/SOLVE software system developed jointly by the US East Coast VLBI groups. At present CALC version 7.6 is used. This software has become a sort of standard against which the other systems can be compared (Ma et al. 1989). Other software systems at the same level of accuracy are the OCCAM package developed by the European VLBI groups in Bonn and Madrid (Zarraoa et al. 1993) running on a personal computer under MS-DOS and the MASTERFIT/MODEST software developed at Jet Propulsion Laboratory, Pasadena (Sovers 1991).

4. Scientific interest, programs and results

A detailed description of the expected geophysical applications of VLBI has been presented as early as 1969 at a conference held in London, Canada on 'Earthquake displacement fields and the rotation of the Earth' (Shapiro and Knight 1970). In following, virtually all of the goals mentioned there and even more could be achieved. Astrometric as well as geodetic and geophysical interest in VLBI is based on the use of an inertial reference frame of highly compact extragalactic radio sources. With the VLBI technique it is possible to measure very accurately the baseline vectors and their change with time between distant points on the Earth's crust. Therefore the primary objectives to be accomplished by geodetic VLBI are:

- the creation of a quasi-inertial extragalactic reference system as a basis for astrometry to study galactic rotation and to improve the distance scale of the universe. Recent results of stellar astrometry with HIPPARCOS are being tied to the extragalactic reference system,

- the realisation of a global terrestrial reference system in order to satisfy the needs of global geodetic and navigational systems (including spacecraft navigation),

- monitoring the Earth rotation parameters (polar motion and UT1 variations) with the highest possible resolution for a better understanding of the kinematics and dynamics of the 'System Earth', *i.e.* the Earth in space, the effects of the atmosphere and the oceans, and the processes in the Earth's interior,

- the determination of improved coefficients for precession and nutation and the estimation of the Earth's elasticity parameters, thereby also contributing to a more comprehensive understanding of the System Earth,

- the determination of regional and global crustal motions to verify the plate tectonic models and to study the processes at the plate boundaries, with the aim to contribute to earthquake prediction research.

Today, the accuracies required to attain these goals have been demonstrated by thousands of VLBI experiments on baselines connecting all major continents of the globe. The global distribution of most of the VLBI stations which are at present used for geodetic applications is shown in Fig. 6. Some of these stations are still in the progress of being fully equipped for geodetic VLBI.

In order to combine the efforts in different countries around the world in realising the goals mentioned above, several programs of international cooperation have been launched in the decade of the 1980s, among which the following are the most important:

The NASA Crustal Dynamics Project (CDP)

This project as part of a US federal program was established by NASA in October 1979 and was completed in December 1991 (Bosworth et al. 1993). It involved several government agencies for the application of geodetic space techniques for measurements of tectonic plate motion, plate stability, regional crustal deformation and Earth rotation dynamics. Cooperative arrangements were made with European and other countries extending the project to a global research program (NASA 1988). The VLBI part of the CDP comprised regular experiments (several dozens each year) of one to three days duration between the major geodetic VLBI facilities in the US, Europe, Asia and in and around the Pacific Ocean. In addition so-called bursts of observations were carried out each year using the mobile VLBI units to occupy tectonically interesting sites in California, Alaska and Europe. The CDP was succeeded in 1992 by a new program: Dynamics of the Solid Earth (DOSE).

IRIS (International Radio Interferometric Surveying), NAVNET (Navy VLBI Network) and DSN (Deep Space Network)

The aim of the IRIS program is to conduct VLBI observations at regular intervals to monitor the Earth rotation parameters, *i.e.* polar motion and UT1 and to improve the models used for nutation and precession. This observational program which began in 1980 under the acronym of POLARIS have been extended in several steps to comprise three networks, the original IRIS-A (Atlantic) network with three stations in the USA and two in Europe (Wettzell and Onsala), the IRIS-S (South) network with Hartebeesthoek added to IRIS-A and the IRIS-P (Pacific) network combining VLBI stations around the Pacific. Additionally, IRIS-Intensive series take place, which consist of short daily observations on a transatlantic baseline nested between the regular (each 5 days, respectively 7 days) IRIS-A runs. Since 1991 the project IRIS has been complemented by the Navy VLBI Network (NAVNET) which is conducted by the U.S. Naval Observatory, Washington D.C. Both IRIS and NAVNET are now part of the American National Earth Orientation

Fig. 6. Present global VLBI network

Fig. 7. European VLBI network (fixed stations are underlined)

Service (NEOS) and contribute to the International Earth Rotation Service (IERS) in Paris.

To be mentioned here are also the VLBI measurements of the ERPs by the NASA Deep Space Network (DSN). The DSN comprises three telescopes in California, Spain and Australia (Steppe et al. 1992).

European geodetic VLBI

In the years from 1988 Europe has seen a rapid expansion of its geodetic VLBI network with the establishment of several new stations in the Mediterranean area. Since that time one of the world's most densely spaced VLBI networks of fixed stations has carried out regular measurements in order to define a European 'zero order' network and also to determine regional crustal motions in an area where strong seismicity betrays the ongoing interaction of the African and Eurasian plates. Thus, the European VLBI network (Fig. 7) forms the backbone of a comprehensive crustal dynamics program using mobile VLBI, mobile lasers and GPS (Campbell et al. 1993).

All of these programs have profited from a broad international cooperation and have produced impressive results. The Earth rotation data, *i.e.* variations in the rotational speed and irregularities in the path of the instantaneous pole of rotation, have begun to show hitherto unseen phenomena. Examples are the influence of the zonal winds of the atmosphere and departures from the normal state, such as the El Nino events in the southern Pacific. Figure 8 shows the remarkably smooth trace of the instantaneous pole from 1987 to 1990. This data set provides an extraordinary opportunity to investigate new phenomena such as rapid polar motions, relatively short period oscillations of the polar path around the main spiral curve. The spiral

is the result of a superposition of the annual and the Chandlerian component of polar motion. The UT1 variations are dominated by the seasonal terms, a one year and a half year oscillation as can be seen in Fig. 9, but there is also a rich spectrum of tidally induced short period oscillations, which could be detected by VLBI for the first time (Robertson et al. 1985, Campbell and Schuh 1986). Even the short period tidal influence of the oceans can be seen in the VLBI measurements using large numbers of data sets (Brosche et al. 1991).

As mentioned in chapter 3 also the precession and nutation angles can be estimated from the VLBI data (Herring et al. 1986). The deviations of the nutation in longitude and obliquity from the IAU 1980 Model (Wahr's theory for an elastic Earth) show a prominent annual oscillation with an amplitude of about 6 mas (Fig. 10a, b) and further periodicities.

The NASA CDP and the IRIS campaigns have been providing baseline length results since 1979 when the MkIII system became operational. As the length of a baseline vector is independent from changes in its orientation, the baseline length series are free from errors in the Earth rotation parameters. Concerning plate tectonics the baseline length changes detected by VLBI are seen to confirm the plate models to a surprisingly good degree. A prominent example is the 6,000 km baseline Westford to Wettzell which now has a record of more than 10 years of uninterrupted observations and displays a very significant trend (1.72 cm $\pm$ 0.02 cm/year) (Fig. 11) which is in good agreement with the predicted relative tectonic motion of the American and Eurasian plates (NUVEL model rate 1.89 cm/year).

The only areas were some disagreements between the VLBI results and the current plate models occur are the critical plate boundary zones such as the western US (Basin and Range province) and the east coast of Japan. There the influence of the processes associated with the formation and subduction of the crust can be clearly seen (Heki et al. 1990). In these areas densification measurements are needed with mobile VLBI, SLR and GPS to establish the detailed motion picture of the complex boundary zones.

The largest relative motions have been detected on baselines in and around the Pacific: Alaska to Hawaii –4.5 cm/year, Japan to Hawaii –6.3 cm/year and Japan to Kwajalein –7.1 cm/year. This demonstrates dramatically the amount of activity along the East Pacific Ridge, offering the Hawaii and Kwajalein islands a dashing ride of just under 10 meters per century in the northwest direction. In the stable inner parts of the plates motions are small, if non-existant. An almost perfect example of this situation is found on the 920 km baseline between Onsala and Wettzell with a 'slope' of –0.6 $\pm$ 0.2 cm/year, a result that also demonstrates the inherent stability of VLBI as a baseline measuring system and of the telescope reference points with respect to the surrounding areas (Fig. 12).

But while the central and northern part of Europe appears to be rather stable, the Mediterranean area comprises some of the tectonically most active regions of the globe. As a result of the collision and continued northward push of the African plate against Eurasia, the Alpine System was thrown up and the smaller plates of Arabia, Italy and the Iberian peninsula were set on individual courses with expected motions of one to two centimeters per year. Recent results of the European VLBI

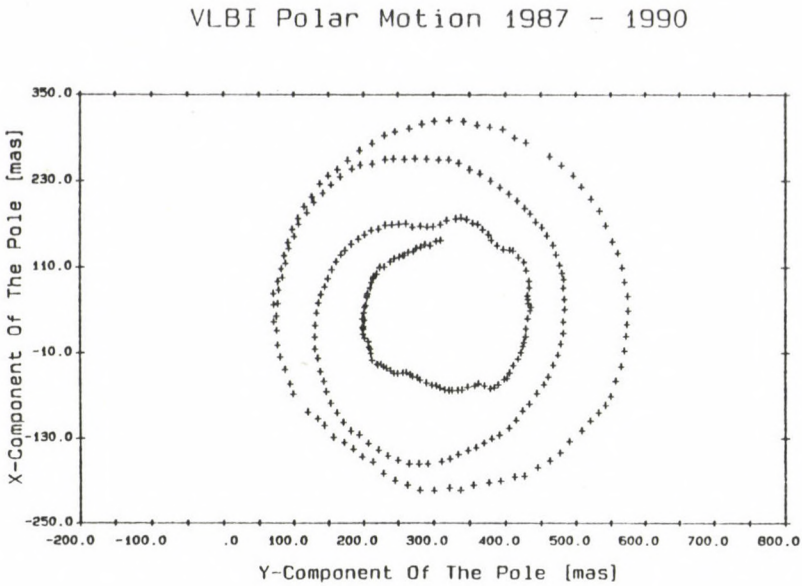

Fig. 8. Polar path as observed by the IRIS VLBI network

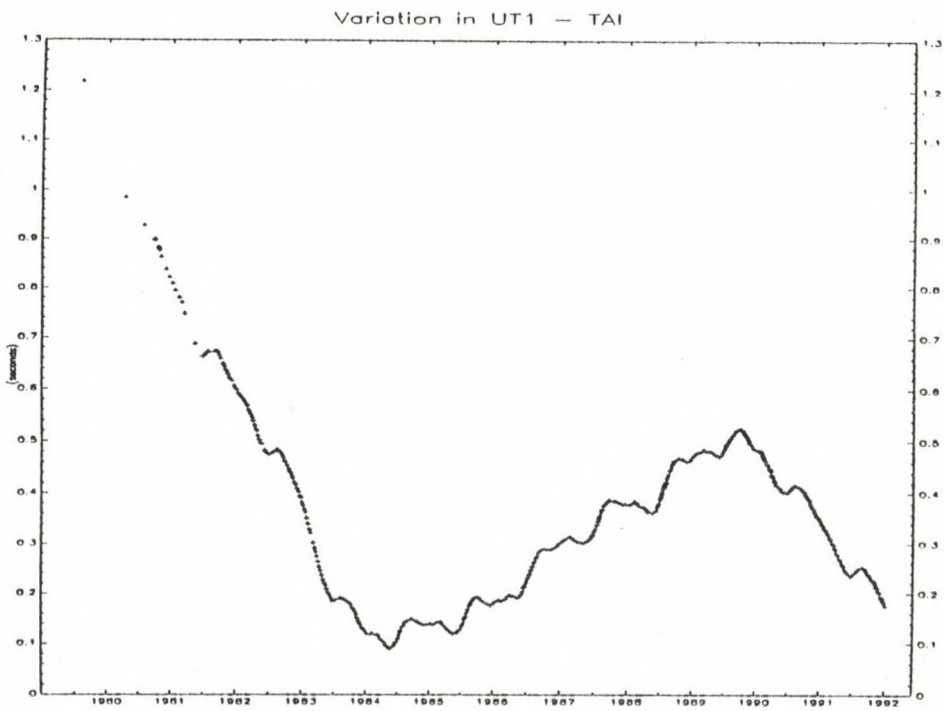

Fig. 9. UT1 variations from IRIS VLBI

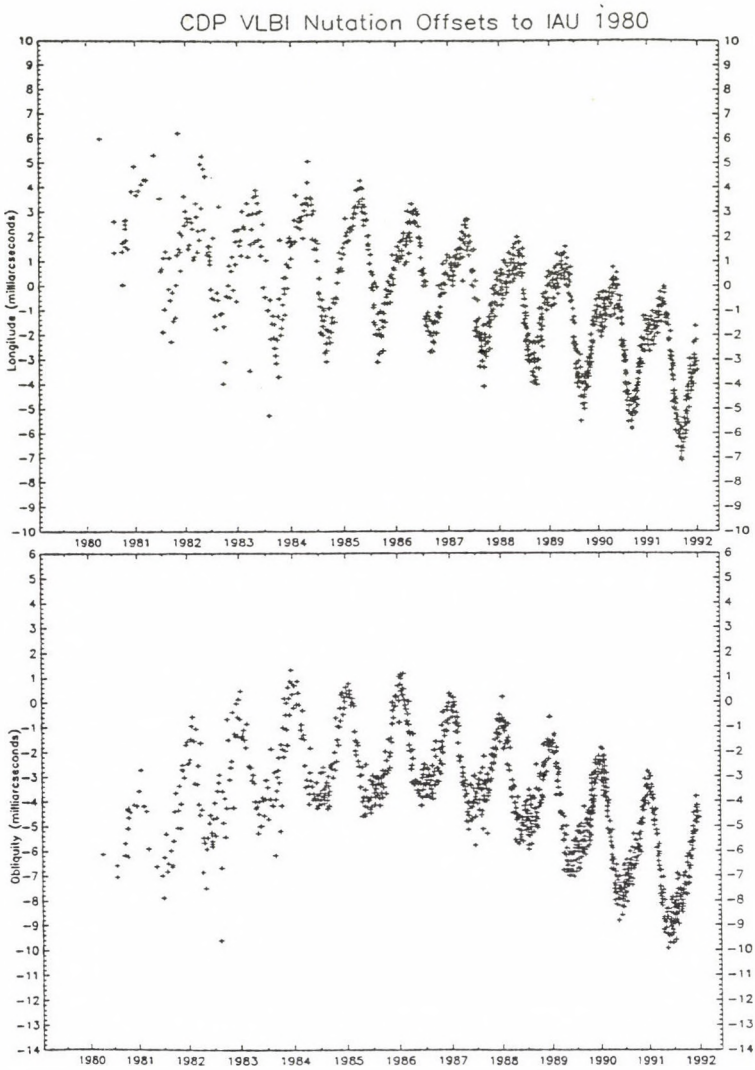

Fig. 10. a, b. Observed nutation corrections to the IAU 1980 model

campaigns have been published by (Campbell et al. 1993, Nothnagel and Campbell 1993, Tomasi 1993) and are included in the final CDP report (Ryan et al. 1993).

The solutions with large data sets are stable enough to produce precise source positions simultaneously with the baseline components and other parameters. The results of several different VLBI solutions are combined to form the celestial system of the International Earth Rotation Service (IERS). Its 1993 realization contains some five hundred radio sources in the declination range $-82°$ to $+86°$. Over 300 objects have positions that are known within ± 0.5 mas (Arias et al. 1993). With the southern stations such as Hartebeesthoek/South Africa, Hobart/Tasmania and

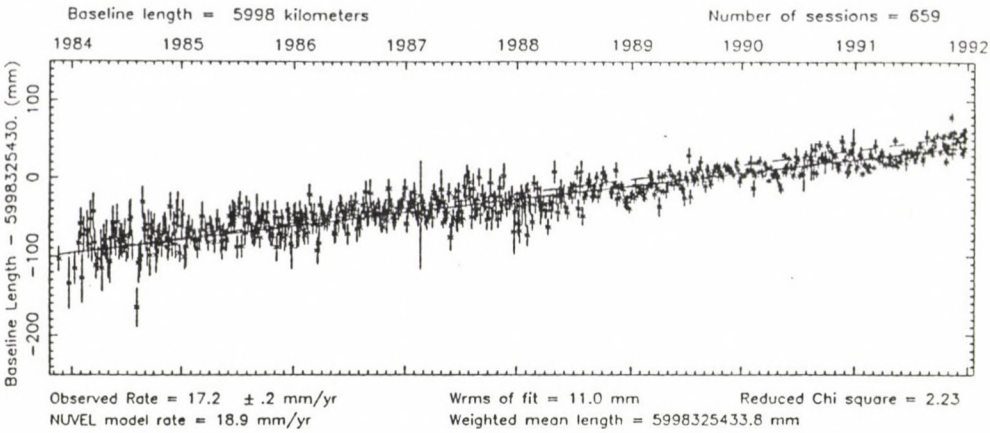

Fig. 11. Transatlantic baseline evolution Westford-Wettzell

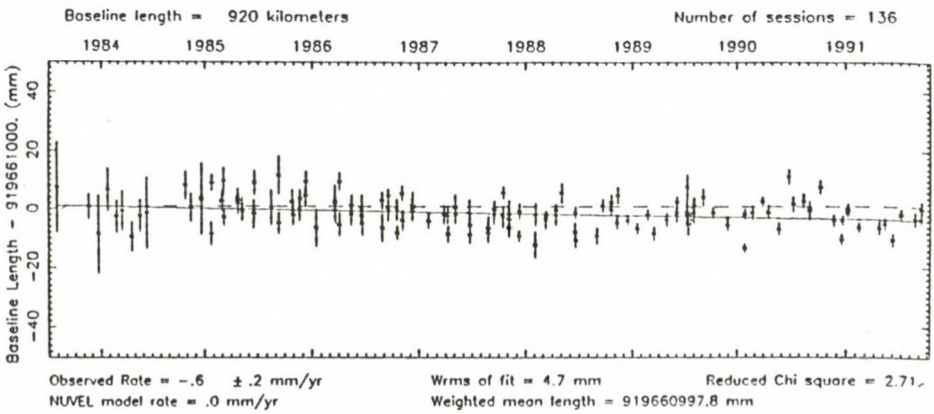

Fig. 12. Baseline evolution in 'stable' Europe: Onsala-Wettzell

O'Higgins/Antarctica the source list is being extended southwards to achieve a uniform coverage of the entire celestial sphere.

In the past two decades geodetic VLBI has grown from its modest beginnings as a mere byproduct of a radio astronomical observing instrument to a dual purpose instrument, where both astronomy and the geosciences have an equal share in a rich harvest of scientific results. With the advent of GPS as a cheaper and more versatile tool for precise relative point positioning, VLBI will retain its importance as the backbone of the inertially based celestial and terrestrial reference frames.

References

Arias E F, Bourquard D, Feissel M 1993: In: Proc. of the 9th Working Meeting on European VLBI for Geodesy and Astrometry, Bad Neuenahr, 1993, Mitteil. aus den Geodätischen Instituten der Rhein. Friedr.-Wilh.-Univ. Bonn, Nr. 81, 140–149.

Bosworth J M, Coates R J, Fischetti T L 1993: In: Contributions of Space Geodesy to Geodynamics: Technology. AGU Geodynamics Series, Vol. 25, 1–20.

Brosche P, Seiler U, Sündermann J, Wünsch J 1989: *Astronomy and Astrophysics*, 220, 318–329.

Brosche P, Wünsch J, Campbell J, Schuh H 1991: *Astronomy and Astrophysics*, 245, 676–682.

Campbell J 1989: In: Systémes de référence spatio-temporels, N Capitaine ed., Observatoire de Paris, 55–63.

Campbell J, Schuh H 1986: In: Proc. Tenth International Symp. on Earth Tides, Madrid, 1985, R Vieira ed., C.S.I.C., 943–951.

Campbell J, Schuh H, Zeppenfeld G 1988: On the Computation of Group Delay Corrections Caused by Radio Source Structure. IAU-Symposium No. 129 "The Impact of VLBI on Astrophysics and Geophysics", M J Reid and J M Moran eds, Kluwer Academic Publ., Dordrecht, p. 427.

Campbell J et al. 1993: In: Contributions of Space Geodesy to Geodynamics: Crustal Dynamics. AGU Geodynamics Series, Vol. 23, 397–406.

Carter W E, Robertson D S, MacKay J R 1985: *J. Geophysical Research*, 90, 4577–4587.

Carter W E, Robertson D S, Nothnagel A, Nicolson G, Schuh H, Campbell J 1988: IRIS-S: *J. Geophys. Research*, 93, 14947–14953.

Carter W E (Convenor) 1991: Proceedings of the AGU Chapman Conference on Geodetic VLBI: Monitoring Global Change held at Washington, D.C. 1991. NOAA Technical Report NOS 137 NGS 49, US Dept. of Commerce, NOAA/NOS, Rockville, MD

Charlot P 1993: In: Proc. of the 9th Working Meeting on European VLBI for Geodesy and Astrometry, Bad Neuenahr, 1993, Mitteil. aus den Geodätischen Instituten der Rhein. Friedr.-Wilh.-Univ. Bonn, Nr. 81, 171–178.

Clark T A et al. 1985: Precision Geodesy using the Mk-III Very-Long-Baseline Interferometer Sy- stem. IEEE Transactions on Geoscience and Remote Sensing, Vol. GE-23, No.4, 438–449.

Davis J L et al. 1985: *Radio Science*, 20, 1593–1607.

DeMets C, Gordon R G, Argus D F, Stein S 1990: *Geophys. Journ.*, 101, 425–478.

DeMets C, Gordon R G, Argus D F, Stein S 1994: *Geophys. Res. Letters* (in press)

Elgered G, Rönnäng B, Askne J 1982: *Radio Science*, 17, 1258–1264.

Eubanks T M 1993: In: Contributions of Space Geodesy to Geodynamics: Earth Dynamics. AGU Geodynamics Series, Vol. 24, 1–54.

Eubanks T M, Steppe J A, Dickey J O 1988: In: Proc. IAU-Symposium No. 128 "The Earth's Rotation and Reference Frames for Geodesy and Geodynamics", A Babcock and G Wilkins eds, D. Reidel, Dordrecht

Eubanks T M ed. 1991: Proceedings of the U.S. Naval Observatory Workshop on Relativistic Models for Use in Space Geodesy. U.S. Naval Observatory, Washington DC

Heki K, Takahashi Y, Kondo T 1990: *Tectonophysics*, 181, 113–122.

Herring T A, Gwinn C R, Shapiro I I 1986: *J. Geophysical Research*, 91, 4745–4754.

MacMillan D S, Gipson J M 1994: Atmospheric Pressure Loading Parameters from VLBI Observations. Preprint

Ma C, Ryan J W, Caprette D 1989: Crustal Dynamics Project Data Analysis-1988, VLBI Geodetic Results 1979-87. NASA Technical Memorandum 100723, Greenbelt, Md.

Minster J B, Jordan T H 1978: *J. Geophysical Research*, 83, 5331–5354.

Mitrovica J X, Davis J L, Mathews P M, Shapiro I I 1994: *Geophys. Res. Letters*, 21, 705–708.

Moran J M 1976: In: Methods of Instrumental Physics, Vol. 12, Part C, Radio Observations, Academic Press, New York

NASA 1988: NASA Geodynamics Program, Summary Report: 1979-1987, Progress and Future Outlook. NASA Technical Memorandum 4065.

Niell A E 1993: In: Proc. International Workshop for Reference Frame Establishment and Technical Development in Space Geodesy, CRL, Tokyo, 61–68.

Nothnagel A, Campbell J 1993: In: Proc. of the 9th Working Meeting on European VLBI for Geodesy and Astrometry, Bad Neuenahr, 1993, Mitteil. aus den Geodätischen Instituten der Rhein. Friedr.-Wilh.-Univ. Bonn, Nr. 81, 42–48.

Preuss E, Campbell J 1992: Very-Long-Baseline Interferometry in Astro-, Geo-, and Gravitational Physics. Lecture Notes in Physics, J Ehlers and G Schäfer eds, Springer-Verlag, 100–130.

Rabbel W, Schuh H 1986: *Journal of Geophysics*, 59, 164–170.

Ray J R, Corey B E 1991: In: Proc. AGU Chapman Conference on Geodetic VLBI: Monitoring Global Change. NOAA Technical Report NOS 137 NGS 49, US Dept. of Commerce, NOAA/NOS, Rockville, MD

Ray J R et al. 1994: NOAA Earth Orientation and Reference Frame Results Derived from VLBI Observations: 1994 Analysis Procedures. subm. to 1993 IERS Annual Report, Paris

Rius A, Rodriguez J, Campbell J 1987: Geodetic VLBI with Large Antennas. Mitt. Geod. Inst. Univ. Bonn, No. 72, J Campbell and H Schuh eds, 59–67.

Robertson D S, Carter W E, Campbell J, Schuh H 1985: *Nature*, 316, 424–427.

Robertson D S, Carter W E, Dillinger W H 1991: *Nature*, 349, 768–770.

Rogers A E E 1970: *Radio Science*, 5, 1239–1248.

Ryan J W, Ma C, Caprette D S 1993: NASA Space Geodesy Program — GSFC Data Analysis – 1992, Final Rep. of the CDP VLBI Geodetic Results 1979–91, NASA Techn. Memorandum 104572

Schalinski C J, Alef W, Campbell J, Witzel A, Schuh H 1988: Results of the VLBI-Investigation of Sources from Geodetic IRIS-Experiments. IAU-Symposium No. 129 "The Impact of VLBI on Astrophysics and Geophysics", M J Reid and J M Moran eds, Kluwer Academic Publ., Dordrecht

Scherneck H-G 1991: *Geophys. J. Int.*, 106, 677–694.

Schuh H 1987: Die Radiointerferometrie auf langen Basen zur Bestimmung von Punktverschiebungen und Erdrotationsparametern. DGK Reihe C, Heft Nr. 328

Schuh H, Mühlmann L 1989: *Geophysical Research Letters*, 16, 1105–1108.

SchuhH et al. 1988: *Physics Letter A*, 129, 299–300.

Shapiro I I, Knight C A 1970: In: Earthquake Displacement Fields and the Rotation of the Earth, L Mansinha, D E Smylie and A E Beck eds, Springer, New York

Sovers O J 1991: Observation Model and Parameter Partials for the JPL VLBI Parameter Estimation Software "MODEST" - 1991. JPL Publication 83-39, Rev. 4.

Sovers O J 1994: *Geophysical Research Letters*, 21, 357–360.

Steppe J A, Oliveau S H, Sovers O J 1992: Earth Rotation Parameters from DSN VLBI, 1992. JPL Geodesy and Geophysics Preprint No. 217, 2–44.

Steufmehl H 1993: In: Proc. of the 8th Working Meeting on European VLBI for Geodesy and Astrometry, Dwingeloo 1991, pp. IV-22 - IV-29, Rep. MDTNO-R-9243, Surv. Dep. of Rijkswaterstaat, Delft

Tamura Y 1993: Periodic Series of DUT1. Geodesy and Physics of the Earth, Montag and Reigber eds, Springer-Verlag, 435-438.

Tomasi P 1993: In: Proc. of the 9th Working Meeting on European VLBI for Geodesy and Astrometry, Bad Neuenahr, 1993, Mitteil. aus den Geodätischen Instituten der Rhein. Friedr.-Wilh.-Univ. Bonn, Nr. 81, 194–198.

Treuhaft R N, Lowe S T 1991: *Astronomical Journal*, 102.

Thomas J B 1972: An Analysis of Long Baseline Radio Interferometry. Deep Space Network Progress Report, Technical Report 32-1526, Vols. VII, VIII, XVI, Jet Propulsion Laboratory, Pasadena

Whitney A R et al. 1976: *Radio Science*, 11, 421–432.

Whitney A R et al. 1991: In: Proc. AGU Chapman Conference on Geodetic VLBI: Monitoring Global Change. NOAA Technical Report NOS 137 NGS 49, US Dept. of Commerce, NOAA/NOS, Rockville, MD

Wünsch J, Seiler U 1992: *Astronomy and Astrophysic*, 266, 581.

Yoder C F, Williams J G, Parke M E 1981: *Journ. of Geophys. Res.*, 86, 881–891.

Zarraoa N, Rius A, Sardón E, Schuh H 1993: In: Proc. of the 8th Working Meeting on European VLBI for Geodesy and Astrometry, Dwingeloo 1991, p. IV-59 - IV-69, Rep. MDTNO-R-9243, Surv. Dep. of Rijkswaterstaat, Delft

Zeppenfeld G 1993: In: Proc. of the 8th Working Meeting on European VLBI for Geodesy and Astrometry, Dwingeloo 1991, p. IV-16 - IV-21, Rep. MDTNO-R-9243, Surv. Dep. of Rijkswaterstaat, Delft

Acta Geod. Geoph. Hung., Vol. 29(3–4), pp. 421–438 (1994)

AUTOMATION OF GEODYNAMICAL OBSERVATIONS

GY MENTES[1]

New developments in electronics and computer techniques, especially in the field of low power consumption, high memory and computation capacity and of diverse input-output capabilities offer the possibility to build fully automatic instruments for the observations of vertical and horizontal motions of the Earth's surface. Therefore the number and the length of gaps in the data records can be diminished what contributes to the best interpretation of the long periodic geodynamical phenomena. A lot of data aéquisition systems are presently available at the market which are apt for the fully automatic measuring, recording, transferring of data and after appropriate programming to control the automatic calibration, to give an alert signal when the measured values are over a pregiven range, *etc.* To apply such an automatic data acquisition system the measuring method and the instrument has to be able to give an electronic output signal, its calibration and other functions (*e.g.* drift or zero-point compensation) have to be made or controlled electrically. Therefore our steady task is to develop such sensors, transducers, instruments which fulfil the requirements of high stability, accuracy and automation. The present paper gives a summary of the main measuring methods used presently for geodynamical measurements with respect of automation.

Keywords: automation; calibration; geodynamics; measurement; sensors; strainmeters; tiltmeter

1. Introduction

To understand the real physical phenomena taking place in the Earth's body we need more and more precise measurements. Table I shows the most typical geodynamical phenomena and the required relative accuracies to obtain them by strain measurements. For the measurement of the long and very long periodic phenomena as *e.g.* tectonic movements, very stable measuring instruments with very high reliability are needed. In the most cases it is very difficult to fulfil these requirements because these instruments are working in rough conditions. Despite careful workmanship in manufacturing of these instruments, they are often reduced to maintenance or what is more characteristic they frequently need a zero (drift) point correction. If these corrections are not made in the right time the output signal of the instrument may go out of the measuring range and there will be a gap in the data record. These gaps can be very long if the instruments are not controlled frequently enough which is the case if the instruments are installed at a distant site. Sometimes the calibration of the instruments is carried out manually. This procedure can also be a source of error. To eliminate a lot of disturbing effects and to diminish the number and length of gaps in the dataflow, a complete automation of the instruments combined with a data transfer capability from the

[1]Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences, H–9400 Sopron, POB 5, Hungary

1217-8977/94/$ 4.00

instrument site to the processing one is needed. This gives later a possibility for a regular control of the state of the instruments and the use of a continuous data processing or — if this would not be — possible the rough or the preprocessed data could be displayed without making a time consuming travel to the instruments.

Table I. Phenomena in the solid Earth

Type of the phenomenon	Period [s]	Relative strain [m]
Earthquake waves	$10^{-1} - 10^{2}$	$10^{-6} - 10^{-9}$
Microseismic waves	$1 - 5 \cdot 10^{1}$	$10^{-9} - 10^{-11}$
Free oscillations	$10^{2} - 10^{3}$	$10^{-8} - 10^{-11}$
Tides	$2 \cdot 10^{4} - 10^{6}$	$10^{-7} - 10^{-8}$
Ocean loading		$9 \cdot 10^{-8} - 10^{-9}$
Seasonal processes	$3 \cdot 10^{7}$	$10^{-7} - 10^{-10}$
Atmosphere		$4 \cdot 10^{-8} - 10^{-9}$
Snow load		$8 \cdot 10^{-9} - 10^{-10}$
Ground water		$4 \cdot 10^{-9} - 10^{-10}$
Rotation of the Earth (estimated elastic strain)	$3 \cdot 10^{7}$	$10^{-9} - 10^{-10}$
Tectonic processes		$10^{-5} - 10^{-7}$

There are a lot of data acquisition systems on the market combined with controllers which are able to sample data within a given sampling rate and accuracy, to control the instruments (calibration, zero point correction, *etc.*), to send and receive data and event reports via modem or by means of a radio transmitter.

The stability requirements against the measuring instrument and so against its main parts as the built-in sensors and transducers are much higher in an automatic measuring system than in a traditional, manually handled one. Otherwise there is no reason for automation.

In this paper, the possibilities of the automation of individual deformation measurement techniques are described giving a brief description of an automatic deformation measuring system. The design and development of the sensors and transducers as the main parts of an automatic measurement system are also given.

2. Deformation measurement techniques

2.1 Space and terrestrial measurement techniques

In the last decades new measurement technologies have been developed for displacement and position measurements. Besides, the stability and the resolution of the widely used tilt- and strainmeters are increased considerably. The deformation measurement techniques used nowadays can be divided into two main groups, namely into space and terrestrial techniques.

The space techniques are:

— the very long baseline interferometry (VLBI),
— the satelite laser ranging (SLR),
— the global positioning system (GPS),
— the planned geodynamics laser ranging system (GLRS).

All of these techniques include a stage of automation. It means that these methods cannot work without considerable automatic functions. The great advantage of all these methods is that they do not need a calibration like tilt- and strainmeters or gravimeters, because these measuring methods are based on fundamental physical constants as the velocity of light or the wavelength of electromagnetic waves of different frequencies. All of these methods give a large amount of data which cannot be processed without automation but certain measuring techniques, e.g. differential measurements by GPS receivers enable a further automation. Disregarding these possibilities in the following we only deal with the automation of the terrestrial deformation measurement methods.

The most important terrestrial techniques of the deformation measurements are:

— geodetic methods:

- distance measurements:
  - measuring rule, tape,
  - geometric and optical distance measurements,
  - electronic distance measurements (EDM),
- leveling:
  - trigonometric,
  - spirit,
  - hydrostatic
- photogrammetry

— strainmeters:

- invar rod,
- invar wire,
- quartz tube,
- laser,

— tiltmeters:

- liquid surface,
- pendulum (horizontal, vertical)

— gravimeters.

2.2 Terrestrial geodetic measurement methods

The terrestrial geodetic measurement methods have a great role in the local deformation measurements e.g. in the surroundings of great tectonic faults. The relatively great displacements in these territories can be measured by geodetic methods in spite of the relative low accuracies of these methods.

The characteristic precision of electromagnetic distance measurements is about 10^{-7} but the measurement accuracy is limited by uncertainties in the integrated refractive index along the measurement path. Several measuring methods were developed to achieve the accuracy of 0.1 ppm. Without corrections for refraction effects in the atmosphere the measurement accuracy is less than 1 ppm.

Optical methods that transfer the horizontal line (plane) of sight to a remote location are limited in accuracy by uncertainties in the refractive index gradients near to the Earth's surface. In spirit leveling the optical path is projected orthogonally to the local plumb line pointed by a pendulum or parallel to a liquid surface. In trigonometrical leveling the optical path may be at an angle to the horizontal. In both cases, errors arising from the ray curvature are minimized through the use of short symmetrical sightings. The application of new technology to both methods has resulted in a reduced measuring and evaluation procedure. However, these improved systems have not substantially better accuracy than the traditional methods. Despite of modern measuring instruments, systematic errors in both systems remain height dependent and require an empirical correction. Thus the errors are accumulated both along the horizontal and the vertical distance traversed. In sum, precise leveling has an uncertainty of approximately 1 cm per vertical kilometer.

The hydrostatic leveling is influenced by dynamic and static density variations of the fluid (water) path and it is difficult to ensure the same air pressure at each end of the instrument. In ideal circumtances the height dependent errors can be reduced to less than 1 mm per vertical kilometer. The practical instruments, however, are not more accurate than the traditional leveling methods.

The measurement processes, handling the instruments, pointing, *etc.* at leveling and electronic distance measurements could not yet be automatized. But a lot of developments were made to automatize the data processing at the measurements (coded leveling rods, automatic data collection built into the instruments, *etc.*).

The main common disadvantage of the current space and terrestrial geodetic measurement methods comes from the measurements made discontinuously. A lot of information concerning long periodic movements between the repetitions of the measurements are lost. Therefore, the noise levels of continuous measurements made by tiltmeters and strainmeters are typically lower than the best available geodetic and space measurements.

The latest developments in the field of robot theodolites enable us continuous measurements. The motorized theodolites equipped with a CCD image sensor camera and an on-line image processing computer are capable of pointing to targets of natural or artificial origin using image processing algorithms. The coordinates of the pointed targets can be calculated from the measured horizontal and vertical angles measured by the theodolite. Using two theodolites, the coordinates can be

calculated from the simultaneously taken stereo image pair by photogrammetric methods. Figure 1 shows the flow-chart of a typical measurement by two automatically controlled motorized theodolites with CCD cameras and photogrammetric image processing. The inner accuracy available by these systems is 0.01 pixel (picture element) which corresponds to an accuracy of about 0.1 mm in object space on a baseline length of about 100 m. At shorter distances the accuracy in the object space is higher, the relative accuracy of such systems is about $10^{-4} - 10^{-6}$. Such systems are wide-spread in the industry and in engineering survey. To make these systems more accurate, CCD cameras with larger sensor area and much more pixels as well as better and faster softwares than previously are needed. Such systems can be used along faults to control the movements continuously.

2.3 Continuously recording tiltmeters and strainmeters

Table I shows that relative motions are very small, in the order of magnitude of about $10^{-7} - 10^{-11}$. The recording instruments have a limited baseline length from a few centimeters to several 100 meters. Therefore, the continuously recording instruments need a very high resolution and accuracy. Since the majority of phenomena causing displacements are long periodic, therefore long term recording is needed for detecting the phenomena in question. This requires recording tilt- and strainmeters of high stability. Generally, these very small displacements are measured continuously by the instruments used for earth tide recording because these instruments have only the required sensitivity. In most cases the records are evaluated as follows: the tidal components are removed from the curve, the remaining residual curve contains the local and global movements, the instrumental drift which consists of signal changes arising from variations of the environmental parameters (temperature, atmospheric pressure, humidity, level changes of the groundwater *etc.*), of the signal variations caused by instrumental parts (mechanical, optical and electrical) and of the instabilities of the recording site. To avoid an interpretation of the above mentioned disturbing effects as a motion during long term measurements, the following requirements should be kept:

1. To record on very stable places, in observatories or boreholes if possible built into or on the bedrock.

2. To reduce the cavity effect as far as possible.

3. A perfectly stable and rigid coupling of the instruments to the bedrock.

4. Application of high accuracy sensors with low drift.

5. To build instruments of very high mechanical stability.

6. To ensure stable environmental conditions at the recording place.

7. Precise measurement of the environmental parameters (temperature, atmospheric pressure, humidity, *etc.*).

8. High precision calibration of the instruments.

9. Parallel recording by several instruments of different and identical types.

The attachment of the continuously recording tilt- and strainmeters to the ground is a very difficult problem, thus a generally valid recipe solving the problem cannot be found. Properly coupled instruments record only deformations which are really present. Deformations are called "real" ones if they are present in great depths: seismic waves, earth tides, surface loads, and crustal deformation. Mea-

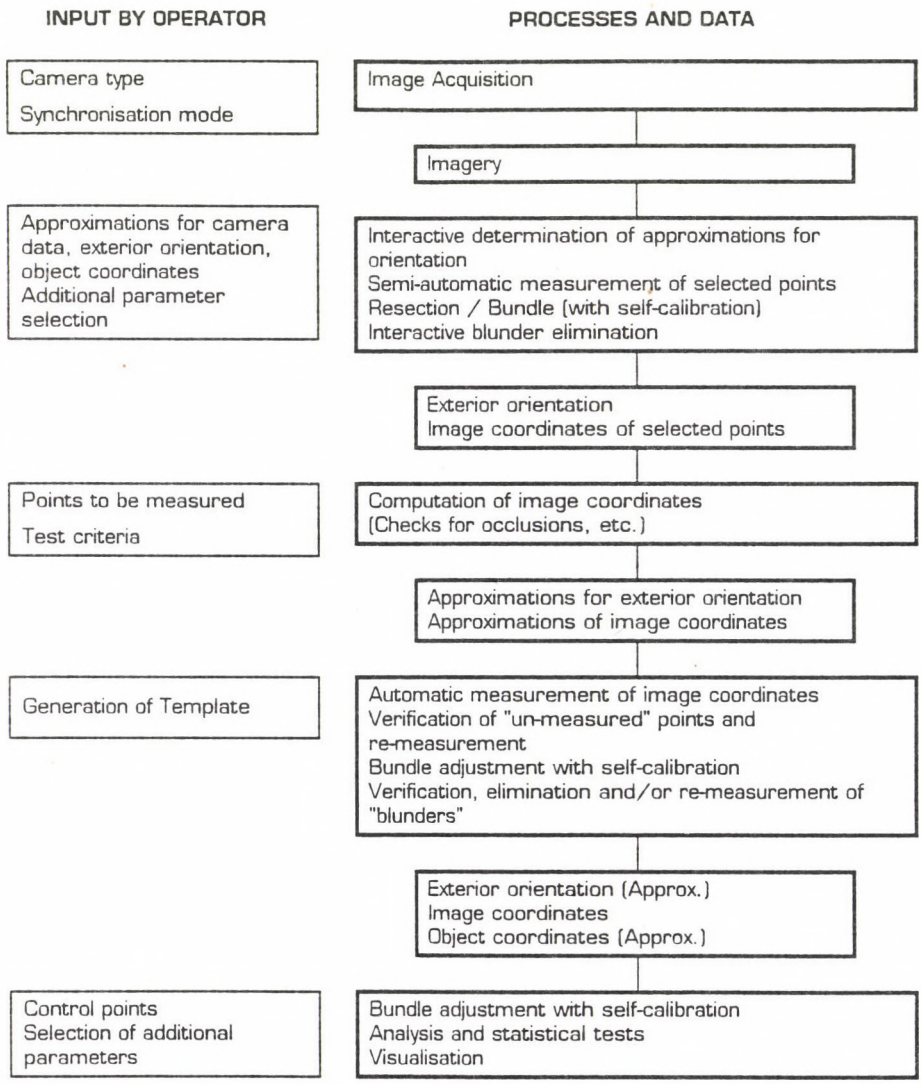

Fig. 1. The flow-chart of a real-time photogrammetric measurement system (Horst 1992)

surements of these phenomena show that the Earth may be assumed to be a nearly uniform elastic body, thus properly coupled instruments record these "real" deformations near the surface of an elastic body. That means that the difficulty of surface installations is rather signals supposed to be caused by other motions of the near-surface layers as *e.g.* by surface motions due to meteorological effects and not the absence of signals due to deeper motions. At deeper installations the groundwater pressure variations may cause problems besides the difficult handling of the instruments *e.g.* in a borehole.

Despite the rigorous requirements against the tilt and strain measurements listed above and the problems of attachment of these instruments to the ground or bedrock, the tilt- and strain measuring instruments play a very important role in the investigation of geodynamical phenomena taking place in the solid crust and interior of the Earth. These instruments are very suitable for automatic measurements.

3. Common construction of automatic deformation measuring instruments

Figure 2 shows a common block-diagram of computer-controlled automatic measuring instruments used for tilt, strain and gravimetric measurements. In some cases the sensor and the transducer of the instruments are inseparable from each other. In the case of the strainmeters a sensor only exists which is at the same time a transducer, too. The quartz tube or the invar rod only transmits the ground displacements from one end of the extensometer to the other, where the relative displacements of both ends are sensed by a capacitive, inductive or optical sensor which transforms the displacement into an electrical or an optical signal. Thus, the sensor is also a transducer. In the case of laser interferometers the laser beam transmits the movement from one end to the other and the laser head functions as a sensor. The sensor of tiltmeters is a horizontal or vertical pendulum (the mass of the pendulum arm) or a liquid surface being always perpendicular to the local plumb. In these cases a separate transducer is needed to transform the mechanical movements of the pendulum arm or the liquid surface into an electrical signal. In the case of gravimeters the movements of the mass have to be transformed into electrical signals by means of a separate transducer.

The transducer of some instruments can be used at the same time for the indirect calibration of the instrument. At gravimeters and pendulums a high D.C. voltage as calibration signal can be given to the plates of the capacitive transducer to produce a displacement of the moving plate of the condenser. By applying another electrostatic force to the trancducer as a feed-back from the output of the transducer amplifier, a zero measuring method can be obtained.

Some of the sensors and transducers need an exciting power for their operation. In case of inductive and capacitive transducers, an oscillator of a frequency range of 5–25 kHz with a magnitude of about 1–50 V gives the exciting voltage. This oscillator also gives the reference signal of a carrier-frequency amplifier.

Because of the measuring method of the most instruments continuously record-

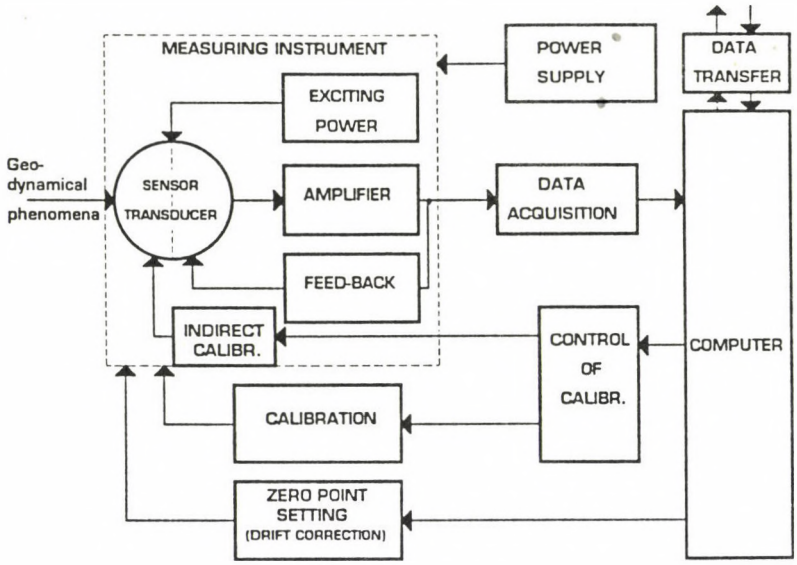

Fig. 2. The block diagram of automatic deformation measurement instruments

ing deformation, the working principle of these instruments are not based directly on fundamental physical constants *e.g.* on the velocity or frequency of electromagnetic waves, hence these instuments need a regular calibration. Some instruments cannot be directly calibrated or it would be very difficult. In these cases an indirect calibration can be used, as mentioned above, to verify the sensitivity and linearity of the instrument (*e.g.* gravimeters).

A direct calibration has an effect on the whole instrument like the geodynamical phenomena to be measured. In the case of tilt- and strainmeters it means that a tilt (e.g. by means of a crapoudine) or a displacement (*e.g.* by means of a crapoudine or a magnetostrictive transducer), respectively, is applied to the instruments. In some cases both kinds of calibration can be combined with each other.

The zero point correction of tiltmeters can be made by a very fine tilting of the suspension of the pendulum or of the liquid level by means of D.C. or stepmotors. The zero point correction of strainmeters can only be made by a displacement of the transducer *e.g.* by shifting the standing plates of the differential condenser.

Both the (direct and indirect) calibration and the zero point correction can be regularly controlled via digital input-outputs of the controller computer of the data acquisition system.

The accuracy of the whole system is determined by the sensor, the transducer, the amplifier and the A/D converter. For higher resolution sensors and amplifiers of very low noise are needed. For very high accuracy high resolution A/D converters (16-bits) are needed. If 16-bits are not sufficient the output voltage of an 8-bit D/A converter can be subtracted from the output voltage of the amplifier. The D/A converter is controlled by the digital outputs of the computer.

The computer can be programmed for prefiltering or preprocessing the measured

data to recognize geodynamical phenomena of higher frequencies (earthquakes, microseisms, *etc.*) and to measure and store these data with a higher sampling rate. The measured data and the status of the whole system can be displayed or transmitted to a remote site via an RS 232 C interface or by means of a modem via telefon line or by a radio transmitter.

Since the computer, other digital and analogous controllers and measuring units, data transfer and processing parts of the automatic measuring system can be purchased or can be assembled from finished parts bought on the market, therefore in the following the sensor and calibration units determining the quality of the whole system will only be discussed.

4. The sensor and calibration parts of the automatic measurement system

4.1 Sensors and transducers

4.11 Electrooptical sensors

The basic principle of these instruments is a reflected light beam to magnify the rotation of a mirror. The reflected beam is directed on a pair of photodiodes whose output signals are subtracted from each other. Thus, the output voltage of the sensor is zero when the light beam is centered and a deflection from the central position produces an output voltage proportional to the displacement of the beam. Such systems can be made very sensitive and stable by means of modern silicon photodiodes. In this case the intensity of the light source must be kept constant. Another disadvantage is that the output signal is analogous and therefore an electrical drift cannot be avoided.

A drift-free sensor can be made by CCD image line or area sensors used for sensing the position of the reflected light beam. A usual CCD line image sensor contains 128–4096 light sensor elements (picture sensor elements, pixels) depending on the type of the line image sensor. The distance between the centres of adjacent pixels is between 7-13 μm at the usual types. These distances are constant and very stable because they depend only on the manufacturing process and the applied silicon semiconductor substrate is mechanically very stable. The position of a projected edge can be determined by reading the illumination (grey) values of each sensor elements, pixel by pixel. The accuracy of CCD sensors is about 0.1 pixel and a higher resolution to 0.01 pixel can be achieved by matemathical methods for a precise edge detection. The displacement transducers built with CCD sensors work with digital signals, thus they have no drift. The disadvantage of these transducers is the complicated electronics usually built with a microprocessor to read out the grey values of the pixels (Fig. 3).

The most common laser sensors used in geodynamics are based on a Michelson interferometer. Figure 4 shows its scheme of principle. A parallel beam of light from a source is sent to a beamsplitter, where it is equally divided and goes to a fixed and a moving reflector. The returned beams interfere at the beamsplitter and go to a detector. Counting frings give an inherently digital output and the value

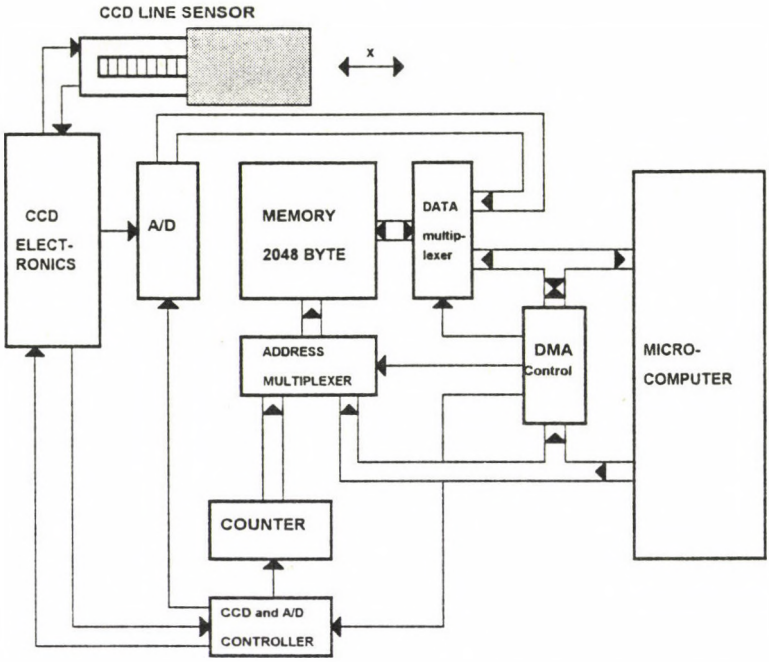

Fig. 3. The block diagram of a displacement transducer built with a CCD image line sensor

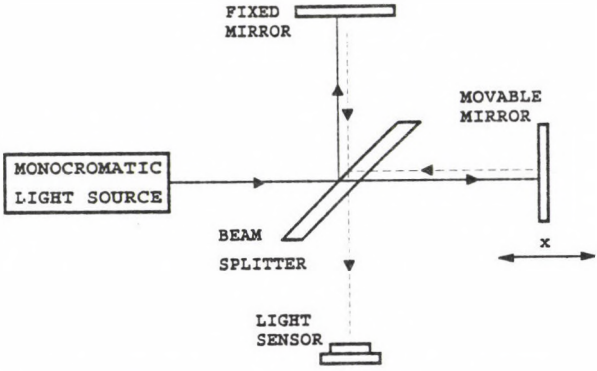

Fig. 4. The principle of the Michelson interferometer

of wavelength (λ) is usually known to 4 or 5 figures. The HeNe gas laser has the best frequency stability (10^{-13}), however, the light produced is not sufficiently narrowband. The available accuracy with laser sensors is a fraction of the wavelength and is about 10^{-7} m in spite of the fact that a lot of efforts were made to increase the resolution to $\lambda/100$ by means of interpolators, but the accuracy is strongly limited by the variations of the refractive index due to changes of temperature and air pressure. They can be eliminated by placing the light path into vacuum. The main

problem at the laser sensors is the short average time between failures of the laser tubes of about 18–48 month.

Nowadays the linear incremental transducers make possible to measure displacements with an accuracy of a few μm. The stability of these transducers is very high because they are made on a glass ruler which has a high mechanical stability and a low coefficient of thermal expansion. The output signal is digital because the impulses due to the displacements are only counted. A new digital method was developed to subdivide the output signal to achieve a resolution of about 0.01 μm (Swatek 1984). The operational life is very long. They have a disadvantage, namely they need a force for their operation, even if a very small one. Applying them in a strainmeter this force is negligible whereas in the high sensitive tiltmeters they are not applicable.

4.12 Electromagnetic (inductive) sensors

These sensors have many different forms, but they can be functionally divided into two main groups: linear variable differential transformers (LVDT) and linear variable reluctance transducers (LVRT). The principle of the LVDT-s is shown in Fig. 5. A sinusoidal voltage is fed to a primary solenoid with an inside movable ferromagnetic core. Secondary coils are put on each end of the primary ones. The magnetization of the core by the exciting field creates a field within the secondary coils. The greater the extension of the core due to its displacement is, the greater are the flux linked by that coil and the voltage induced in it. A large motion of the core is possible and the LVDT is insensitive to the transverse motion of the core. The core is completely passive, and the coils can be shielded from it and from the outside world inside a nonmagnetic housing. Therefore, humidity or even liquid water has no effect on the operation of the LVDT-s. Due to eddy currents and hysteresis, the output voltage is not in phase with the input and also contains higher harmonics and a quadrature signal that do not vanish when the core is in its middle position. The output voltage of the LVDT varies with the frequency and the amplitude of the exciting voltage. Variations in the transformer temperature affect the winding resistance, the magnetic properties of the core, and the dimensional relationships and so the sensitivity of the transducer. Figure 5 shows a solution for the electronics of LVDT where the amplitude of the oscillator is controlled by an additional feedback for compensation of temperature changes and other effects.

The linear variable reluctance transducer (LVRT) is physically very similar to the LVDT-s just discussed. The only difference is the coil arrangement; the three transformer coils of the LVDT are replaced by a single, center-tapped coil as shown in Fig. 6. The coil is usually connected as one half of an inductive bridge. The bridge output voltage is linear with position, its phase is shifted by 180° as the core crosses the null position. The output voltage of the bridge is amplified by a carrier-frequency amplifier similarly to the capacitive transducers.

The available accuracy of LVDT-s and LVRT-s is about 0.5 % of the measuring range or somewhat better. Their main advantages are:

- Relatively high accuracy, sensitivity and linearity

- Cross-axis rejection
- Low sensitivity to disturbing electrical signals
- Not sensitive to humidity

Their main disadvantages are:

- Moving mass (inertia)
- Friction of the moving core
- High reaction force

Because of the friction and the high reaction force these trancducers cannot be used in instruments where the deflecting force or torque of the measured geodynamical phenomena is low *e.g.* at horizontal pendulums.

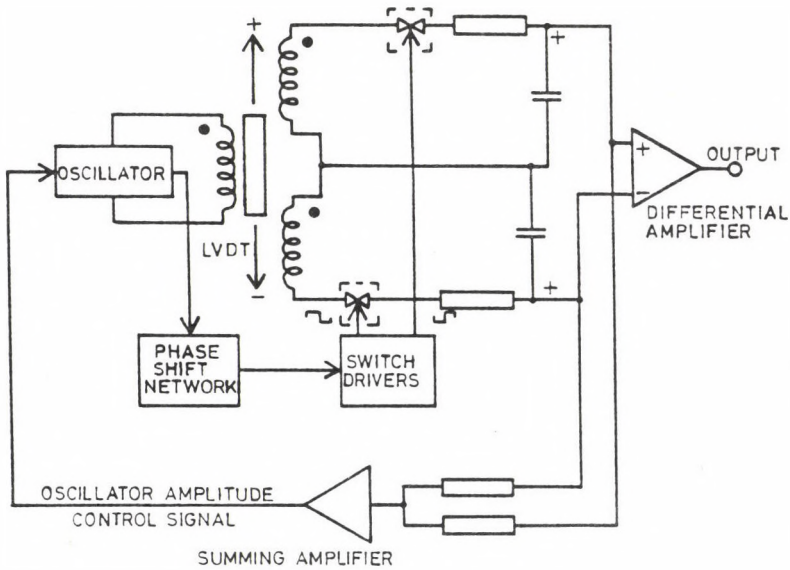

Fig. 5. The principle of the linear variable differencial transformer (LVDT)

4.13 Capacitive transducers

The most sensitive transducers used in geodynamical instruments are capacitors changing their electrical capacitances in consequence of mechanical displacements. The most familiar capacitor consists of two parallel plates separated by an air space or by a dielectric (insulating) material. The capacitance is affected not only by the spacing and size of the plates, but also by the dielectric between them. The nonlinearity of a simple capacitor and its sensitivity due to the changes of environmental parameters can be reduced by three-plate differential condensers.

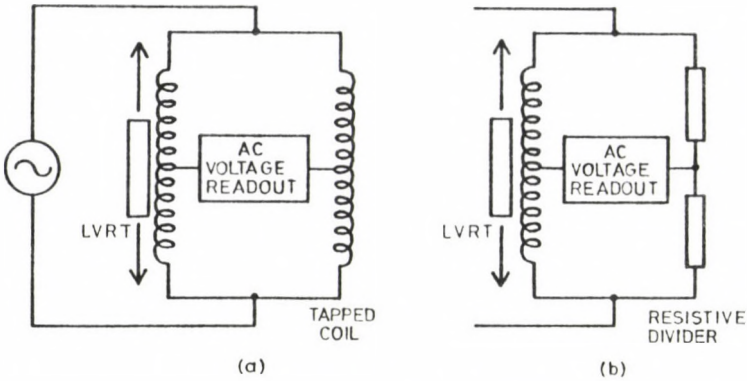

Fig. 6. The principle of the linear variable reluctance transducer (LVRT). a) tapped coil, b) resistive divider

The change of capacitances can be measured in a bridge circuit the bridge is usually supplied by an AC voltage of constant frequency and amplitude (Fig. 7a). The output voltage is proportional to the amplitude of the supply voltage of the bridge, therefore it must be very stable. The output impedance of the bridge depends on the frequency of the supply voltage. In case of the usual bridge capacitances of 10–50 pF, the bridge is fed by a supply voltage of about 5–15 kHz to hold the output impedance at an acceptable value (some kΩ's).

Figure 7b shows the equivalent circuit of the bridge together with the spurious capacitances arising from the mounting and from cable capacitances in the case when the moving plate of the differential condenser is the output of the bridge and the middle point of the two constant bridge capacitances is grounded. When the constant capacitors (C) have a large value of about 1 nF, then the spurious capacitances of the order of 20 pF do not effect the output signal of the bridge, only their change causes a noise in the output voltage of the bridge. The load capacitance between the cable connected to the output of the bridge and the ground is about 20–40 pF/m depending on the type of the cable, therefore it forms together with the input impedance of the preamplifier and the output impedance of the bridge a voltage divider and reduces the gain of the sensor. That is the reason why the preamplifier should have a high input impedance and should be placed close to the bridge.

Unfortunately, the moving plate of the differential condenser must be directly connected at most instruments to the moving part of the instrument *e.g.* at pendulums and gravimeters. In this case the spurious capacitances are connected parallel to the two parts of the differential condenser (Fig. 7c). Since all of the capacitances are in the same order, the spurious capacitances cause large disturbances in the output signal of the bridge. To avoid this effect the solution shown in Fig. 7d can be chosen by using a part of the differential condenser as a coupling capacitor of constant value. This coupling capacitor causes an additional voltage division (Mentes 1983).

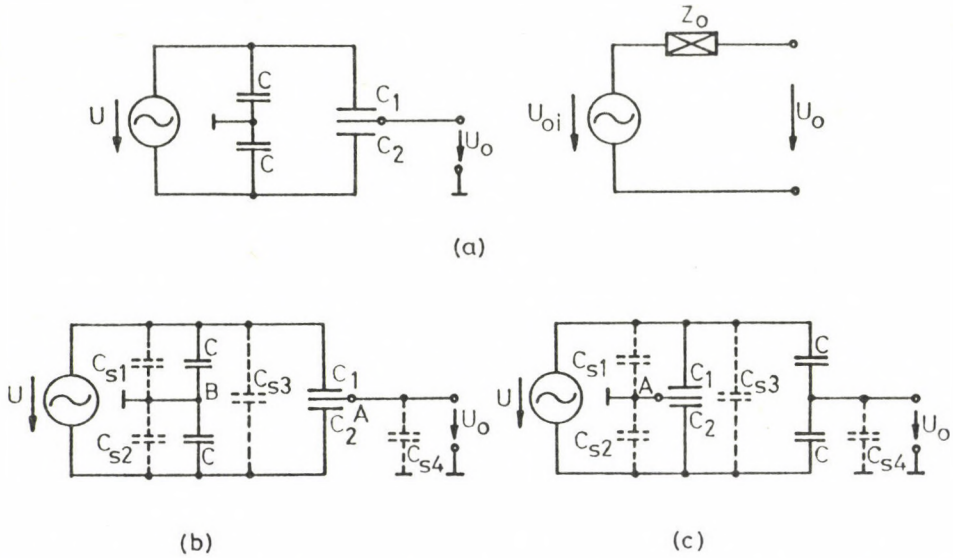

Fig. 7. Capacitive bridges: a) the usual bridge circuit and its equivalent circuit, b) the spurious capacitances of the bridge, when the middle plate is not grounded, c) the spurious capacitances of the bridge, when the middle plate is grounded

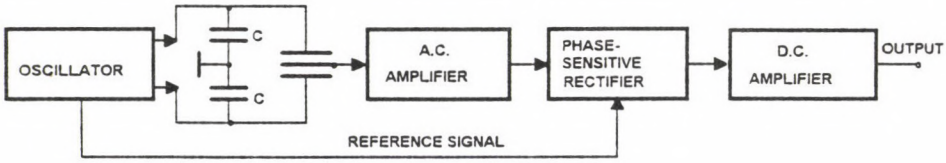

Fig. 8. The carrier-frequency amplifier

The best method to measure the output voltage of the bridge is to use a carrier-frequency amplifier (also called a lock-in amplifier), as shown in Fig. 8. An oscillator with high amplitude stability and with frequency ω supplies the capacitive bridge (the outer plates of the differential condenser). An amplifier of high input impedance amplifies the output signal of the bridge and a phase-sensitive rectifier only detects that part of the signal which is in phase with the supply voltage of the bridge. Averaging this signal makes the detector insensitive to noise with the exception of a narrow frequency band around ω. In this manner a resolution of $10^{-10} - 10^{-14}$ can be achieved. Another method is shown in Fig. 9. The two parts of the differential condenser are connected with two oscillators. The frequencies of the oscillators are subtracted from each other and the resulting frequency is proportional to the displacement of the moving plate of the transducer. This solution has a lower resolution than the above mentioned carrier-frequency method. Its advantage is that it has a direct digital output signal.

Other electrical methods to detect the output voltage of the bridge using *e.g.*

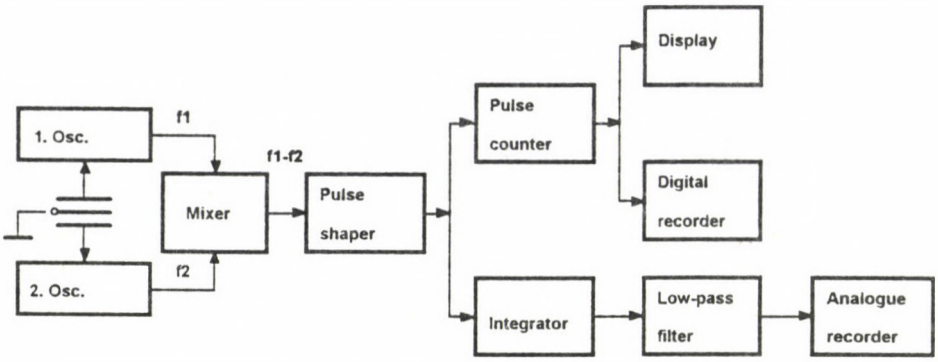

Fig. 9. The two-oscillator circuit

digital integrated circuits have additional error sources (*e.g.* jitter) notwithstanding these solutions being theoretically bridge circuits, too.

4.14 Vibrating wire sensors

The mechanical eigenfrequency of a wire depends on the tension in it. By measuring this frequency physical quantities can be determined causing this tension. The principle of the sensor is shown in Fig. 10. In the intermittent mode an electromagnet gives pulses into the mesuring wire and the same magnet senses the vibrations in the wire (Fig. 10a). In the continuous mode an oscillator feeds a continuous sinusoidal mechanical vibration into the wire. An electromagnetic transducer senses this vibration and the signal of this transducer is fed back to the oscillator for controlling its frequency to the eigenfrequency of the wire (Fig. 10b). Nowadays, some instruments (tilt- and strainmeters) are available for industrial purposes (Maihak 1988). These are of lower sensitivity than needed in geodynamics, but some new developments in material technology enable us to achieve a relative accuracy of 10^{-7}. The main advantage of these transducers is that they have a direct digital output signal (Mentes 1989).

4.2 Calibration units

Electrically controlled calibration units can only be used in automatic measuring systems. The indirect calibration used in the sensors or transducers is made in the most cases by means of an electromagnetic or electrostatic force acting on the moving part of the sensor (*e.g.* on the mass of a gravimeter). The calibration current or voltage can be kept very stable and can be controlled by the computer of the automatic system. In this case the calibration is used only for controlling the constancy of the sensitivity of the measuring system.

The magnetic elongation (magnetostrictive effect), the piezoelectric effect and the deformation of a closed membrane (crapoudine) are used for direct calibration. The piezoelectric materials are hygroscopic and therefore not applicable. Thus,

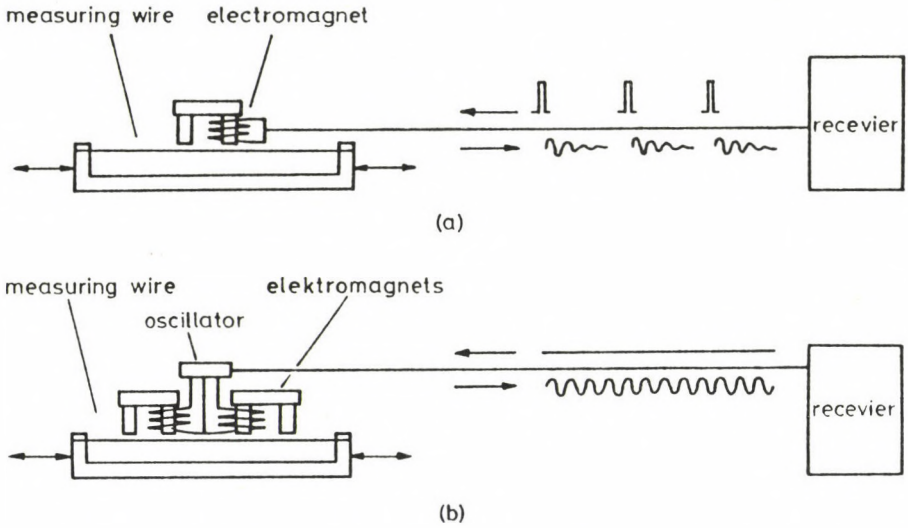

Fig. 10. The principle of the vibrating wire transducers; a) intermittent mode, b) continuous mode

the most usable calibration units are crapoudines and magnetostrictive coils. Both can be very easily controlled electrically and therefore they are suitable for the calibration of automatic measuring systems.

4.21 Crapoudine

The crapoudine consists of a thick-walled box made of stainless steel at which the surface of the box to be deformed, the so-called pressure face is about 5 mm thick. Thus, it produces very small displacements (less than 1 μm) at high pressure variations. The box is connected via a flexible pressure-tight tube to a vessel and the whole system is filled with mercury. The pressure in the box and so the bulge of the pressure face producing the displacement can be changed by lifting or sinking the mercury vessel (Fig. 11). The moving of the mercury vessel is made by a D.C. or a step motor and so it is very easily controllable by a computer. The calibration of the pressure box (crapoudine), producing the displacement is made by a laser interferometer (Melchior 1978) or with a better resolution and accuracy by the calibration device developed in the Geodetic and Geophysical Research Institute of the Hungarian Academy of Sciences (Mentes 1993) before the crapoudine is built into the instrument. But it is to be remarked that this calibration device must be calibrated in large steps by means of a laser interferometer, too.

The sensitivity of the crapoudines depends on the thickness of the pressure face of the box. Thus, the sensitivity can be set during the manufacturing. The characteristics of the crapoudines are linear in a wide pressure range. The linearity error is usually less than one percent (Mentes 1993). According to the tests carried out regularly for a long time, they have a small hysteresis, a very slow ageing and a very small elastic after-effect.

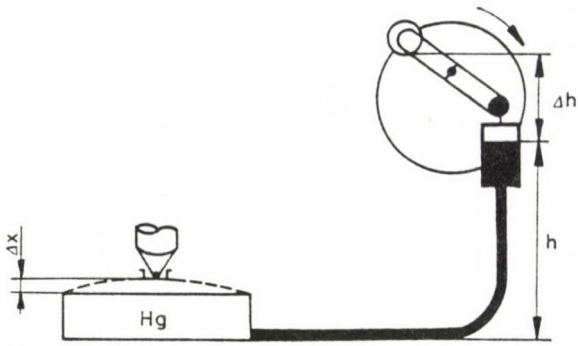

Fig. 11. The principle of the crapoudine

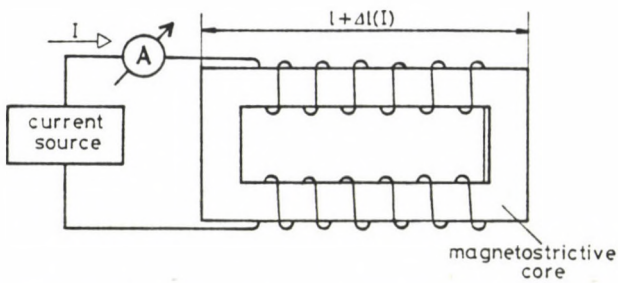

Fig. 12. The principle of the magnetostrictive calibration unit (magnetostrictive coil)

The main disadvantage of the crapoudine is, however, that the mercury is very injurious to the health and therefore it must be very carefully handled. Moreover the calibration units built with crapoudines need a relatively complicated mechanics and an electric drive.

4.22 Magnetostrictive coils

These calibration units consist of a core made of a very high quality magnetostrictive material, usually of a nickel alloy. If this material is placed in a magnetic field induced by the current flowing through the coil reeled up on the core, this latter changes its length as a function of the magnetic field strength and hereby of the current in the coil (Fig. 12).

The characteristics of the magnetostrictive coils are linear functions of the current, but they have usually a little bit greater linearity errors than the crapoudines. Despite of this, the errors are whithin one percent. The hysteresis is also somewhat greater than that of the crapoudines.

If the coil is induced always with the same current intensity, the repeatability errors of the core elongations are within 0.1 percent over many years according to the measurements.

In contrast to the crapoudines the magnetostrictive coils have a thermal effect when the current impulse driving the coil is too long, due to a heat produced by

the current in the coil. The less is the sensitivity of the magnetostrictive material, the higher current is needed to produce the same elongation of the core and this higher current produces more heat causing an observable thermal expansion of the core. Therefore, the usual calibration impulse-width should not be more than some minutes.

The control and handling of the magnetostrictive coils is more easier than that of the crapoudines.

5. Conclusion

The recent space measuring techniques can compete neither with the 10^{-9} resolution of the continuous strain and tilt measurements, nor with the data amount continuously recorded since several decades. In other respect the space measurements provide a global data basis for unambiguous interpretation of strain and tilt measurements.

It is a big problem at tilt- and strainmeters to decide whether the instrument gives useful measuring data or not. Without this information the interpretation of the measured data is impossible and it cannot be decided if the instrument is correct. To solve this problem as many instruments should be put close together as many is possible. Therefore, the main task will be in the future to develop low cost instruments besides the development of sensors, transducers and calibration units of high stability and accuracy. In other respect it is very advantageous to use the same automatic measuring system to several instruments of the same and different types at a measuring site what remarkably reduces the costs of the automation per instrument.

References

Horst A B 1992: Geometric and Radiometric Analysis of a CCD-Camera Based Photogrammetric Close-Range System. Institut für Geodäsie und Photogrammetrie an der Eidgenössischen Technischen Hochschule, Zürich, Mitteilungen Nr. 51.

Melchior P 1978: The Tides of the Planet Earth. Pergamon Press, Oxford

Maihak A G 1988: Data sheets and MDS equipment list. Hamburg

Mentes Gy 1983: *Acta Geod. Geoph. Mont. Hung.*, 18, 359-368.

Mentes Gy 1989: In: Proceedings of the Eleventh International Symposium on Earth Tides, Helsinki, 41-47.

Mentes Gy 1993: *Marées Terrestres Bulletin d'Informations*, 115, 8467-8471.

Swatek H 1984: *Der Wirtschaftsingenieur*, 16, 38-41.

Acta Geod. Geoph. Hung., Vol. 29(3–4), pp. 439–442 (1994)

PERTURBATIONS OF INITIALLY CIRCULAR SATELLITE ORBITS CAUSED BY THE ZONAL HARMONICS OF THE GEOPOTENTIAL

A PAL[1], V MIOC[2], M STAVINSCHI[3]

First order perturbations caused by *all* zonal harmonics of the geopotential in the motion of artificial satellites with initially circular orbits, over one nodal period, are analytically determined. Considered separately, the even zonal harmonics produce orbit precession, while the odd ones produce orbit deformation. The results can be particularized to every individual harmonic.

Keywords: artificial satellites; Earth's gravitational field

The perturbed orbital motion of artificial satellites in the terrestrial gravitational field constitutes a well studied problem. Very significant contributions to this problem have brought *e.g.* King-Hele (1958), Brouwer (1959), Kozai (1959), Izsák (1960, 1961), Zhongolovich (1960a, b), Kaula (1966), and many others. Such studies take into account only one or few harmonics of the geopotential. In this paper we shall consider the influence of all zonal harmonics of the geopotential.

So, consider that the Earth has geometrical and dynamic symmetry with respect to its rotation axis (hence only the zonal harmonics of its gravitational potential are taken into account). In this case the geopotential will be written as

$$U = (\mu/r)(1 - \sum_{n=2}^{\infty} J_n (R/r)^n P_n(\sin\phi)) , \tag{1}$$

where μ = Earth's gravitational parameter, R = mean equatorial terrestrial radius, J_n = features the n-th zonal harmonics, r = geocentric radius vector of the satellite, ϕ = latitude, P_n = Legendre polynomials.

We shall determine the first order perturbations for satellites with initially circular orbits. The variations of the orbital elements caused by the zonal harmonics of the geopotential will be determined over one nodal period of the satellite. So, we describe the perturbed motion by means of Newton-Euler equations written with

[1]Babes-Bolyai University, Faculty of Mathematics and Computer Science, Str. M. Kogălniceanu, RO–3400 Cluj-Napoca, Romania

[2]Astronomical Institute of the Romanian Academy, Astronomical Observatory Cluj-Napoca, Str. Ciresilor 19, RO–3400 Cluj-Napoca, Romania

[3]Astronomical Institute of the Romanian Academy, Astronomical Observatory Bucharest, Str. Cutitul de Argint 5, RO–75212 Bucharest 28, Romania

1217-8977/94/$ 4.00

respect to the argument of latitude (u) in the form (*e.g.* Mioc 1991):

$$\begin{aligned}
dp/du &= 2(Z/\mu)r^3T\,,\\
dq/du &= (Z/\mu)(r^3kBCW/(pD)+r^2T(r(q+A)/p+A)+r^2BS)\,,\\
dk/du &= (Z/\mu)(-r^3qBCW/(pD)+r^2T(r(k+B)/p+B)-r^2AS)\,,\\
d\Omega/du &= (Z/\mu)r^3BW/(pD)\,,\\
di/du &= (Z/\mu)r^3AW/p\,,\\
dt/du &= Zr^2(\mu p)^{-1/2}\,,
\end{aligned} \tag{2}$$

where $Z = (1 - r^2C\Omega/(\mu p)^{1/2})^{-1}$, p = semilatus rectum, $A = \cos u$, $B = \sin u$, $C = \cos i$, $D = \sin i$ (i = inclination), Ω = longitude of ascending node, $q = e\cos\omega$, $k = e\sin\omega$ (e = eccentricity, ω = argument of perigee, S, T, W = radial, transverse and binormal components of the perturbing acceleration respectively.

Since the elements $y \in Y = \{p, q, k, \omega, i\}$ have small perturbations over one revolution of the satellite, they can be considered constant and equal to $y_0 = y$ ($u = 0$) in the right-hand sides of (2), hence every equation may be separately integrated. The variations of y over one revolution are:

$$\Delta y = \int_0^{2\pi} (dy/du)du, \quad y \in Y\,, \tag{3}$$

with integrands provided by (2). Integrals (5) are estimated by successive approximations, with $Z \approx 1$. We shall stop the process at the first order perturbations over one nodal period.

Taking into account all these considerations, we shall anticipately omit the factor Z; also we shall drop the subscript "0" in the right-hand side of motion equations (but retaining that there the element y represents in fact the constant y_0).

By (1), the perturbing function has the expression:

$$R_Z = -\mu \sum_{n=2}^{\infty} J_n(R^n/r^{n+1})P_n(\sin\phi)\,. \tag{4}$$

The components of the perturbing acceleration are

$$\begin{aligned}
S &= \partial R_Z/\partial r\,,\\
T &= (1/r)(\partial R_Z/\partial\phi)(A/B)\tan\phi\,,\\
W &= (1/r)(\partial R_Z/\partial\phi)C\sec\phi\,.
\end{aligned} \tag{5}$$

Performing the calculations, and with $\sin\phi = DB$, (5) becomes

$$\begin{aligned}
S &= \mu \sum_{n=2}^{\infty}\sum_{j=0}^{n'} J_n(R^n/r^{n+2})H_{nj}D^{n-2j}B^{n-2j}\,,\\
T &= \mu \sum_{n=2}^{\infty}\sum_{j=0}^{n'} J_n(R^n/r^{n+2})K_{nj}D^{n-2j}AB^{n-2j-1}\,,\\
W &= \mu \sum_{n=2}^{\infty}\sum_{j=0}^{n'} J_n(R^n/r^{n+2})K_{nj}CD^{n-2j-1}B^{n-2j-1}\,,
\end{aligned} \tag{6}$$

where $n' = [n/2]$ (integer part), and we abbreviated

$$\begin{aligned} H_{nj} &= (-1)^j (n+1)(2n-2j)!/(2^n j!(n-j)!(n-2j)!)\,, \\ K_{nj} &= (-1)^{j+1}(n-2j)(2n-2j)!/(2^n j!(n-j)!(n-2j)!)\,. \end{aligned} \tag{7}$$

Replacing (6) in the first five equations (2) considered separately, and taking into account the fact that the initial orbit is circular (of the radius $r = p$), after long but straightforward algebra we get the integrands of (3) expressed only in terms of explicit functions of u (through A and B) and quantities considered constant over one revolution of the satellite.

The integrations were separately performed for even and odd n. For $n = 2s$ we obtain

$$\begin{aligned} \Delta y &= 0, \quad y \in Y - \{\Omega\}\,, \\ \Delta\Omega &= \pi \sum_{s=1}^{\infty} \sum_{j=0}^{s} J_{2s}(R/p)^{2s} C D^{2s-2j-2} F_{sj}\,, \end{aligned} \tag{8}$$

where we denoted

$$\begin{aligned} F_{sj} &= (-1)^j (4s-2j)!(2s-2j+1)!! \cdot \\ &\quad \cdot (2^{3s-j-2} j!(s-j-1)!(2s-j)!(2s-2j+1)!)^{-1}\,. \end{aligned} \tag{9}$$

For $n = 2s+1$ we get

$$\begin{aligned} \Delta y &= 0, \quad y \in Y - \{q\}\,, \\ \Delta q &= \pi \sum_{s=1}^{\infty} \sum_{j=0}^{s} J_{2s+1}(R/p)^{2s+1} D^{2s-2j+1} G_{sj}\,, \end{aligned} \tag{10}$$

with the abbreviation

$$\begin{aligned} G_{sj} &= (-1)^{j+1} s(4s-2j+2)!(2s-2j+1)!! \cdot \\ &\quad \cdot (2^{3s-j} j!(s-j+1)!(2s-j+1)!(2s-2j+1)!)^{-1}\,. \end{aligned} \tag{11}$$

Therefore, for initially circular orbits, the even zonal harmonics causes first order perturbations only in the longitude of ascending node (resulting orbital precession). The odd zonal harmonics cause first order perturbations only in eccentricity (by the parameter q) and (by the constancy of p) in semimajor axis; hence, under the separate influence of these harmonics, after one nodal period, the orbit will no longer be circular.

Of course, most of the results obtained by the other authors as to the perturbations caused by the zonal harmonics of the geopotential in artificial satellite motion are more general from different standpoints: nonzero initial eccentricity, determination of higher order perturbations, *etc.* However, all such studies consider only one or few harmonics. From this point of view, our results (8)–(11) are more complete, since they allow the determination of first order effects caused by all zonal harmonics of the geopotential, even or odd, considered separately or together.

References

Brouwer D 1959: *Astron. J.*, 64, 97.
Izsák I 1960: *SAO Spec Rep.*, No. 52.
Izsák I 1961: *Astron. J.*, 66, 129.
Kaula W M 1966: Theory of Satellite Geodesy. Blaisdell Publ. Co., Waltham, Mass.
King-Hele D G 1958: *Proc. Roy. Soc.*, A 247, 49.
Kozai Y 1959: *Astron. J.*, 64, 367.
Mioc V 1991: *Astron. Nachr.*, 312, 127.
Zhongolovich I D 1960a: *Byull. Inst. Teor. Astron.*, 7, 521.
Zhongolovich I D 1960a: *Byull. Inst. Teor. Astron.*, 7, 743.

Acta Geod. Geoph. Hung., Vol. 29(3–4), pp. 443–456 (1994)

TECHNIQUES AND METHODS IN GEODYNAMICAL RESEARCH: SPACE VLBI

I FEJES[1]

Space VLBI is a future technique which will be realised in the second half of this decade by putting radio telescopes as interferometer elements into orbit. Besides the main astronomical interests, this new development has some very interesting potential applications in satellite dynamics, geodesy and geodynamical research which qualify space VLBI a potential new technique in space geodesy. The paper reviews the developments led to the basic concepts, the requirements for application in reference frame unification and precise orbit determination. The properties of space VLBI observables, modelling results and the limitations of the present systems under development are also discussed. Finally a dedicated satellite concept (SURF) for unification of reference frames is presented. SURF may overcome the limitations of the first generation space VLBI satellites in this particular field.

Keywords: orbit; reference frame; space geodesy; space VLBI

1. Introduction

Space VLBI is an emerging new space technique which will extend baselines into space in the second half of this decade. What role space VLBI will play in geodesy and geodynamics in the future? What are the new characteristic of space VLBI which makes it special? The purpose of this paper is to explain what is space VLBI in plain terms, why it is different from ground based VLBI, and what are the potential applications in geodesy, geodynamics, and orbital dynamics.

It is assumed thorough the paper that the reader is familiar with the basic principles of ground based VLBI and its astronomical and geodetic applications. Otherwise the reader is referred to the paper of Schuh (1994).

2. What is space VLBI?

Space VLBI is the extension of the Very Long Baseline Interferometry technique into space by putting one (or more) radio telescope(s) in orbit and using them as radio interferometer elements. This can be a combination between ground and space telescopes or just between space telescopes as well. The feasibility of space VLBI has been demonstrated in 1986, when a TDRS communication satellite was operated shortly as a radio interferometer element in combination with ground radio telescopes in the USA and Japan (Levy et al. 1986)

A schematic diagram of a space VLBI system is illustrated in Fig. 1. It consists of four basic components:

– the space radio telescope(s),

[1]FÖMI Satellite Geodetic Observatory, H–1373 Budapest, POB 546, Hungary

1217-8977/94/$ 4.00

- the telemetry and control stations,
- ground radio telescope(s),
- processing facility.

Note that the first two points are the concerns of the space agencies, while the last two points are technically covered by the ground-based VLBI facilities, thus in order to operate space VLBI, a very complicated interaction (and understanding) between the two groups are necessary.

Currently two dedicated space VLBI satellites are being prepared for launch in the second half of this decade with radio telescopes of 8–10 m diameter. The VSOP (VLBI Space Observatory Program) is being developed in Japan (Inoue 1993) and the Radioastron (Kardashev and Slysh 1988), which is being constructed by the Russian Space Agency in international collaboration with seven other countries or agencies.The main characteristics are given in Table I.

There are already preliminary discussions about the next generation space VLBI satellites with improved characteristics to be realised after the turn of the century (Ulvestad 1994, Kardashev et al. 1994).

3. Why is space VLBI different?

From astronomical point of view the most important point is that space VLBI provides higher angular resolution and better imaging potential with respect to ground based VLBI. *From the point of view of geophysical and geodetic applications*, however, the most important fact is that the space antenna is free falling in the gravity field of the Earth while the ground VLBI stations are rigidly fixed to the surface. From this fact some very interesting consequences follow for the information

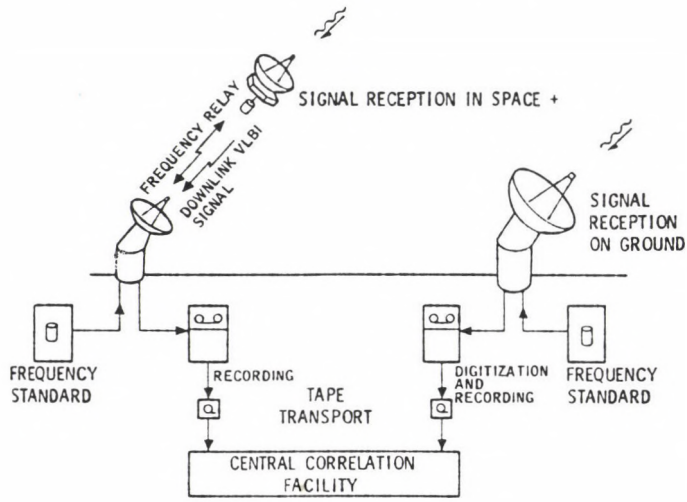

Fig. 1. Schematic diagram of a space-ground VLBI system

Table I. Characteristics of planned space VLBI missions

Space VLBI experiments	VSOP	RADIOASTRON
Country	Japan	Russia +7 countries
Launch date	1996	1997
Expected lifetime	5 years	3 years
Height at apogee	20 000 km	75 000 km
Maximum baseline length	30 000 km	85 000
Observing frequencies	22 GHz, 5 GHz, 1.6 GHz	22 GHz, 5 GHz, 1.6 GHz, 0.3 GHz
Maximum angular resolution	90 micro arc sec.	30 micro arc sec.
Telescope diameter	8 m	10 m
Sensitivity	40 mJy	20 mJy

content of the space VLBI observables, the delay, delay-rate and phase, referred to the system of orbital elements — ground based telescope positions — radio source positions. A more detailed discussion on space VLBI kinematic, dynamic and measurement properties is given by Fejes and Mihály (1991), here only a short summary of these follows.

3.1 Kinematic properties

In the rotating frame of the Earth the relative motions between the ground stations amount to maximum a few cm/year while the space element moves with velocities up to 9 km/s. This is 13 orders of magnitude difference, which indicates the huge range of model quantities which are to be applied in space VLBI. In an inertial frame the range of the space VLBI delay, delay-rate observables can be more than an order of magnitude larger, the accelerations more than 2 orders of magnitude larger than at ground based VLBI.

3.2 Dynamic properties

The orbit of the space antenna defines an inertial frame dynamically. The focus of the orbital ellipse is the mass centre of the Earth. At the other hand the space antenna is tied by interferometric baselines to the Terrestrial Reference Frame defined by the network of ground observatories while observing extra-galactic radio sources. This can be exploited in two different ways:

- for tying the terrestrial reference frame directly to the inertial reference frame defined by the observed quasars,
- to use the space interferometric observables for improving orbit determination accuracy of the space telescope.

3.3 Measurement properties

Because the raw interferometric data are first recorded and only later correlated, the VLBI data processing and evaluation allows for rather large a priori geometrical

Table II. Data types of satellite geodetic techniques

Satellite technique	Data type (observable)	Relation
Satellite Laser Ranging	range	ground station - satellite
Doppler (TRANSIT)	range-rate	ground station - satellite
Altimetry	elevation	surface - satellite
GPS	pseudo-range	ground station - satellite
Space VLBI	delay, delay-rate	ground station - satellite - radio source

model errors. The so called "fringe search" process serves to narrow down these errors and squeeze them into a standard error window of the correlator. After the interferometric fringes have been found, a much higher positional, velocity, acceleration and timing accuracy will be available which can be further improved by residual analysis. In the case of space VLBI, precise positional, velocity, acceleration and timing information should be supplied continuously to the correlator during operation because these parameters are changing very rapidly. Therefore space VLBI data processing needs an additional module which provides these quantities.

3.4 The place of space VLBI among other satellite geodetic techniques

At the presently available satellite techniques for geodesy — SLR, Doppler, GPS, altimetry — the basic observables are the range and the range-rate (elevation measurements above the surface is also a kind of range measurement). These observables carry information related to the ground to satellite geometry only, without any reference to the Celestial Reference Frame defined by quasar positions. Space VLBI observables, the delay and delay-rate, at the other hand, *contain information related to the geometry of ground station — satellite — radio source positions simultaneously.* In this respect space VLBI is unique among satellite geodetic techniques and in principle, capable of complementing them by providing external reference.

Table II summarises the properties of current satellite geodetic techniques.

It should be added, that due to the two way radio link between the tracking stations and the space telescope a high precision range-rate observable is also available beside the traditional satellite tracking data.

4. The basic concept of space VLBI application

The space VLBI delay and delay-rate observables are functions of the position of the observed radio source ($\overline{X_Q}$), the space telescope position, the ground telescope position ($\overline{x}_g$), the Earth orientation parameters (EOP_k) and the epoch of observation (t). Furthermore, the space telescope position can be expressed as a function of the orbital elements and additional perturbations (E_i, P_j):
For the space VLBI delay

$$\tau = F_1(\overline{X_Q}, E_i, P_j, x_g, EOP_k, t)$$

similarly for the delay-rate

$$\dot{\tau} = F_2(\overline{X_Q}, E_i, P_j, x_g, EOP_k, t) \, .$$

Using *a priori* values for the parameters, the delay and delay-rate can be calculated by the F_1, F_2 functions. Differences between the observed (O) and calculated (C) space VLBI observables will occur due to the a priori parameter errors, modelling errors, and observational errors. Assuming that both the $O - C$ errors and the parameter errors are small, a least squares adjustment can be applied for the parameter estimation. The observation equation for the delay observable will have the following general form (Fejes and Mihály 1991):

$$\frac{\partial F_1}{\partial X_Q} dX_Q + \sum_i \frac{\partial F_1}{\partial E_i} dE_i + \sum_j \frac{\partial F_1}{\partial P_j} dP_j + \frac{\partial F_1}{dx_g} dx_g + \sum_k \frac{\partial F_1}{\partial EOP_k} dEOP_k + L = V$$

where L is the vector of $\tau_O - \tau_C$ differences and V is the vector of residuals. For the delay-rate observable F_2 and $L = \dot{\tau}_O - \dot{\tau}_C$ should be substituted into the observation equation (see in more detail in Ádám 1990b).

Based on the unique characteristics of space VLBI observables described by the above equations the following applications have been suggested in the field of geodesy and geodynamical research:

- unification of the terrestrial and the celestial reference frames, because the observation equation contains the source co-ordinates, station co-ordinates, and the EOP parameters which are the basic components of transformation between these two systems,
- orbit determination accuracy improvement of space VLBI satellites, because the E_i orbit parameters can be improved by space VLBI delay, delay-rate observables,
- combination of space VLBI data with LAGEOS data in order to avoid the rank defect of ascending node at laser geodynamic satelliteÆs orbit (this follows from the previous point) and improve gravity field models,
- monitoring of the geocenter position in an inertial reference frame.

The interrelation between the conventional inertial system (CIS), the conventional terrestrial system (CTS) and the techniques which are applied to realise them are shown in Fig. 2. The unique place of the space VLBI technique is pointed out by its central position. It provides inherently the data type of both, the ground based VLBI and the geocentric satellite techniques like the LAGEOS, GPS *etc.* It has also an unusual duality on the system levels. Previously the observing ground networks and their celestial targets were clearly separated. The GPS and the SLR ground networks targeted the GPS and SLR satellites in orbit, similarly the ground based VLBI networks (*e.g.* IRIS, JPL, *etc.*) targeted their own list of extra-galactic radio sources. In the case of space VLBI these dual roles are merged because a space VLBI satellite is an observing station and an observed target at the same time. This is why its box is crossing the ground — space division line. It provides a kind of *space collocation.*

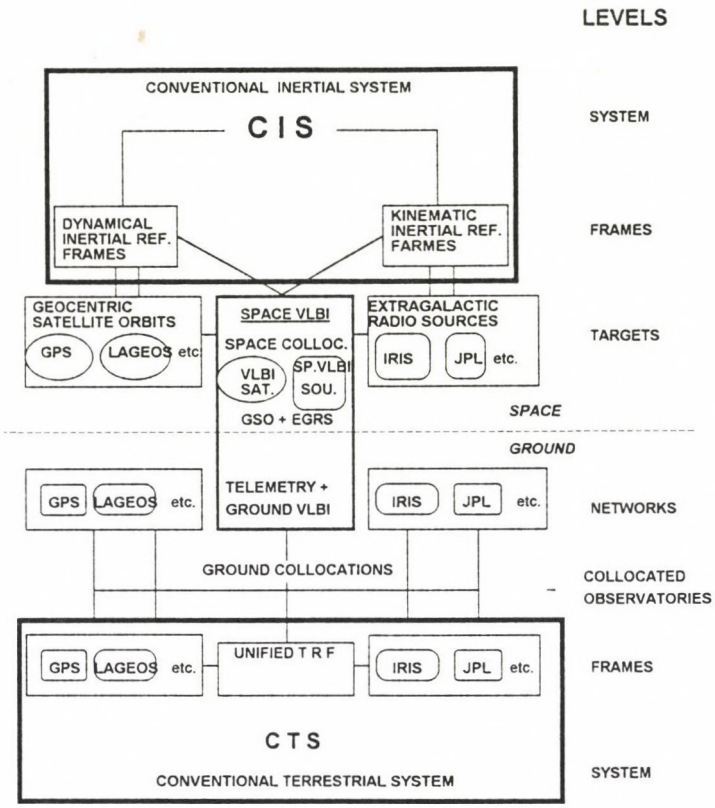

Fig. 2. Inter-relations between the Conventional Inertial System and the Conventional Terrestrial System

5. Milestones of concept development

In this paragraph a concise historical description is given about the main milestones of concept development and about the work which has been carried out in this field.

5.1 The beginning : QUASAT Workshop 1984 June

In June 1984, ESA organised a workshop on the QUASAT project where the application of space VLBI for astronomy and astrophysics were extensively discussed. There were no discussions, however, on geodetic, geodynamic applications during the meeting. In the proceedings a short note by Tang (1984) appeared who suggested to use the space VLBI observables to improve orbit determination of QUASAT. Shortly after the Workshop Fejes (1984) and independently Ádám (1985) suggested to use space VLBI for tying inertial and terrestrial reference frames. These suggestions were the initial elements of further studies on geodetic, geodynamic applications of space VLBI.

5.2 Concept formulation 1985-86

In January 1985, a working group was established in the SGO which participated in the ESA QUASAT Assessment Study (J Ádám, I Almár, T Borza, I Fejes, Sz Mihály) and formulated the concept of geodetic applications in more detail (ESA SC 85, 1985) The full concept was first presented at the Toulouse COSPAR meeting (Fejes et al. 1986). Here beyond the original suggestion of tying reference frames, the space VLBI data process flow and a simplified mathematical model was also presented. The paper also introduced the concept of "primary" and "secondary" tracking of space VLBI satellites.

Soon thereafter the IUGG recognised the importance of space VLBI studies and established an IAG special study group (SSG 2.109) in 1987, which dealt with the "Application of space VLBI in the field of Astrometry and Geodynamics" (Fejes 1992). The term of this SSG was extended in 1991 for another four years period (Ádám 1995).

5.3 Software developments and early concept tests by simulations 1987–89

The next step was to develop software in order to test and prove the concept by simulations. For this purpose the SGOFAKE orbit simulation program has been developed by Mihály and Szánthó at the FÖMI Satellite Geodetic Observatory in 1987–88. Using this software, the concept of Tang has been proved by simulations (Fejes et al. 1989), namely, that the delay, delay-rate observables can indeed be used for orbit determination improvement particularly when range rate data is also available simultaneously (which is the case anyway). Ulvestad (1992) at the JPL independently arrived to a similar conclusion. Orbit simulations using the ORAN program package at the Delft University of Technology by Borza et al. (1989) showed that the dominant part of the orbital errors of space VLBI satellites originates from errors in the solar reflectivity models.

Software developments in this field were presented at the COSPAR Workshop on "Space VLBI Related Software", Budapest, in October 1991 (Fejes and Schilizzi 1991).

5.4 Theoretical studies 1989–93

At the end of the eighties theoretical studies on space VLBI have been started at the Ohio State University, Department of Geodetic Science and Surveying which were supported by the NASA Goddard Space Flight Centre (Ádám 1990a). Two important works have been reported in this field in the series of the OSU Reports:

Ádám (1990b) has carried out a comprehensive estimability analysis for astrometric, geodetic and geodynamic parameters from space VLBI observations. Here he introduced a unified notation for the ground and the space VLBI parameters and worked out a detailed mathematical model for the delay and delay-rate observable on the ground to ground, ground to space and space to space baselines. He has also carried out an analytical rank defect analysis and pointed out that the UT1, the

ascending node of the space antenna and the right ascension of the radio source are inseparable parameters, if we consider data only on a ground to space baseline.

Kulkarni (1992) has continued this work, which culminated in a Ph.D. dissertation at the OSU with the title: Feasibility Study of Space VLBI for Geodesy and Geodynamics. He has developed software for estimability simulations and found that in an optimum observing scenario with Radioastron the polar motion parameters can be estimated to 0.7 milli arc second accuracy and the length of the day (LOD) to 0.05 ms accuracy from the ground to space delay observable. The results were also presented at he IAG General Meeting in Beijing (Kulkarni et al. 1993).

As a next step, the sensitivity of the satellite's orbital elements to observing configuration, to errors in other geodetic parameters and to solar radiation pressure was investigated (Kulkarni and Ádám 1993). Numerical simulations indicated that the VSOP satellite may be more suitable for geodesy programs than Radioastron, due to orbit characteristics.

6. The orbit problem

For space VLBI data processing and also for geodynamic, astrometric applications the orbit of the space telescope should be known with high precision. Therefore in this paragraph the orbit determination process, the associated terms, the requirements and simulation results are briefly reviewed.

6.1 What is orbit determination?

Orbit determination is *fitting* an *orbit* through *observations*. This definition has three basic elements : the orbit, the observations and the fitting.

The word *orbit* is often used in two different ways. One is the *real (true) orbit* which we never know exactly in the real life, only approximately. The other one is the *orbit model* which is used to describe the real (true) orbit approximately. The orbit model contains the dynamic relationships. Its starting point is the initial state vector at epoch t_0 (or the orbital elements) and from this a state vector at epoch t can be computed according to the equation of motion.

The word observations means *real data*, but accordingly we also need *model observational data* which is produced by the *measurement model*. The measurement model contains the geometric relationships and physical effects which effect the measurements, such as the tracking network, the data types, the atmospheric effects etc. The calculated data will be corrupted with model errors of course.

Fitting is carried out using an estimation scheme, where the orbit model and the measurement model parameters are improved or solved for and formal errors are calculated. This is where the residuals give a direct indication of the quality of our models and measurements. The solution for parameters can be devided in two groups (Dow 1988). *Arc dependent parameters* (*e.g.* state vector at epoch, force scaling factors, station measurement biases *etc.*) and *common parameters* (*e.g.* tracking station co-ordinates, coefficients of geopotential and tide models, Earth rotation parameters *etc.*) In a multi satellite and multi arc estimation scheme which

is implemented in several sophisticated program packages (*e.g.* Geodyn, Bahn) the parameters can be split into arc dependent and common parameters and estimated separately using a partitioned solution of the normal equations.

Paradoxically, orbit simulations provide the only method, where the "true orbit" is known exactly and the corruption of the measurements by errors can be precisely controlled. Therefore simulations are very effective in analysing error propagation and test or optimise measurement strategies. If there are no real data, orbit simulation is the only way to assess orbit accuracy as well.

6.2 What is orbit determination accuracy?

The key question to the astrometric, geodetic, geodynamic application of space VLBI is the orbit determination (OD) accuracy of the space telescope. But what is OD accuracy? Usually an accuracy figure is given like 10 cm accuracy or 50 m accuracy etc. This type of characterisation of the OD accuracy can be very misleading if the method or the definition is not stated clearly with the figure. Most common confusion occurs when no distinction is made between error of an orbit arc, and error of a single point at epoch t. An other case of misunderstanding is quite often the confusion between the orbit prediction and the orbit reconstruction errors. It should be made clear that there are lots of different types of OD errors. One possible classification is the following :

1. formal errors obtained from stochastic or deterministic models as
   - single point errors of reconstruction,
   - time dependent OD error along an orbit arc reconstruction,
   - average OD error of an orbit arc reconstruction,
   - time dependent OD prediction errors, and
2. OD errors from real observations as
   - orbit repeatability, in which positions are compared of orbits determined independently without using common measurements,
   - difference between orbit predictions and orbit reconstruction for the same epoch or time interval,
   - error estimation from baseline solution inter comparisons,
   - error estimation from space interferometric fringe characteristics.

From our point of view the average OD error along one or more reconstructed full orbit arcs should be considered. This characterises namely the accuracy of the realisation of the co-ordinate origin – the geocenter and also the overall orientation errors in the inertial space.

6.3 Requirements and numerical simulations

For space VLBI astronomical data processing purposes, orbit determination accuracy requirements are high, but not extreme. Radioastron or VSOP reconstructed average orbit arc errors should not exceed 40 m in order to operate smoothly the processing facilities. This requirement is achievable within the present tracking concepts relatively easily.

For geodynamic and astrometric measurement purposes the OD requirements are much more stringent if we want results comparable or better than achieved by ground techniques. Here the requirements are extreme indeed. The average OD error along one or more reconstructed full orbit arc should be in the decimetre or cm range for most applications. In only a few selected cases can this requirement be relaxed.

Several numerical OD simulation studies have been carried out in order to assess the achievable maximum OD accuracy of the presently planned space VLBI satellites using different types of tracking techniques, observing strategies and scenarios (*e.g.* Borza et al. 1989, Fejes et al. 1990a, Mihály 1990, Ulvestad 1992). One can conclude from these simulations, that Radioastron and VSOP are not well suited for geodynamic or astrometric experiments because orbits of these satellites cannot be determined with sufficient accuracy within the presently valid tracking concepts. Should be additional tracking devices (*e.g.* PRARE or GPS) installed on board the satellites, OD accuracy in the decimetre range could be achieved.

7. The need for further investigations

The simulation studies are not yet complete and a lot of questions remained open (see, for instance, the list of Kulkarni and Ádám 1993). No multi satellite and multi arc estimation scheme has been applied to space VLBI solutions where the parameters can be split into arc dependent and common parameters and estimated separately using a partitioned solution of the normal equations. What we would need most in this area is the integration of space VLBI delay, delay-rate observations into a sophisticated space geodesy program software package, like the Geodyn or the Bahn program. This would open the way for applications, like the combination of SLR data with space VLBI data in order to improve the stability of LAGEOS orbit solutions or improve gravity field parameter solutions.

Numerical investigations until now have neglected the space VLBI delay-rate observables, probably because in ground based VLBI this observable has limited use, due to the fact that it contains no information about the vector components parallel to the Earth's rotation axis. It should be pointed out, however, that the space VLBI delay-rate carries information for all the three components of the ground station co-ordinates (the only exception when the satellite's inclination is equal to zero). Consequently the space VLBI delay-rate can be as useful as the delay observable.

Application assessments of space VLBI sometime overlook the fact that this is not a "stand alone" technique, but rather an "add on" technique. This means that space VLBI should not be considered only as a single ground to space or space

to space baseline interferometer for geodynamic applications, because a ground network belongs always to the observing scenario. Therefore space VLBI can do everything which can be done by the ground based VLBI plus in addition exploit the space options. At parameter estimations for example parameters which can be estimated from the ground based data can be clearly separated from those parameters which are characteristic of the orbiting station. For example, the UT1 and the ascending node of the satellite should be estimated separately.

8. A proposal for concept demonstration

To prove the concept of reference frame unification using space VLBI, an observing proposal is being prepared for the Radioastron or the VSOP satellites. In this experiment using specially designed observation sequences of highly compact extra-galactic radio sources the consistency of terrestrial frame and its geocentric origin defined by the CTS locations of the participating ground VLBI stations at the one hand, and by the satellite orbit derived from space VLBI observations at the other hand, will be analysed. It is expected that by using real space VLBI data, new insights will be obtained concerning space VLBI applications even if we cannot expect direct improvement of the transformation parameters, due to limitations in OD accuracy. The experience obtained, we hope, can be used in the design of the next generation space VLBI satellites.

Three to five large telescopes distributed globally should take part in the space VLBI observations. The selected sources should satisfy the following conditions

- astrophysically interesting very compact objects,
- having large flux and flat spectra,
- they should be within the observable region of Radioastron or VSOP,
- large angular separation,
- are included in astrometric catalogues (JPL, NGS, IRIS *etc.*).

The first condition will ensure that the observations can be exploited for both astrophysical and geodetic interests. The second and third conditions are necessary for observability, while the fourth and fifth are necessary for geodetic analysis.

Four to six sources should be observed in one 24–36 hours session, each for 3–5 hours duration.

9. A next generation space VLBI satellite for geodynamics

Following the assessment of the limitations of Radioastron and VSOP for geodynamic and astrometric research, one can ask the challenging question: how should a space VLBI satellite, optimised for these area, look like?

Among the next generation space VLBI satellite concepts discussed at JPL in March 94, the only satellite fully dedicated to astrometry, geodesy and geodynamics was the SURF. SURF stands for Satellite for Unification of Reference Frames.

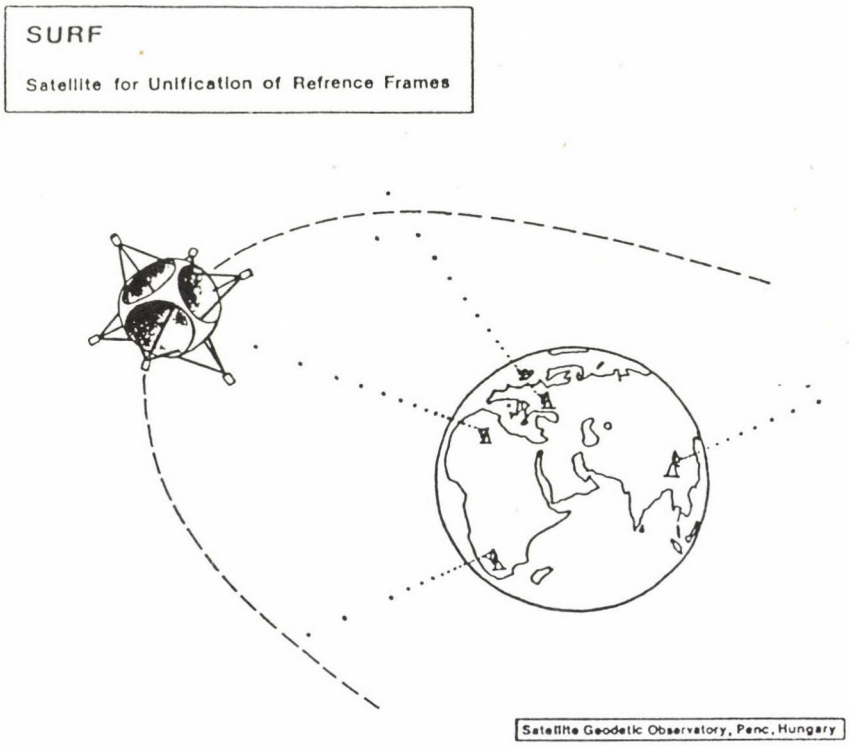

Fig. 3. An early version of the SURF concept (Fejes et al. 1990b)

Generally, geodetic satellites have regular shapes, small area over mass ratio, precise tracking capability resulting cm orbit accuracy and long operational lifetime. These characteristics are quite in contrast with the presently planned space VLBI satellites RADIOASTRON or VSOP. They all have irregular shape, large area over mass ratio, 50–100 m OD accuracy and short lifetime. To make things even worse they have slow slew rates and limited capability in switching many times between sources in large angular distances, which would be desirable for geodetic observing programs.

SURF is a space VLBI satellite concept without the above mentioned limitations (Fejes et al. 1990b). The basic design characteristic is a multiple beam phased array antenna which is built on a regularly shaped satellite structure (*e.g.* a tetrahedron or hexahedron). The beams can be switched electronically, therefore no slewing is required. The attitude can be fixed in the inertial space. For elimination of the disturbing effects of non-gravitational forces a "discos" type drag free system or a less expensive on board micro accelerometer is proposed.

The primary goal of SURF, according its name, is the unification of reference frames, but additional interesting applications can also be considered. Continuous real time monitoring of the Earth orientation parameters, or geocenter position could be carried out with this system. For high precision radio astrometry large

angular separation phase referencing could be realised which has cosmological applications as well.

10. Conclusion

Space VLBI is an emerging new space geodetic technique with applications in the field of geodesy, geodynamical research and astrometry. In principle it is capable of tying the Inertial (Celestial) and the Terrestrial reference frames directly. The space VLBI observables the delay and the delay-rate can be applied as a new type of tracking data for orbit determination improvement of the satellite, as has been suggested by Tang (1984). The full potential of space VLBI in this field can be exploited only if high precision orbit determination methods and new satellite tracking concepts will be introduced. Further studies of parameter estimation schemes are necessary in order to assess the importance of the space VLBI delay-rate observable. It has been pointed out that multisatellite multiarc partitioned solutions of arc dependent and common parameters of interest should be further investigated including ground as well as space VLBI delay and delay-rate observables.

Acknowledgements

Contribution of the SGO space VLBI team to this review is highly appreciated. This work was financially supported be the Hungarian Research Fund under contract OTKA–3/1925.

References

Ádám J 1985: Presentation at the Hungarian QUASAT Team meeting, 1985

Ádám J 1989: A Possible Application of the Space VLBI Observations for Establishment of a New Connection of Reference Frames. In: Theory of Satellite Geodesy and Gravity Field Determination. F Sansó and R Rummel eds, Lecture Notes in Earth Scences, p 459, Springer Verlag, Berlin - Heidelberg

Ádám J 1990a: *Geodézia és Kartográfia*, 6, 426.

Ádám J 1990b: Estimability of Geodetic Parameters from Space VLBI Observables. Report No. 406. Department of Geodetic Science and Surveying. The Ohio State University, Columbus, Ohio, USA

Ádám J 1995: Application of Space VLBI in the Field of Astrometry and geodynamics. Report of the president, IAG. SSG. 2. 109. Travaux de L'Association Internationale de Geodesie (in preparation)

Borza T, Fejes I, Ambrosius B A C 1989: *Adv. Space Res.*, 10. 321.

Dow J M 1988: Ocean Tides and Tectonic Plate Motions from Lageos. Deutsche Geodatische Komission bei der Bayerischen Akademie der Wissenschaften, Reihe C, Heft No. 344.

Fejes I 1984: Letter to R. T. Schilizzi, concerning geodetic applicatons of QUASAT. ESA, QUASAT Documents No. 81.

Fejes I 1990: Tying extragallactic celestial and dynamical frames by space -VLBI. Presented at the IERS Workshop, 23-27 April, Paris, France

Fejes I 1992: Application of Space VLBI in the Field of Astrometry and geodynamics. Report of the president, IAG. SSG. 2. 109. Travaux de L'Association Internationale de Geodesie. Tome 29. p. 189. Ed. P. Willis. Paris

Fejes I, Mihály Sz 1991: *Adv. Space Res.*, 11, 429.

Fejes I, Schilizzi R T 1991: The COSPAR Workshop on Space-VLBI Related Software, (Report to COSPAR)

Fejes et al. 1985: QUASAT Potential Geodynamic Applications. Contribution for QUASAT; a Space VLBI Satellite, Assessment Study. ESA SCI(85)5.

Fejes I, Almár I, Ádám J, Mihály Sz 1986: Space VLBI: *Adv. Space Res.*, 6, 205.

Fejes I, Borza T, Mihály Sz, Szánthó L 1990a: Orbit determination accuracy improvement by space-VLBI observables as tracking data. in Global and Regional Geodynamics. Proc. IAG Symp. 101. p. 64. P Vyskocil, C Reigber, P A Cross eds, Springer Verlag

Fejes I, Almár I, Ádám J, Mihály Sz 1990b: Space VLBI: A new technique for unufication of reference frames. in Earth Rotation and Reference Frames. Proc. IAG symp. 105. p. 158. C Boucher, G A Wilkins eds, Springer Verlag

Inoue M 1993: Space VLBI Project, VSOP. in Sub-arcsecond Radio Astronomy, p. 434. R J Davis and R S Booth eds, Cambridge University Press

Kardashev N S, Slysh V I 1988: The Radioastron Project. in The Impact of VLBI on Astrophysics and Geophysics. M J Reid and J M Moran eds, Kluwer Academic Publishers, Dordrecht

Kardashev N S, Andreyanov V V, Gvamichava S F, Likhachev S F, Slysh V I 1994: Orbiting Very Long Base Interferometer (OVLBI) for Radio and Optical Astronomy. IAF - 94 - Q.2.335. Jerusalem, Israel

Kulkarni N M 1992: Feasibility Study of Space VLBI for Geodesy and Geodynamics. Report No. 420. Department of Geodetic Science and Surveying. The Ohio State University, Columbus, Ohio, USA

Kulkarni N M, Ádám J 1993: *Acta Geod. Geoph. Mont. Hung.*, 28, 441.

Kulkarni N M, Mueller I I, Schaffrin B, Ádám J 1993: A Feasibility Study of Space VLBI for Geodesy and Geodynamics. Paper presented at the IAG General Meeting, Beijing, China

Mihály Sz 1990: Radioastron tracking concept with PRARE system. Paper presented at the 10th Radioastron review meeting, Espoo, Finland

Levy G S et al. 1986: *Science*, 234, 187.

Schuh H 1994: *Acta Geod. Geoph. Hung.* (this issue)

Tang G 1984: A short note on high precision navigation of QUASAT. in ESA SP213.

Ulvestad J S 1992: Use of the VLBI Delay Observable for Orbit Determination of Earth — Orbiting VLBI Satellites. TDA Progress Report 42–110, Jet Propulsion Laboratory, Pasadena, California, USA

Ulvestad J S 1994: Planning for future Space VLBI missions. Summary of meeting held at Jet Propulsion Laboratory

Acta Geod. Geoph. Hung., Vol. 29(3-4), pp. 457–458 (1994)

Book reviews

F HENNECKE, H J MECKENSTOCK, G PALLMER: *Vermessung im Bauwesen.* Neunte überarbeitete Auflage. Dümmlers Verlag, Bonn, 1993, 176 pages, 158 figs, 21 tables

This book is a well-written text-book especially for building engineers but it is very useful for surveyors and university students in both fields, too.

In spite of the fact that the book is limited to the fundamentals of surveying, it describes all of the significant measuring methods and refers also to the newest instruments. A lot of examples help to understand the different measuring methods for both practicians and students. In this book all of the measuring methods from the simplest to the most up-to-date ones are found which are necessary to establish a building from planning to the control and supervision measurements of ready buildings.

The book consists of four chapters. The first chapter deals with point marking, alignments, pegging out of boundaries and building lines, squaring by means of cord triangles and prismatic reflectors, plotting, planimetry and area computation.

The second chapter describes levelling by means of bubble levels, hydrostatic levels and level instruments. It familiarizes with the handling, examination and readjustment of the level instruments.

The third chapter gives the definition of horizontal and vertical angles, describes the construction and handling of the theodolites and introduces to the optical and electrical tachymetry.

The fourth chapter describes different plumbing and special laser instruments used in the building industry. Different surveying methods are given for measurements of building constructions and walls as well as for measurements and calculations of curve ranging and controlling. Some examples are also given for measurements in road construction and structural engineering.

Gy Mentes

S HEITZ, E STÖCKERT-MEIER: *Grundlagen der physikalischen Geodäsie.* Zweite, durchgesehene Auflage. Dümmlers Verlag, Bonn, 1994, 436 pages, 57 figs

The book presents the fundamental principles of geodesy in a new conception according to the recent state of the physics. It is done by means of physical-mathematical modelling of geodetic principles. In accordance with this, the first part of the book gives an introduction to the fundamentals of nuclear physics and the three and four dimensional geometric models. Thereafter, the elements of the classical field theory and quantum mechanics are described to help a deeper understanding of the theory of geodesy.

The second part of the book summarizes the fundamental geodetic methods "from the measuring-rod to the nuclear spin resonance" as a collection of examples of the applied physics. The examples include the application of the methods of the classical physics of

solid bodies, fluids, the electrodynamics as well as the application of the quantummechanical nuclear theory. This latter is the basis of the modern time and distance measurements made by means of atomic clocks and instruments based on laser interferometry.

The book is written for students but it can be very well used by scientists and practicians in field of geodesy and moreover by physicists.

Gy Mentes

ACTA GEODAETICA et GEOPHYSICA
A Quarterly Journal
of the Hungarian Academy of Sciences

VOLUME 29

AKADÉMIAI KIADÓ, BUDAPEST
1994

CONTENTS

Numbers 1–2

Book reviews

Numbers 3–4

Book reviews

AUTHOR INDEX

KEYWORD INDEX

PRINTED IN HUNGARY

Akadémiai Kiadó és Nyomda, Budapest

Manuscripts should include the following components which should be presented in the order listed (tables and illustrations should be separated in case of manuscripts on disk, too).

1. Title, name(s) of the author(s), affiliation, dateline, abstract, keywords
2. Text, acknowledgements
3. References
4. Footnotes
5. Legends
6. Tables
7. Illustrations

1. The *affiliation* should be as concise as possible and should include the complete mailing address of the authors. The *date of receipt* will be supplied by the editors. The *abstract* should not exceed 250 words and should clearly and simply summarize the most important methods and results. 5–10 significant expressions describing the content are used as *keywords*. Authors may recommend these keywords.

2. The *text* should be in English and as short and clear as possible. In case of typewritten manuscripts, please note the following:

— avoid possible confusion between o, O (letters) and 0 (zero), l (letter) and 1 (one), ν (Greek nu) and u, v (letters), etc.
— explain ambiguous and uncommon symbols by making marginal notes in pencil
— formulas are to be numbered consecutively with the number in parentheses to the right of the formula. Text references to the equations may then be made by the number in parenthesis. The word equation in this context is to be abbreviated to Eq. and Eqs in the plural
— the International System of Units (SI) should be used.

3. *References* are accepted only in the Harvard system. Citations in the text should be as:

...(Bomford 1971)...or Bomford (1971)...
...(Brosche and Sündermann 1976)...
...(Gibbs et al. 1976b)...

The list of references should contain names and initials of all authors (the abbreviation et al. is not accepted here): for journal articles, year of publication, the title of the journal in abbreviated form, volume number, first and last page. For books or chapters in books, the title is followed by the publisher and place of publication.

All items must appear both in the text and in the list of references.

4. *Footnotes* should be typed on separate sheets.

5. *Legends* should be short and clear. The place of the tables and figures should be indicated in the text, on the margin.

6. *Tables* should be numbered serially with Roman numerals. Vertical lines are not used.

7. All the *illustrations* should contain the figure number and author's name in pencil on the reverse. The most important point with figures is clearness. Photographs and half-tone illustrations should be sharp and well contrasted. Colour photographs will be accepted, but the extra cost of reproduction in colour must be borne by the authors (approx. US$ 260 per page).

Only original papers will be published and a copy of the Publishing Agreement will be sent to the authors of papers accepted for publication. Manuscripts will be processed only after receiving the signed copy of the agreement. Information is sent to the first author if no other wish is expressed. Fifty offprints are supplied free of charge, for further copies please contact the Editorial Office.