MAGYAR ÁLLAMI
EÖTVÖS LORÁND
GEOFIZIKAI INTÉZET

GEOFIZIKAI
KÖZLEMÉNYEK

ВЕНГЕРСКИЙ
ГЕОФИЗИЧЕСКИЙ
ИНСТИТУТ
ИМ Л. ЭТВЕША

ГЕОФИЗИЧЕСКИЙ
БЮЛЛЕТЕНЬ

GEOPHYSICAL

T R A N S A C T I O N S

EÖTVÖS LORÁND GEOPHYSICAL INSTITUTE OF HUNGARY

CONTENTS

Practical definition of robustness	*F. Steiner, B. Hajagos*	193
Investigations concerning resistance — importance of the choice of the formula determining the scale parameter	*B. Hajagos, F. Steiner*	211
Comment on an old dogma: 'The data are normally distributed'	*P. Szűcs*	231
Comparison of the Karhunen-Loève stack with the conventional stack	*L. Bruland*	239
Interconnecting gravity measurements between the Austrian and the Hungarian network	*G. Csapó, B. Meurers, D. Ruess, G. Szatmári*	251

VOL. 38. NO. 4. JUNE 1994. (ISSN 0016-7177)

BUDAPEST

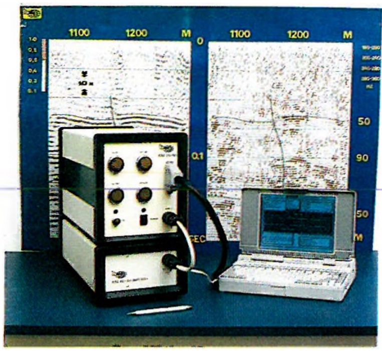
TARTALOMJEGYZÉK

A robusztusság mérőszámának definíciója	*Steiner F., Hajagos B.*	210
Rezisztencia-vizsgálatok. A skálaparaméter-formula megválasztásának fontossága	*Hajagos B., Steiner F.*	229
Megjegyzés egy régi dogmához: „Az adatok Gauss-eloszlásúak"	*Szűcs P.*	238
A Karhunen-Loève és a hagyományos stacking eljárás összehasonlítása	*L. Bruland*	249
Összekapcsoló gravitációs mérések Ausztria és Magyarország gravimetriai alaphálózatai között	*Csapó G., B. Meurers, D. Ruess, Szatmári G.*	259

ESS - 03 - 24

PC based 24 channel portable seismograph

ESS-03 was designed for high-resolution subsurface imaging using up-to-date electronics and computer technology. Compared with excavation or dense drilling networks, it offers a highly cost-effective solution for geotechnical problems.



It is an excellent tool

- for raw material prospecting;
- for determining static corrections for surface seismic methods;
- for acquiring information when designing large scale foundations such as factories, dams or roads;
- for seismic investigation of well-sites;
- for in-mine exploration;
- for detection of destructive building-vibrations.

ESS-03 can be used not only for rapid and inexpensive data acquisition, but the built-in PC and software package offer excellent on-site processing, quick-look interpretation, and visual display of the results.

Specifications:

Frequency response: 1-8000 Hz
Resolution: 12 bit A/D+42 dB IFP
Built-in electronic roll along switch
Microprocessor control 80386-SX-20
Data recording on floppy disc 1.44 MB, hard disc 80 MB
Menu-driven operating commands
LCD display with VGA resolution
Self-testing programs and parameter checking
Weight: 10 kg
Dimensions: 200*220*450 mm

Founded in 1919

Eötvös Loránd Geophysical Institute of Hungary

Budapest, XIV. Columbus u. 17-23.
Letters: H-1440 Budapest, POB. 35.
Phone: (36-1) 184-3309
Fax:(36-1) 163-7256
Telex: (61) 22-6194 elgi h
E-mail: H6123 TIT@ELLA.HU

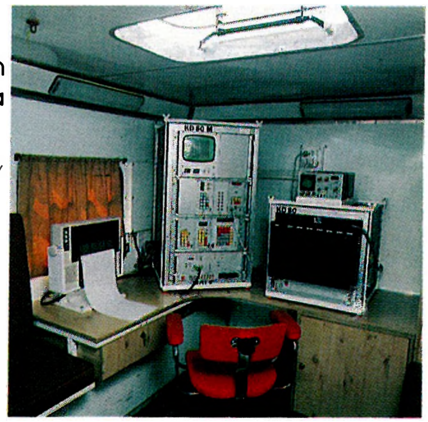
DON'T BUY EQUIPMENT OR SERVICES UNTIL YOU KNOW THE FACTS

ELGI's Well Logging Divison has put its 25 years of experience to work again in the new line of well logging technology in

water,
coal,
mineral,
geotechnical

prospecting

HERE'S WHAT WE OFFER



- Complete series of surface instruments from portable models to the PC controlled data logger
- Sondes for all methods: electrical, nuclear, acoustic, magnetic, mechanical, etc.
- Depth capacity down to 5000 m
- On-site or office computer evaluation
- International Metrological Base for calibration to true petrophysical parameters
- Training and in-house courses
- Design laboratory for custom-tailored assemblies

Just think of us as the scientific source of borehole geophysics you may never have heard of

SALES ✤ ✤ ✤ ✤ RENTALS ✤ ✤ ✤ ✤ SERVICES

Well Logging Division of ELGI
POB 35, Budapest, H-1440 Hungary
Phone: (361) 252-4999, Telex: 22-6194,
Fax: (361) 183-7316

ASSOCIATION OF HUNGARIAN GEOPHYSICISTS

23th ANNUAL MEETING 13–14 October 1994,

SÁROSPATAK

CALL FOR PAPERS

The Association of Hungarian Geophysicists in cooperation with the Hungarian Geological Society holds its 23th Annual Meeting on October 13-14, 1994 in the historical city of Sárospatak

The topics of the Annual Meeting are:

- Geological-geophysical research in the Pannonian Basin
- Recent advances in geophysical exploration techniques

The deadline of submitting abstracts in English or Hungarian not exceeding 300 words is September 1, 1994. Abstracts should be taped in a 15 x15 cm space as follows:

TITLE
Authors and Institutions
Text

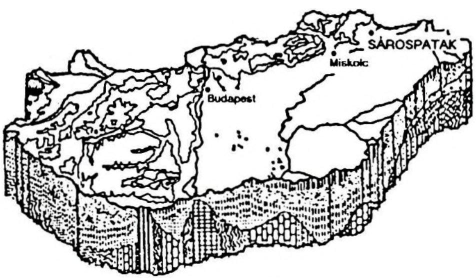
Both poster and oral papers are accepted. The official languages are English and Hungarian.

For more information please write to our postal address:

Association of Hungarian Geophysicists
H-1371 Budapest,
POB 433
Hungary
Fax and phone: 36-1-2019815

or to the

Local Organization Committee
1994 Meeting of Assoociation of
Hungarian Geophysicists
Miskolc University
Geophysical Department
H-3515 Miskolc-Egyetemváros
Fax and phone: 36-46-361936



REGISTRATION

Registration fee: appr. 6000 HUF (60 USD).

Hotel expenses: appr. 30 USD/night/person.

Registration should be done in letter addressed to the Association of Hungarian Geophysicists before September 15, 1994.

Your registration letter should be accompanied with a copy of the bank transfer order about 60 USD for

MHB Rt
H-1052 Budapest
Gerlóczy utca 11.
323-10195

The Annual Meeting takes place in the City Cultural Centre, where exhibitiors are welcome, too.

INVITATION

The Association of Hungarian Geophysicists decided at its annual meeting to establish the "Foundation for Hungarian Geophysicists" and elected its first Advisory Board for 3 years. The foundation has been started with a moderate initial capital of 300 000 HUF, which has by now increased to more than 3 million and it is open for everybody.

The aim of the foundation is to help Hungarian geophysicists. There are two main target groups whose application for grants will be accepted with preference: young geophysicists needing assistance (travels, participation at conferences, publications,post-graduate education etc.) at the beginning of their professional life as well as retired and unemployed colleagues whose economic and social position became especially unfavourable.

The nine members of the Advisory Board invite everybody to join this foundation; donations should be communicated with the Board. Organisations and persons donating sums exceeding the initial capital will have the opportunity to delegate representatives into the Board. Detailed information is available at the following address:

Advisory Board of the
"Foundation for Hungarian Geophysicists"
H–1371 Budapest, P.O.B. 431
Budapest, I., Fő u. 68.
Telephone 201-2011/590
Telex 22-4343
Telefax 156-1215

GEOPHYSICAL TRANSACTIONS 1993
Vol. 38. No. 4. pp. 193- 210

PRACTICAL DEFINITION OF ROBUSTNESS

Ferenc STEINER* and Béla HAJAGOS*

The paper defines the index of robustness (r) as a weighted average efficiency belonging to a statistical estimating procedure. The weights are the occurrence probability densities of the various model types which can be accepted as adequate for a given discipline. The value r can simultaneously take very different probability distribution types into consideration. Instead of deciding categorically 'robust' — 'not robust' the examples show robustnesses in the interval from r=36 % to r=96 %. In geophysics practice quantitative comparisons are unavoidable.

Some of the figures demonstrate the original efficiency curves ($e(t)$-s), figuring in Eq. 12 given for r, too, thereby enabling so that the changes in the efficiencies can be analysed in detail.

Keywords: robustness, index of robustness, statistical efficiency, probability density, error distribution

1. Introduction and preliminaries

The definition of robustness by theoretical experts of mathematical statistics [see e.g. HAMPEL et al. 1986] does not result in numerical values (thereby facilitating the near-optimum choice of the statistical algorithm,) and/or it belongs to very narrow (or even infinitesimal) neighbourhood of a distribution type. Let one comment be cited from the Summary of the article of DONOHO and LIU [1988], i.e., from a paper written by mathematicians: 'Of course, this robustness is formal because μ-contamination neighbourhoods may not be large enough to contain *realistic departures from the model*' (enhancement was not made in the original text). Here we propose the acceptance of a measure of robustness which is also suitable for practical applications. The discipline of geophysics particularly needs quantitative comparisons made on the grounds of large type-intervals.

* University of Miskolc, Department of Geophysics, H-3515 Miskolc-Egyetemváros
Manuscript received: 20 July, 1993

1.1. Various estimations of the location parameter (a brief enumeration)

A chronological enumeration of different statistical procedures is given below with some comments. In every case below the task is to determine (estimate) on the grounds of a given sample the most characteristic value of the actual probability distribution (this is naturally the symmetry point if the distribution is symmetrical). — In the first and second case it is impossible to determine how old these estimations are (at least two hundred years old):

arithmetic mean		
sample median		
α-trimmed mean	1821	see e.g. FEGYVERNEKI [1992] — but may be as old as the arithmetic mean itself
Hodges-Lehmann estimate	1963	
Huber estimate	1964	
M^*-estimate	1965	this is the minimum place of the P^*-norm, see Eq. 36 in HAJAGOS and STEINER [1991]
M-estimate	1973	this is the minimum place of the P-norm, see Eq. 30 in HAJAGOS and STEINER [1991]
L_p-estimate ($p>0$, $p\neq1$, $p\neq2$)	1990	this is the minimum place of the generalized L_p-norm, see e.g. TARANTOLA [1987] (it is well known that for $p=1$ we would get the sample median and for $p=2$ the arithmetic mean). The date of L_p is given here in accordance with SOMOGYI and ZÁVOTI [1990], as the authors do not know any earlier article in applied statistics that deals in detail with a p value which is not an integer.

Where no explanation is given or no reference is cited, see e.g. the monograph HUBER [1981] or the original papers HODGES, LEHMANN [1963] and HUBER [1964] (in the present paper 'Proposal 2' of HUBER is treated). It should be mentioned that both M^*- and M-estimates are called '*m*ost *f*requent value' therefore in the case of more unknown parameters the corresponding statistical algorithm is called '*MFV* procedure' (and the simple estimate can also be called '*MFV*-value' instead of M- or M^*-estimate). Some characteristics of the M-estimate are given in a comprehensive manner in the Table at the end of the book STEINER (ed.) [1991]; in the bibliography of this book are cited the paper and thesis where M- and M^*-estimates were first defined.

1. 2. How to calculate the efficiencies

If certain conditions for the density function are fulfilled and the sample range (n) tends to infinity, the distribution type of the estimates becomes

Gaussian (see e.g. HUBER [1981]; the overwhelming majority of the following can also be found in the same monograph). This means that the dispersion can adequately be characterized by the variance ($VAR = \sigma^2$) of the estimates. To be independent of n, it is convenient to introduce the notion 'asymptotic variance' (A^2) with the equation

$$A^2 = \lim_{n \to \infty} n.\sigma^2. \tag{1}$$

It is often easy to find statistical algorithm that leads to the minimum asymptotic variance ($A^2_{\min}$) for the probability distribution in question.

The efficiency (e) of an arbitrary statistical algorithm having an asymptotic variance A^2 for a well defined probability distribution, is defined as

$$e = \frac{A^2_{\min}}{A^2}, \tag{2}$$

(where $A^2_{\min}$ obviously belongs to the same probability distribution). Often e is expressed in per cent.

Eq. 2 says that e per cent of the data would be sufficient for the same estimation accuracy if we were to use an optimum algorithm instead of the one actually used. In practice therefore, from the viewpoint of the cost of measurements it is of crucial importance that the statistical efficiency e is as great as possible.

How does one calculate the asymptotic variance A^2? If the so-called influence function $IC(x)$ is known for the statistical algorithm and for the actual probability distribution defined by the density function $f(x)$, A^2 can be determined as

$$A^2 = \int_{-\infty}^{\infty} IC^2(x) \,.\, f(x)\,\mathrm{d}x. \tag{3}$$

If primarily the $\psi(x)$-function is given (the ψ-function plays a key-role in the best elaborated part of the robust statistics), the influence function can be calculated as

$$IC(x) = \psi(x) \,.\, \left[\int_{-\infty}^{\infty} \psi'(y) \,.\, f(y)\,\mathrm{d}y\right]^{-1}. \tag{4}$$

In some cases A^2 can be calculated directly by means of a simple formula. *Table I* gives either A^2-formulas, or IC-, or ψ-functions (always choosing the simplest alternative) for the statistical procedures yet enumerated in 1.1. (for probability distributions symmetrical to the origin). The asymptotic variance A^2 can be calculated in every case without difficulty.

1. 3. The supermodel $f_a(x)$

The supermodel $f_a(x)$ was introduced by the density functions

$$f_a(x) = \Gamma\left(\frac{a}{2}\right) \cdot \Gamma^{-1}\left(\frac{a-1}{2}\right) \cdot \pi^{-\frac{1}{2}} \cdot (1+x^2)^{-a/2} \quad (a>1) \tag{5}$$

[see e.g. CSERNYÁK, STEINER 1991]; this standard form can be generalized by replacing x by $(x-T)/S$ and dividing by S (T and S are the parameter of location and parameter of scale, respectively). Here, we mention some types of this supermodel: the distribution type a=5 is called geostatistical or simply statistical having clearly the density function

$$f_{st}(x) = 0.75(1+x^2)^{-2.5} \ . \tag{6}$$

(according to DUTTER [1987] this is a very commonly occurring distribution type in geostatistics, but in the opinion of the authors its acceptance as a model is justified more generally in the practice of statistics). If short flanks are guaranteed, the so-called Jeffreys-type (a=9) can serve as an adequate model for the distribution:

statistical procedure (estimate)	characterization of the procedure from the view-point of the asymptotic variance of the estimates
arithmetic mean	$IC(x) = x$, *i.e.*, $A^2 = VAR = \sigma^2$ (VAR means the variance, σ the scatter of the mother distribution)
sample median	$A^2 = \dfrac{1}{4 \cdot f^2(0)}$
α-trimmed mean	$IC(x) = \begin{cases} \dfrac{1}{1-2\alpha} F^{-1}(\alpha), & \text{if } x < F^{-1}(\alpha) \\ x/(1-2\alpha), & \text{if } \lvert x\rvert \le F^{-1}(1-\alpha) \\ \dfrac{1}{1-2\alpha} F^{-1}(1-\alpha) & \text{if } x > F^{-1}(1-\alpha) \end{cases}$
Hodges-Lehmann estimate	$A^2 = \dfrac{1}{12\left\{\int_{-\infty}^{\infty} [f(x)]^2 \, dx\right\}^2}$

Huber-estimate		$A^2 = \dfrac{\int_0^{cS} x^2 f(x)\mathrm{d}x + (cS)^2 \int_{cS}^{\infty} f(x)\mathrm{d}x}{2\left(\int_0^{cS} f(x)\mathrm{d}x\right)^2}$; the value S fulfils the condition $\frac{1}{S^2}\int_0^{cS} x^2 f(x)\mathrm{d}x + c^2 \int_{cS}^{\infty} f(x)\mathrm{d}x = \int_0^{c} f_G(x)\,\mathrm{d}x + c^2 \int_c^{\infty} f_G(x)\,\mathrm{d}x$; ($f_G(x)$ represents the Gaussian density function)
M^*-estimate	most frequent values	$\psi_{M^*} = \dfrac{x}{[3\,(k\varepsilon)^2 + x^2]^2}$ The dihesion ε fulfils in both cases the condition $\int_{-\infty}^{\infty} \dfrac{3x^2 - \varepsilon^2}{[\varepsilon^2 + x^2]^2} f(x)\,\mathrm{d}x = 0$
M-estimate		$\psi_M = \dfrac{x}{(k\varepsilon)^2 + x^2}$
L_p-estimate		$\psi_p(x) = sign\ x \,.\, \lvert x \rvert^{p-1}$

Table I. Charaterization of some statistical procedures
I. táblázat. Statisztikai eljárások jellemzése

$$f_J(x) = \frac{35}{32}(1+x^2)^{-4.5} \quad . \tag{7}$$

It can easily be shown that for the supermodel $f_a(x)$ the minimum asymptotic variance is given by the simple formula

$$A^2_{\min} = \frac{a+2}{a(a-1)} \quad . \tag{8}$$

For integer values of a we get Student distribution types characterized by $(a-1)$ degrees of freedom; the so-called Jeffreys interval of distribution types defined by $6 \le a \le 10$ was primarily given also by limits expressed as 5 and 9 degrees of freedom. Obviously $(a-3)^{-1/2} . f_a\,[x\,.\,(a-3)^{-1/2}]$ tends to the standard Gaussian density function $f_G\,(x) = (2\pi)^{-1/2}.\ \exp(-x^2/2)$ if $a \to \infty$. For a=2 we trivially get the Cauchy distribution.

The probability density functions of the Cauchy-, (geo)statistical-, Jeffreys- and Gaussian type are shown in *Fig. 1*; in all four cases the probable error (i.e., the semi-interquartile range q) equals unity (choosing the parameter of scale S always appropriately). We find these curves visually very similar — although statistical procedures can behave very differently if the actually occurring error distribution type is, say, geostatistical instead of Gaussian. Some statistical procedures (first of all the classical ones) are extremely sensitive to the behaviour of the flanks but Fig. 1 (and other such commonly used visualizations, too) does not characterize these parts of the distributions very well (the small values of $f(x)$ at both ends of the $f(x)$-curve can result in misjudging the weight of the flanks measured in the occurrence probability of x of the neglected sides). The authors therefore prefer the plotting of the density function versus $F(x)$-curve since this does not depend upon the parameter of scale and, moreover, it enhances the behaviour of the tails (as usually,

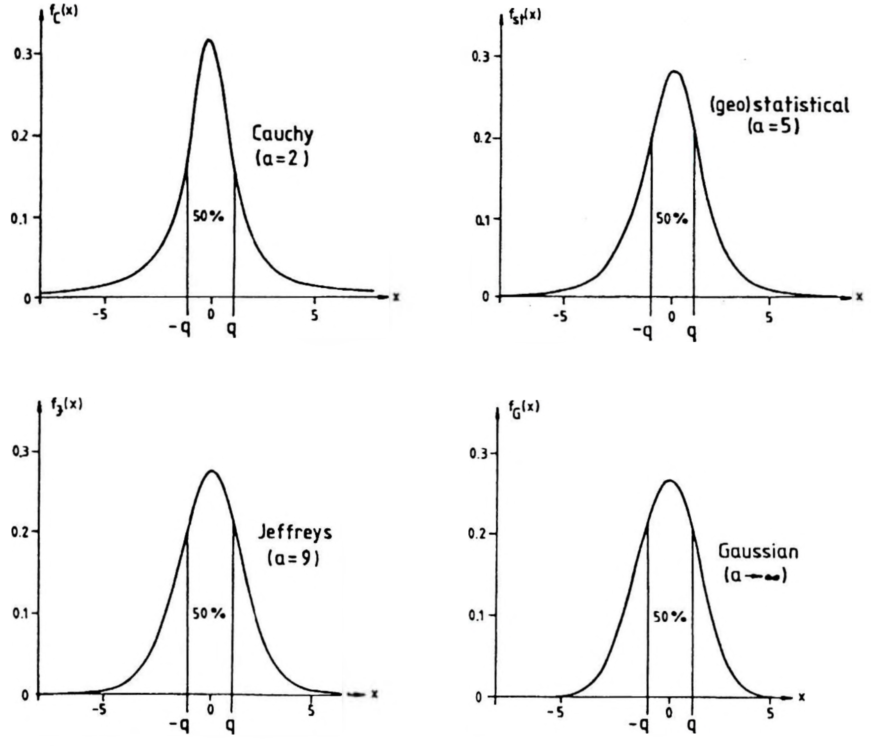


Fig. 1. Four probability density functions of x from the supermodel $f_a(x)$ (see Eqs. 5-7). With appropriately chosen parameter of scale the probable error (semi-interquartile range) q equals unity in every case

1. ábra. Az $f_a(x)$ szupermodell négy valószínűségsűrűség-függvénye (ld. az 5-7 egyenleteket). A skálaparaméter megfelelő választásával a q valószínű hiba (azaz az interkvartilis félterjedelem) egységnyi nagyságú mind a négy esetben

$F(x) = \int_{-\infty}^{x} f(x)\,dx$ represents the distribution function). It is advantageous to 'norm' the densities to their maximum value; this was done in *Fig. 2.* where the great difference between the flanks and the general features of the Cauchy-, (geo)statistical and the Gaussian type are visualized. (For Laplace- and uniform distributions the $f(x)/f_{max}$ versus $F(x)$-'curves' consist of straight lines, see the dashed lines in Fig. 2.) It should be mentioned, too, that Fig. 2 clearly shows: that there are distribution types that are characterized by much heavier flanks, than those of the Cauchy-type.

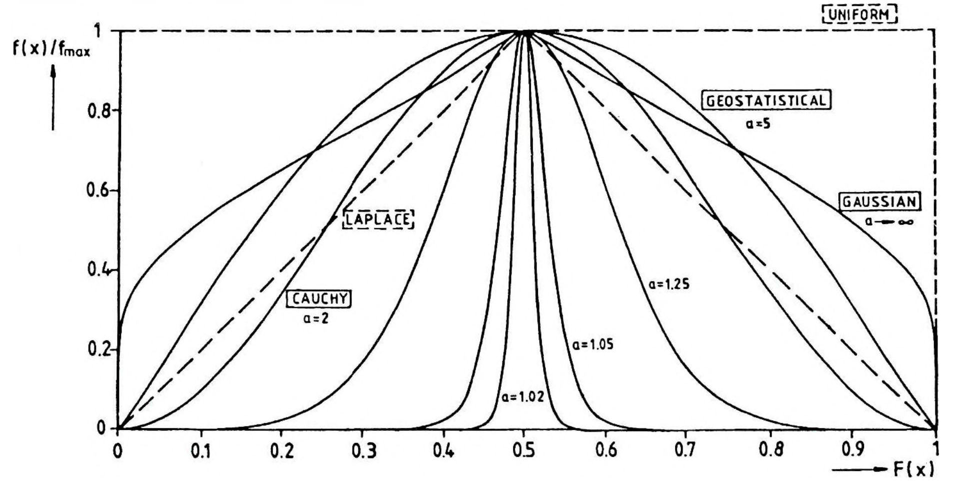


Fig. 2. Probability densities (normed to their maximum value) versus distribution function $F(x)$ as a visualization which is independent both of the parameter of scale and the parameter of location. The different behaviour of the flanks is satisfactorily accentuated here

2. ábra. Maximális értékükre normált valószínűségsűrűségek az $F(x)$ eloszlásfüggvény értékeinek a függvényében. Ily módon mind a hely-, mind a skálaparamétertől független görbéket nyerünk, amelyek jól láthatóan fejezik ki az eloszlások szárnyainak különböző viselkedését

2. Quantitative characterization of robustness

2. 1. Inherent supposition of the maximum likelihood-principle from the practical viewpoint. Occurrence probability densities ($f_J(t)$, $f_D(t)$) of type t distribution.

Statistical procedures can be derived on the basis of the *m*aximum *l*ikelihood-principle (but these procedures are usually applied *not only* for the distribution type which was supposed in the first step). The ML-principle originally *postulates that the type of the actual distribution is a priori known* (with probability = 1). Good Heavens! Indeed, the statistician working in a practical environment *never* a priori knows the type of the actual probability distribution exactly.

Let us suppose, however, just for a moment, that this supposition is fulfilled and this a priori known type is the Jeffreys distribution (see Eq. 7). It is easy to verify that the maximum likelihood method results in the calculation of the M-estimate with k=3. This latter value is a slightly rounded one consequently the efficiency is not exactly 100 % but 'only' 99.9999 %. Obviously the practical statistician would tolerate perhaps a 'loss' of say, 2–3 %, too (and a loss of 1% would certainly be accepted as insignificant even by the most rigorous mathematician).

The question arises if other estimation procedures can approximate the maximum efficiency or not. *Fig. 3* shows the efficiencies of the L_p-estimates versus p for the Jeffreys distribution; if p=1.6 is chosen the efficiency is greater than 98 %. It can be demonstrated in a similar way that the Huber estimate has maximum efficiency for the Jeffreys distribution if c=1.4 is chosen. Briefly, the efficiencies of six estimating procedures (to an accuracy of two decimals) are summarized in *Table II.*

statistical procedure	efficiency for the Jeffreys distribution
M-estimate; k=3	100.00%
M^*-estimate; k=3	99.87%
Hodges-Lehmann estimate (H.L.)	99.86%
Huber; c=1.4	99.60%
α-trimmed mean ($\bar{x}_\alpha$); α=0.1	99.54%
L_p-estimate; p=1.6	98.19%

Table II. Efficiencies of various statistical procedures if the errors are Jeffreys-distributed
II. táblázat. Statisztikai eljárások Jeffreys-eloszlásra vonatkozó hatásfokai

From the practical viewpoint, all six procedures turned out to be equally good if the samples come from the Jeffreys distribution. It should be emphasized that the first five estimates show efficiencies even greater than 99.5 %.

Introducing $t = (a-1)^{-1}$ as the type parameter, the assumption of the maximum likelihood-principle says nothing less than that the density function of the occurrence probabilities of various $f_a(x)$ types is

$$f_{ML}(t) = \delta(t-0.125) \tag{9}$$

(δ means Dirac-δ). For practical purposes, this is unacceptable. We can require at least that the occurrence probability density must be maximum for the type

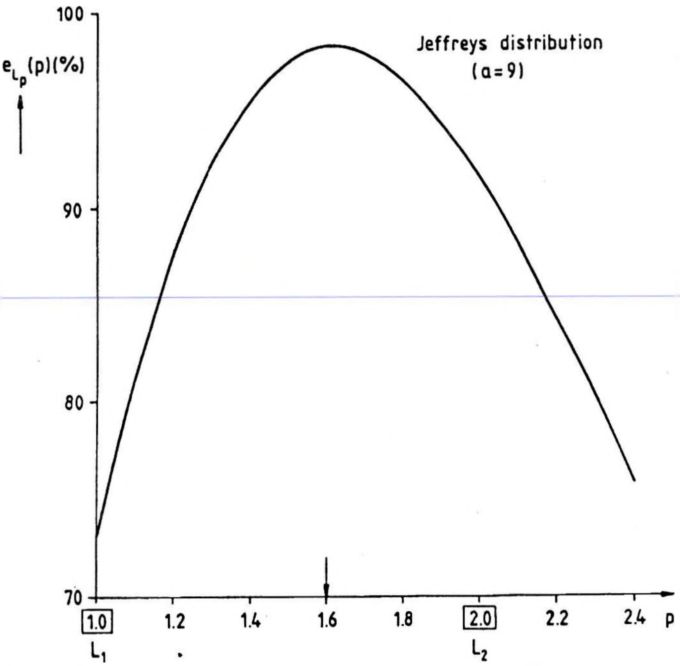


Fig. 3. Efficiency curve for different L_p-estimates for the Jeffreys distribution (see Eq. 7)
3. ábra. Különböző L_p-becslések hatásfokai a Jeffreys-eloszlásra vonatkozóan (ld. a 7 formulát)

t=0.125 (and not significantly less for the neighbouring types). If outliers seldom occur then one per cent probability density of the maximum value should be enough for the Cauchy-type to model somehow such situations, too; and finally we require $f(0)=0$ (see SZÜCS 1993 and references therein). Consequently, instead of Eq. 9 it is not only convenient but also justifiable to accept

$$f_J(t) = 64.t.e^{-8t.}, \tag{10}$$

the letter J in the index refers to the fact that $f_J(t)$ has its maximum position at t=0.125, i.e., at the Jeffreys distribution.

(A comment seems to be appropriate here: although $f(0)=0$ holds — in agreement with the modern statistical literature — the following zero hypothesis: 'the error distribution is Gaussian' is generally accepted at the commonly used significance levels even if Eq. 10 characterizes the occurrence probabilities of each type-interval, see SZÜCS [1993].)

The so-called Jeffreys interval of probability distribution types around t=0.125 shows the shortest flanks which can realistically be hoped for in nature. For example, in geostatistics, it can be stated [after DUTTER 1987] that we can accept as the most common type an $f_a(x)$ with a=5, i.e., with t=0.25. On the other hand, STEINER (ed.) [1991] shows examples proving that in the geosciences the Cauchy-type really occurs, i.e., the probability density of the types can not be a negligible value around t=1 compared with the maximum one.

These conditions are fulfilled (and $f(0) = 0$ also) if we accept as a probability density function for the distribution type t:

$$f_D(t) = 16.t.e^{-4t} \tag{11}$$

(compare Eq. 12 in STEINER 1991). Generally speaking, it is of crucial importance that we must at least be approximately imformed about the probability densities of the types of supermodel which can be accepted for adequate modelling of the error distributions occurring in a given discipline. It is the duty of the expert of the discipline in question to give an acceptable density function formula for the types which are able to model the actual error distributions in his territory of science or application. Both $f_D(t)$ and $f_J(t)$ curves are visualized in *Fig. 4.*

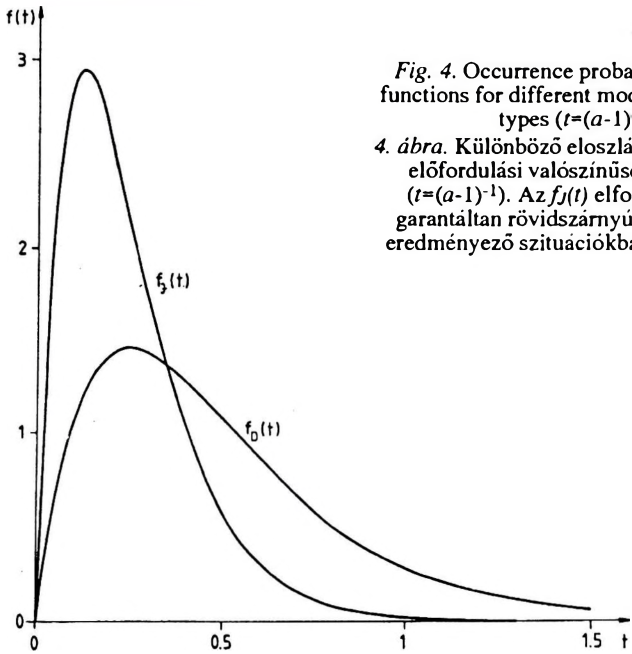


Fig. 4. Occurrence probability density functions for different model distribution types (t=(a-1)$^{-1}$)

4. ábra. Különböző eloszlástípusmodellek előfordulási valószínűségsűrűségei (t=(a-1)$^{-1}$). Az $f_J(t)$ elfogadása csak garantáltan rövidszárnyú eloszlásokat eredményező szituációkban javasolható

2.2. *Efficiency curves to visually demonstrate the different robustnesses of various statistical procedures*

One can find, in the literature of robust statistics, statements of the form: 'procedure A is robust, procedure B is not robust'. By the authors opinion such categorical distinctions are hard to justify — to say nothing about the contradiction that BOX [1953] introduced the notion 'robustness' for a method of conventional statistics (based on the L_2-norm) which letter is quite uniformly classified as 'not robust' by robust statistics (in the last three decades).

The efficiency curves versus t are shown in *Figs. 5–8* for all six statistical procedures figuring in Table II (in Figs. 7 and 8 the $e(t)$-curve for the median is also given). The speed of the decrease of e is different for increasing t from the nearly equal maximum value: it is most rapid for L_p p=1.6; at $t \geq 0.8333$ even e=0 holds. (It is easy to demonstrate also for the general case that $e > 0$ can hold only if $t < (2p-2)^{-1}$.) It is curious that two pairs of estimates behave similarly (M and M^* both for k=3; Huber c=1.4 and $\bar{x}_\alpha$ $\alpha = 0.1$; see Figs. 6 and 7) though the definitions of the corresponding statistical procedures are different.

Qualitatively the order concerning the robustness of the six procedures seems to be the following: L_p p=1.6; $\bar{x}_\alpha$ $\alpha = 0.1$ and Huber c=1.4; Hodges-Lehmann estimate; M and M^* both for k=3. The interesting behaviour of the latter $e(t)$-curves is that for $t \to \infty$ $(a \to 1)$ the efficiency seems to tend to an asymptotic value of 33-34% (see Fig. 8); Fig. 2 shows that these distributions have extremely heavy flanks. In *Figs. 9* and *10* also for k=2 the efficiency curves are shown both for M and M^*; the corresponding asymptotic values here are 48 and 50%, respectively. It should be mentioned that k=2 is accepted as the 'standard version' of the most frequent value (*MFV*-) calculations, in full agreement with the fact that maximum efficiencies are to be obtained very near to t=0.25 (i.e., to a=5) where $f_D(t)$ reaches its maximum (see Eq. 11).

The asymptotic behaviour of the $e(t)$ curves is a hint that *MFV*-procedures are not only robust to a high degree but are also extremely outlier-resistant. The two notions robustness and resistance, must be distinguished although there exists some interconnection between them. The oft occurring opinion, however, that robustness = outlier-resistance, is misleading and unacceptable.

2. 3. Average efficiencies as adequate indices of robustness in practice

Definition. Let us take the probability density function $\varphi(t; x)$ for t-values in the inverval $T_1 \leq t \leq T_2$ and let it be supposed that the probability density function of the type parameter t (i.e., $f(t)$) is also given. The index of the robustness of an estimation procedure according to $f(t)$ is defined as

$$r = \int_{T_1}^{T_2} e(t).f(t)\mathrm{d}t \tag{12}$$

where $e(t)$ is the efficiency of the estimation procedure in question if the data are distributed according to $\varphi(t; x)$.

Comment 1. The existence of $e(t)$ anticipates the existence of the Fisher-information of $\varphi(t; x)$ to the fixed value t, on the one hand and, on the other, it

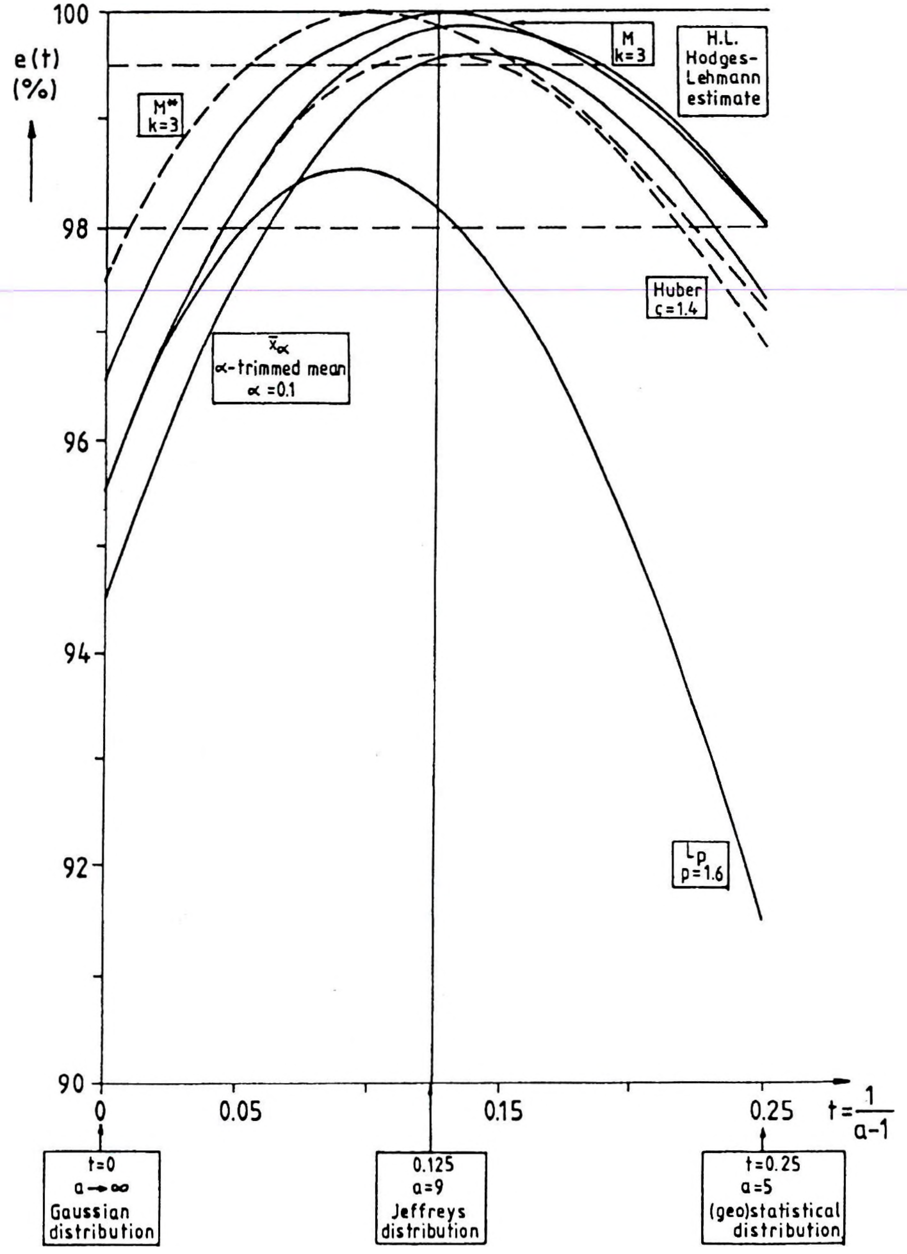


Fig. 5. Efficiency curves for six estimating procedures in the type interval $0 \leq t \leq 0.25$

5. *ábra.* Hatásfokgörbék hat becslési eljárásra a $0 \leq t \leq 25$ típustartományban

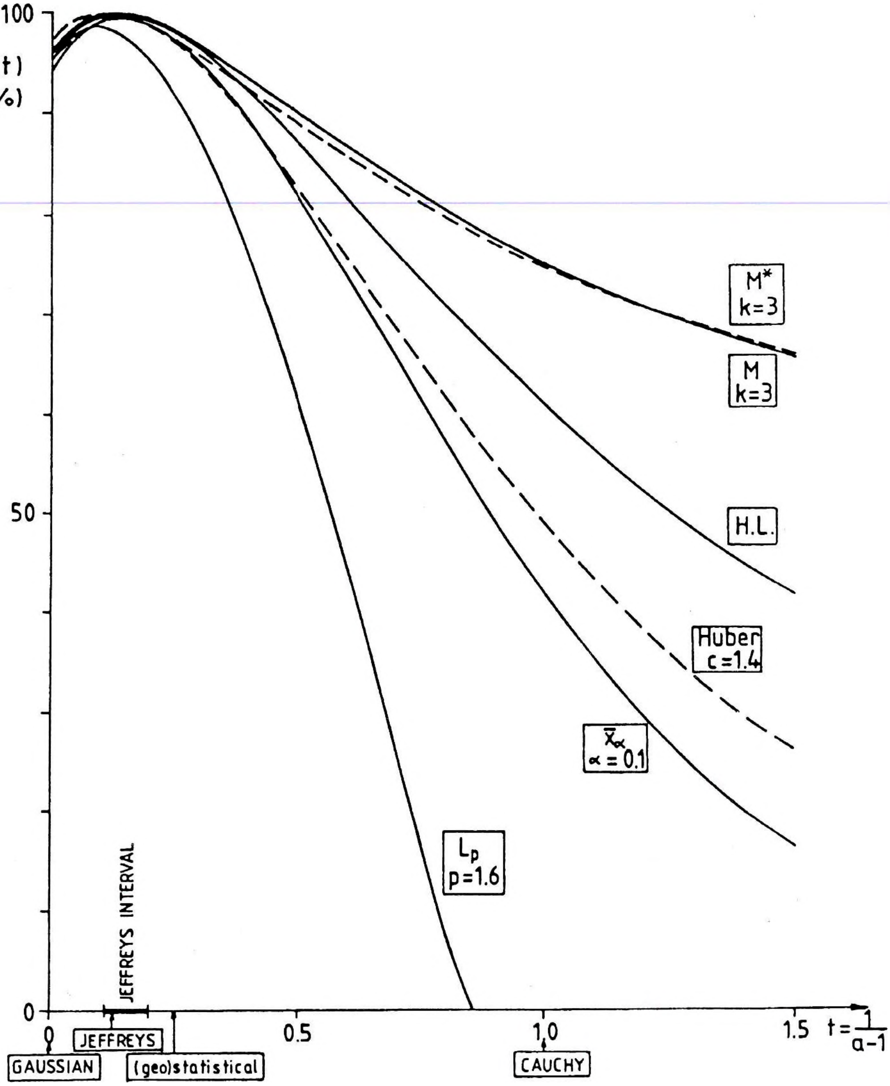


Fig. 6. Efficiency curves for six estimating procedures in the type interval $0 \leq t \leq 1.5$
6. ábra. Hatásfokgörbék hat becslési eljárásra a $0 \leq t \leq 1{,}5$ típustartományban

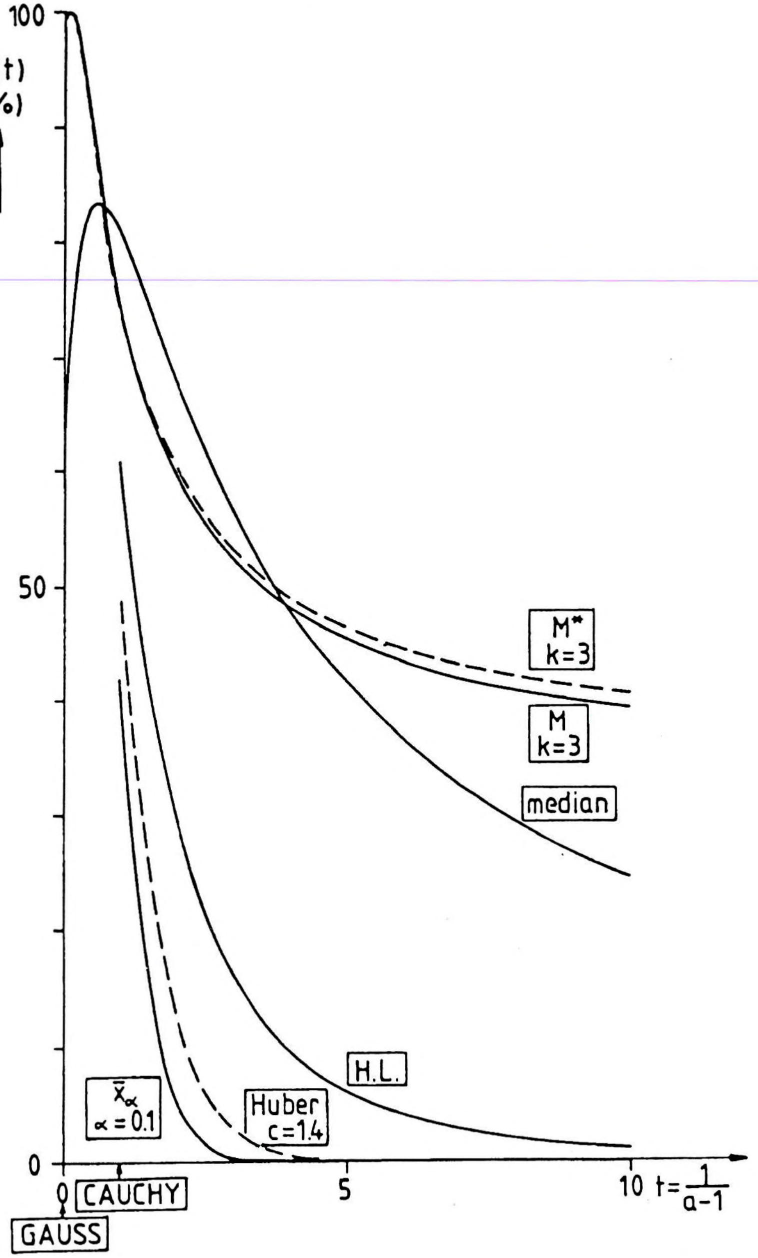


Fig. 7. Efficiency curves for six estimating procedures in the type interval $0 \leq t \leq 10$

7. ábra. Hatásfokgörbék hat becslési eljárásra a $0 \leq t \leq 10$ típustartományban

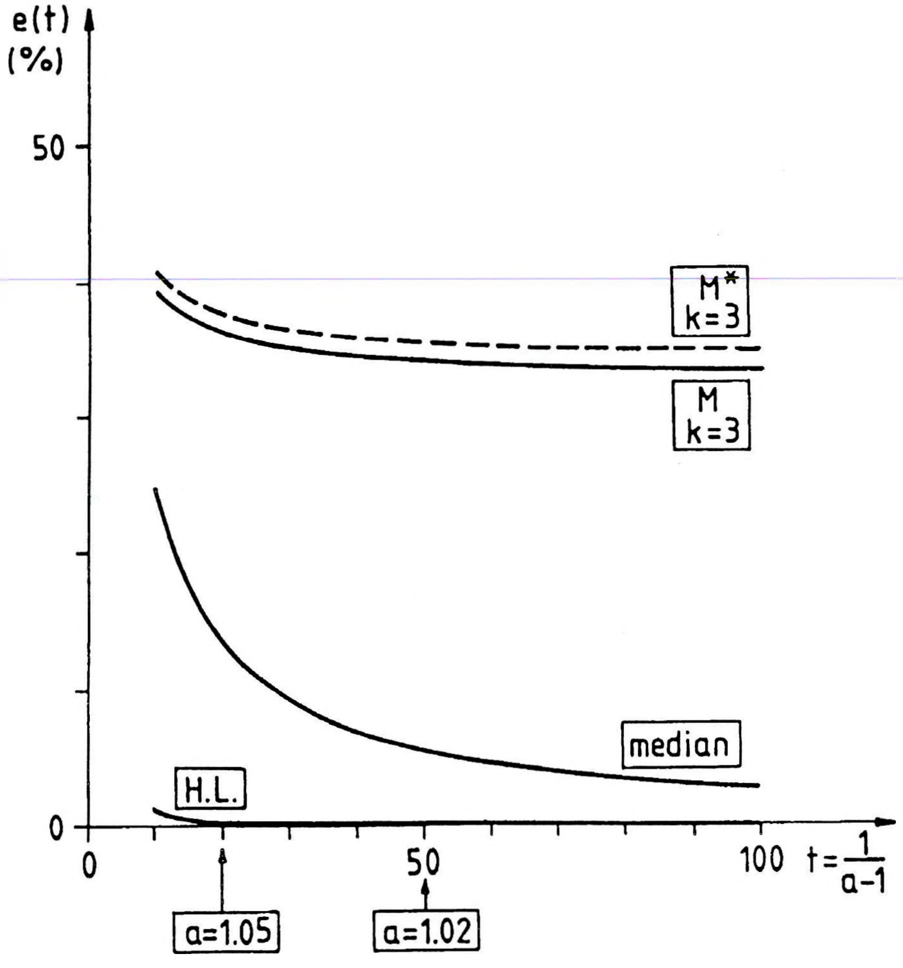


Fig. 8. Efficiency curves for four estimating procedures in the type interval $0 \leq t \leq 100$
8. ábra. Hatásfokgörbék négy becslési eljárásra a $0 \leq t \leq 100$ típustartományban

also anticipates the existence of the asymptotic variance of the estimates if the data are distributed according to $\varphi(t;x)$ ($T_1 < t < T_2$).

Comment 2. It is the task of the expert of a discipline (and not the task of the mathematician) to define a function $f(t)$ which can be accepted as an adequate one for the discipline in question. The choice $f(t)=f_D(t)$ (see Eq. 11) seems to be an adequate one in the geosciences (but the authors of the present paper suppose that this choice may be all right in other territories of statistics, too). The choice $f(t)=f_J(t)$ (see Eq. 10) seems to be a 'quasi-classical' one as

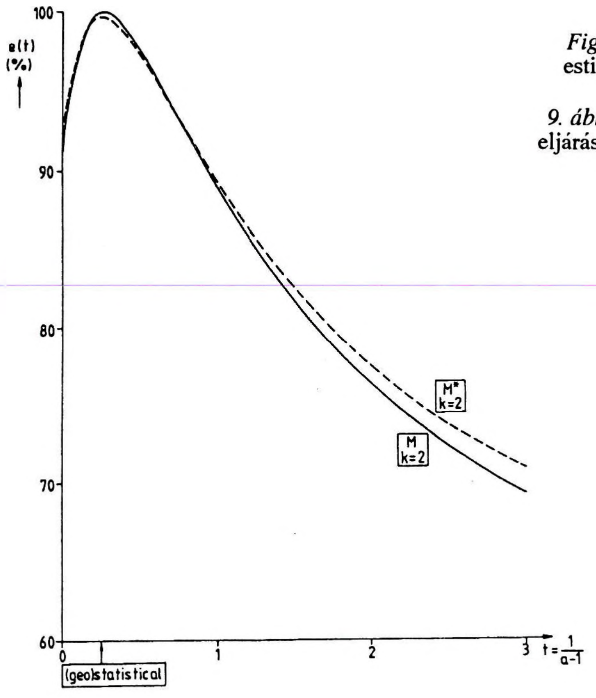


Fig. 9. Efficiency curves for two estimating procedures in the type interval $0 \leq t \leq 3$

9. ábra. Hatásfokgörbék két becslési eljárásra a $0 \leq t \leq 3$ típustartományban

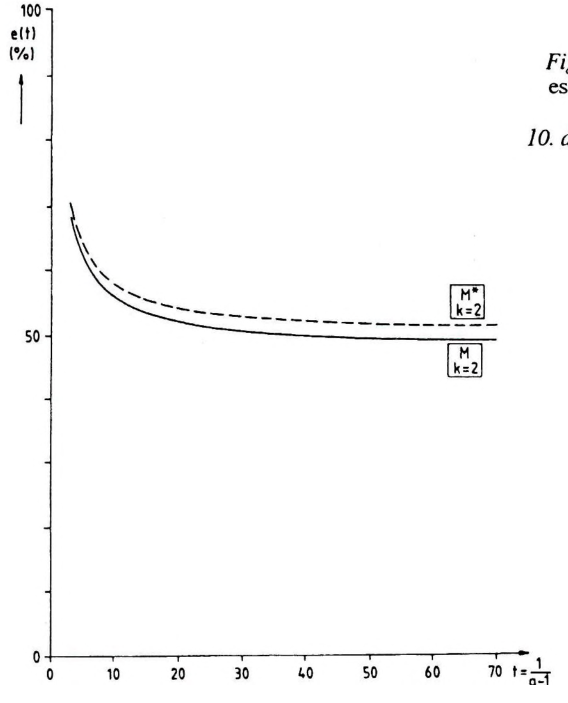


Fig. 10. Efficiency curves for two estimating procedures in the type interval $0 \leq t \leq 70$

10. ábra. Hatásfokgörbék két becslési eljárásra a $0 \leq t \leq 70$ típustartományban.

the tails of the distributions in the overwhelming majority of the cases are very short.

Comment 3. The definition of r given in Eq. 12 based on a supermodel $\varphi(t;x)$, i.e., for a case of only one type parameter, can be trivially generalized if more than one type parameter exist in the supermodel used.

In *Table III.* for ten statistical estimating procedures the indices of robustness are given (in per cent), calculated for both $f(t)=f_J(t)$ and $f(t)=f_D(t)$; the ordering was made according to the latter one.

statistical estimate		index of robustness (r) concerning the supermodel $f_a(x)$ if the occurrence probability of the various error distribution types are characterized by the density function	
name	symbol	$f_J(t)$ (Eq. 10)	$f_D(t)$ (Eq. 11)
arithmetic mean	$\bar{x}$ (L_p; $p = 2$)	67%	**36%**
	L_p; p=1.6	85%	**60%**
α-trimmed mean	$\bar{x}_\alpha$; $\alpha = 0.1$	93%	**79%**
sample median	med (L_p; p=1)	77%	**80%**
Huber-estimate (Proposal 2)	Huber; c=1.4	94%	**81%**
Hudges-Lehmann estimate	H. L.	96%	**85%**
most frequent value (*MFV*)	M^*; k=3	96%	**89%**
	M; k=3	97%	**90%**
	M^*; k=2	98%	**96%**
	M; k=2	98%	**96%**

Table III. Indices of robustness for various statistical estimates
III. táblázat. A robusztusság mérőszámai különböző statisztikai becsléseknél

REFERENCES

BOX G. E. P. 1953: Non normality and test on variances. Biometrika **40**, pp. 318–335

CSERNYÁK L., STEINER F. 1991: Monte Carlo studies on the rate of fulfillment of the law of large numbers. *In:* The Most Frequent Value. Introduction to a modern conception of statistics. (ed. F. STEINER) Chapter XII. pp. 173–185, Akadémiai Kiadó, Budapest

DONOHO D. L., LIU R. C. 1988: The 'automatic' robustness of minimum distance functionals. The Annals of Statistics **16**, No. 2 pp. 552–586

DUTTER R. 1987: Mathematische Methoden in der Montangeologie. Vorlesungsnotizen 1986/87. Leoben (Manuscript), 187 p
FEGYVERNEKI S. 1992: Robust estimates according to the model distribution and their applications (in Hungarian). Thesis. Manuscript. (University of Miskolc), 70 p
HAJAGOS B., STEINER F. 1991: The P^*-norm. Acta Geod. Geoph. Mont. Acad. Sci. Hung. **26**, (1-4) pp. 153-182
HAMPEL F. R., RONCHETTI E. M., ROUSSEEUW P. J., STAHEL W. A. 1986: Robust Statistics. Wiley, New York, 502 p
HODGES J. L., Jr. LEHMANN E. 1963: Estimates of location on rank tests. Ann. Math. Statist. **34**, pp. 518-611
HUBER P. J. 1964: Robust estimation of a location parameter. Ann. Math. Statist. **35**, pp. 73-101
HUBER P. J. 1981: Robust Statistics. Wiley, New York, 308 p.
SOMOGYI J., ZÁVOTI J. 1990: Die Anwendung der L_p-Norm-Schätzung für Ähnlichkeitstransformationen. Zeitschrift für Vermessungswesen **115**, (1) pp. 28-36
STEINER F. 1991: Average efficiency of statistical procedures. Acta Geod. Geoph. Mont. Acad. Sci. Hung. **26**, (1-4) pp. 135-151
STEINER F. (ed.); 1991: The Most Frequent Value. Akadémiai Kiadó, Budapest (Hungary), 315 p
SZŰCS P. 1993: Comment on an old dogma: 'The data are normally distributed' (present issue)
TARANTOLA A. 1987: Inverse Problem Theory. Elsevier, Amsterdam, 613 p

A ROBUSZTUSSÁG MÉRŐSZÁMÁNAK DEFINÍCIÓJA

STEINER Ferenc és HAJAGOS Béla

A dolgozat megadja a robusztusság r-rel jelölt mérőszámának a definícióját. A definíció szerint r a szóban forgó statisztikai eljárás hatásfokainak a súlyozott átlagaként számítandó; a súlyok valamely tudományág szemszögéből adekvátnak minősülő hibaeloszlástípusoknak az előfordulási valószínűségsűrűségei. A „robusztus" — „nem robusztus" kategórikus megítélés helyett, amely ma már túlhaladottnak tekintendő, a bemutatott példák az r=36 %-tól r=96 %-ig terjedő intervallumba eső robusztusság-értékeket mutatnak. A geofizika gyakorlatának különösen szüksége van ezen a téren is arra, hogy kvantitatív összehasonlításokat tehessen.

A dolgozat hat ábrája azokat az $e(t)$ hatásfokgörbéket is bemutatja, amelyek alapján az r számítása történik. Az olvasónak így módja van arra, hogy esetleges speciális szempontok szerint is vizsgálat tárgyává tegye a különböző statisztikai eljárások hatásfokainak a hibaeloszlástípus szerinti változásait.

GEOPHYSICAL TRANSACTIONS 1993
Vol. 38. No. 4. pp. 211-230

INVESTIGATIONS CONCERNING RESISTANCE — IMPORTANCE OF THE CHOICE OF THE FORMULA DETERMINING THE SCALE PARAMETER

Béla HAJAGOS* and Ferenc STEINER*

If statements are made only in a summary manner, '« distributionally robust » and « outlier resistant », although conceptually distinct, are practically synonymous notions' [HUBER 1982]. If, however, quantitative comparisons are necessary (especially in the practice of geophysics) on the grounds of an outlier model, an estimation procedure can turn out to be more resistant (compared with any other one) even though its index of robustness is significantly less. The estimation-pair of 'sample median' and 'α-trimmed mean' (α = 0.1) can serve as example.

The paper shows, too, that the chosen scale parameter generally plays a key role in the estimation of the location parameter regarding both the resistance and the robustness. For example, in the case of far lying outliers the estimate *MFV* (a variant of the *m*ost *f*requent *v*alue calculations) is to a significant degree more resistant than *CML* (frequently used abbreviation for *C*auchy *m*aximum *l*ikelihood, inasmuch as also the scale parameter is determined on the basis of the maximum likelihood principle), although the formula for determining the location parameters *MFV* and *CML* is just the same in both cases. It should be noticed, too, that the permissible rate of outliers (the classical breakdown bound) is also greater if *MFV*s are calculated and not *CML* values.

Keywords: resistance, scale parameter, statistical efficiency, outlier models, breakdown bound

1. The effect of one single Cauchy flank

Seldom can it be guaranteed that our data are outlier-free moreover there are countless types of outlier. Every investigation can consider only some of it.

One possibility is to accept the opinion of TARANTOLA [1987 p. 303]: 'the Cauchy function $1/(1+x^{2})$ '...'seems to be adequate for modeling suspected outliers by an unknown amount'. Bias, however, is generated by outliers only

* University of Miskolc, Department of Geophysics, H-3515 Miskolc-Egyetemváros
Manuscript received: 20 July, 1993

if the far lying values behave asymmetrically. Consequently, the simplest way is to investigate the estimating procedures on the outlier model given by the density funtion

$$f_{out}(x) = \begin{cases} 0, & \text{if } x \leq F_c^{-1}(p) \\ \dfrac{1}{\pi(1-p)\,(1+x^2)}, & \text{if } x > F_c^{-1}(p) \end{cases} \tag{1}$$

where $F_c\,(x)$ is the distribution function of the standard Cauchy distribution. The interpretation of Eq. 1 is the following:

in the interval $[F^{-1}(p), F^{-1}(1-p)]$ defined 'clear' distribution is distorted by a positive Cauchy tail of the weight $p/(1-p)$.

As the real value is assumed to be zero, the resulting T-values for given p-s have the meaning of bias caused by outliers greater than $F^{-1}(1-p)$. The T-curves for six estimating procedures are given in *Fig. 1*; besides the generally known α-trimmed mean ($\bar{x}_\alpha$ for $\alpha = 0.1$) and the sample median (med) four versions of the most frequent values are characterized by T-curves (M- and M^*-values for k=2 and k=3, see the corresponding ψ-functions and the condition for ε in Table I. in STEINER, HAJAGOS [1993]). The greater the increase of the bias (i.e., of the T-values) the less the resistance of the statistical procedure in question against such an occurrence of outliers. Fig. 1. shows that the sample median is more resistant than the α-trimmed mean in the conventional case of $\alpha = 0.1$, and the resistance of M^* for k=2 is even greater than the resistance of the sample median.

It is shown in STEINER, HAJAGOS [1993] that the α-trimmed mean for $\alpha = 0.1$ is more robust than the sample median — and we have just seen that the opposite relation is valid for the resistances if the outliers occur according to the 'Cauchy tail model'. The questions 'which is more robust?' 'which is more resistant?' must be answered in some concrete situations also giving the numerical values of the indices of the robustness and characterizing somehow quantitatively also the difference of the resistances (in different ways, e.g. by the quotient of two biases, i.e., of two T values for the same p which is actually of interest to us). Even if a given estimation A is more robust and more resistant than estimation B, quantitative comparisons can naturally differ significantly (e.g. A is twice as resistant as B but A is only a little bit more robust than B).

The foregoing shows that in respect of *quantitative comparisons* the notions 'robustness' and 'resistance' differ essentially from each other. If, however, only summarizing statements are made qualitatively (having only the possibility to say 'yes' or 'no'), we can agree that '« distributionally robust » and « outlier resistant », although conceptually distinct, are practically synonymous notions' [HUBER 1982]). Unfortunately, practical problems can seldom be solved satisfactorily with only 'yes' or 'no' answers; we are obliged to know which method is '*more* robust' and/or '*more* resistant' in concrete situations

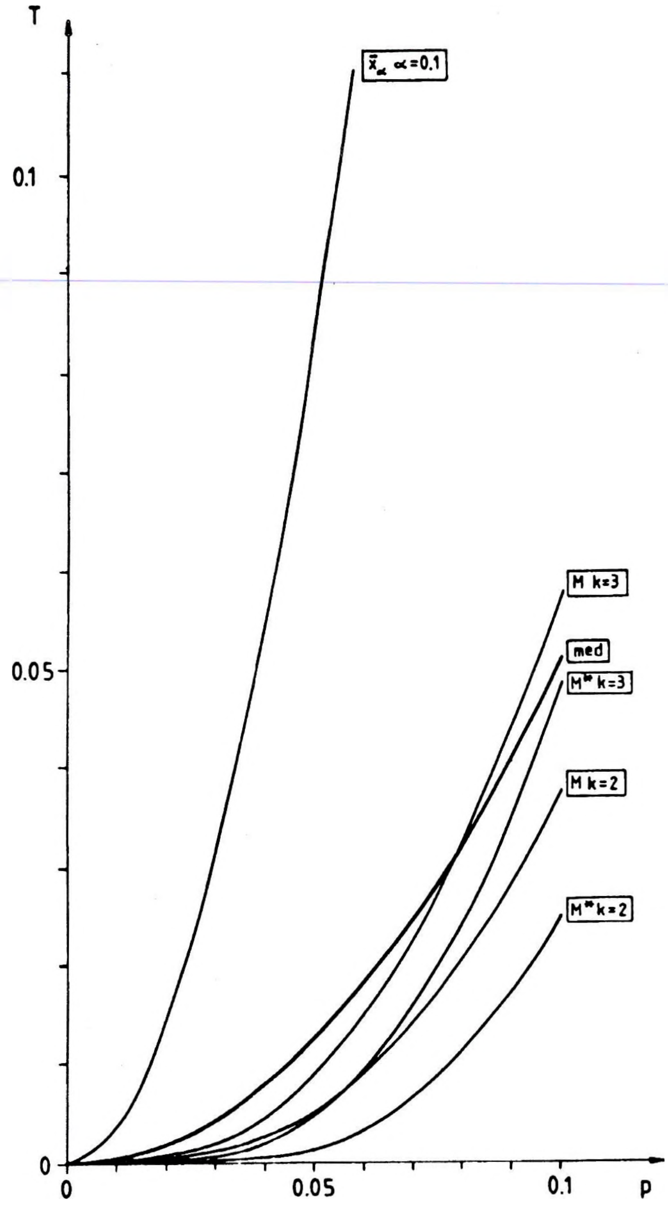


Fig. 1. Bias (T) versus 'weight' p of a single Cauchy-flank for α-trimmed mean, sample median, and some most frequent values (M and M^* for k=2 and 3)

1. ábra. Torzulások („bias", T-vel jelölve) egyetlen Cauchy-szárny p „súlyának" a függvényében α-levágott átlag, mintamedián és néhány leggyakoribb érték esetén (M és M^*; k=2 ill. 3) and — in addition — if these differences are significant or not. For instance engineering practice always needs measurable characteristics.

2. Comparison of *MFV*- and *CML*-estimates

Let *MFV* (*m*ost *f*requent *v*alue) in this paper be the notation for the *M*-value for $k=1$. The simplest form of its ψ-function is:

$$\psi_{MFV}(x) = \frac{x}{1+x^2}\ ; \tag{2a}$$

the so-called *CML*-estimate (*C*auchy *m*aximum *l*ikelihood), however, has just the same ψ-function:

$$\psi_{CML}(x) = \frac{x}{1+x^2} \tag{2b}$$

(see the iteration formula for μ in ANDREWS et al. 1972, p. 17). The curves characterizing the resistance (measured on a Cauchy tail) do not fully coincide (see *Fig.* 2.) because *CML*- and *MFV*-estimates differ from each other in the accepted scale parameter. (The med-curve in Fig. 2. is shown for comparison, and the *T*-curve for M^* $k=1$ is also given, showing a significantly greater resistance for this type of most frequent value calculations.)

The conditions for the just mentioned scale parameters denoted by ε_{MFV} and ε_{CML} — in integral form and reduced to zero — are the following:

$$\text{for}\quad \varepsilon_{MFV}\text{:}\ \int_{-\infty}^{\infty} \frac{3x^2-\varepsilon^2}{[\varepsilon^2+x^2]^2} f(x)\mathrm{d}x = 0\ , \tag{3}$$

$$\text{for}\quad \varepsilon_{CML}\text{:}\ \int_{-\infty}^{\infty} \frac{x^2-\varepsilon^2}{\varepsilon^2+x^2} f(x)\mathrm{d}x = 0 \tag{4}$$

(see HAJAGOS [1991], and ANDREWS et al. [1972] p. 17, respectively; in the latter case a convenient iteration formula is given on the second line of p. 17 resulting in ε_{CML} if $n\rightarrow\infty$). For the standard Cauchy type both formulae give unity (i.e., the semi-interquartile range of the standard Cauchy distribution) — but what about other distribution types?

To investigate this and similar questions we introduce the '$f_t(x)$-supermodel' by the probability densities

$$f_t(x) = \frac{\Gamma\left[\dfrac{1}{2t}+\dfrac{1}{2}\right]}{\sqrt{\pi}\ .\ \Gamma\left(\dfrac{1}{2t}\right)} (1+x^2)^{-\frac{1}{2t}-\frac{1}{2}} \tag{5}$$

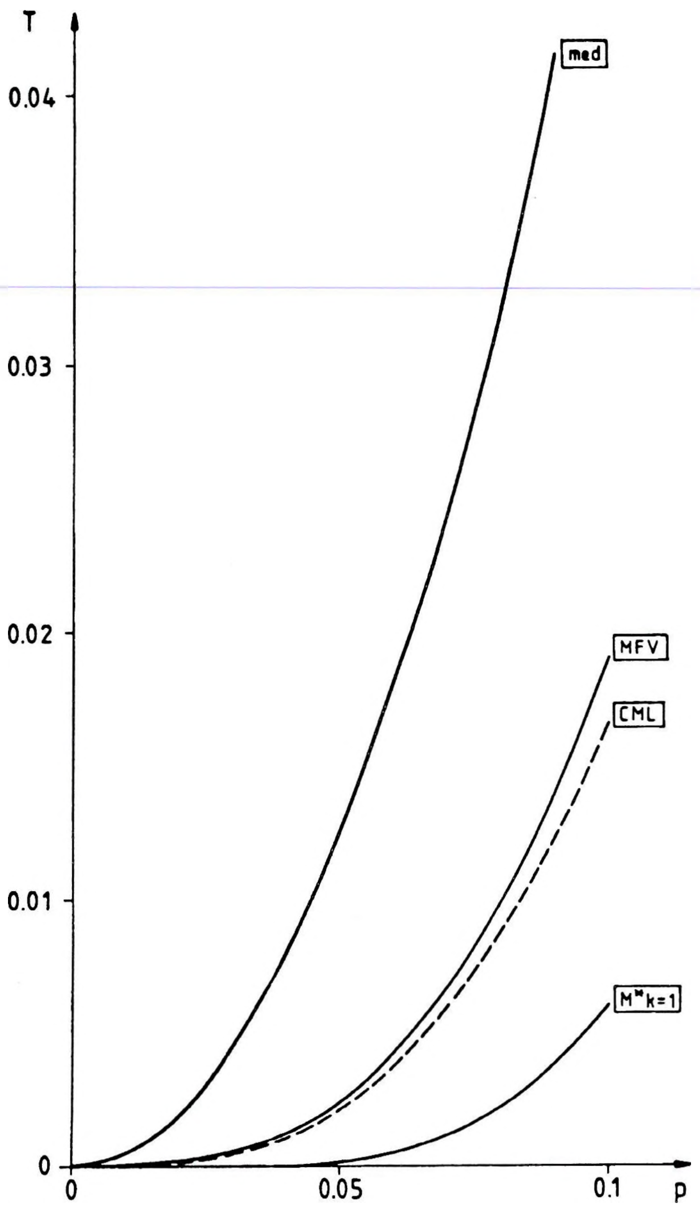


Fig. 2. Bias (T) versus 'weight' p of a single Cauchy-flank for *MFV*-, *CML*- and M^{*} $k=1$ estimations. For comparison, the curve for the sample median is demonstrated, too

2. *ábra.* Torzulások („bias", T-vel jelölve) egyetlen Cauchy-szárny p „súlyának" a függvényében *MFV*-, *CML*- és M^{*} $k=1$ becslésekre. Összehasonlítás kedvéért itt is feltüntettük a mintamedián görbéjét

With $a = 1+1/t$ this is the same expression as Eq. 5 in STEINER, HAJAGOS [1993] but in the present paper $f_t(x)$ means the actual density; in contrast, $f_a(x)$ will be the *model* density function being in some cases very far from the actual one. In such a way the treatment will be easy and will not lead to any misunderstanding.

For $t=1$ we obviously get the Cauchy density function $f_C(x)$, for $t=1/4$ the $f_{st}(x)$ and for $t=1/8$ the $f_j(x)$ densities (see Eqs. 6 and 7 in STEINER, HAJAGOS [1993]). If $t \rightarrow 0$ and the scale parameter simultaneously varies as $\sqrt{1/t-2}$, the limit density is the standard Gaussian one (given by $f_G(x) = (2\pi)^{-1/2} \cdot \exp(-x^2/2)$). All four specially mentioned types are visualized by their density functions in Fig. 1. of the just cited paper.

Fig. 3. shows the ε_{MFV}- and ε_{CML}- curves versus t, Eq. 3 and Eq. 4, respectively. The different scale parameters coincide only at $t=1$; for great values of t ε_{MFV} tends to a constant value (to 2.592), whereas, ε_{CML} increases exponentially, see *Fig. 4.* (It should be noted that if $f(x)$ is not symmetrical to the origin, naturally $(x-MFV)^2$ and $(x-CML)^2$ figure instead of x^2 in Eqs. 3 and 4.)

From the viewpoint of determining of the location parameter the definition of the scale parameter is usually treated as a second order question, or even one that can be neglected. The question arises if this method of treatment is justified or not with respect to both the resistance and the robustness as the ε-curves are quite different. The simplest way is to show the efficiency curves $(e(t))$ for both estimations (see *Figs. 5* and *6*). The significance of the differences is obvious

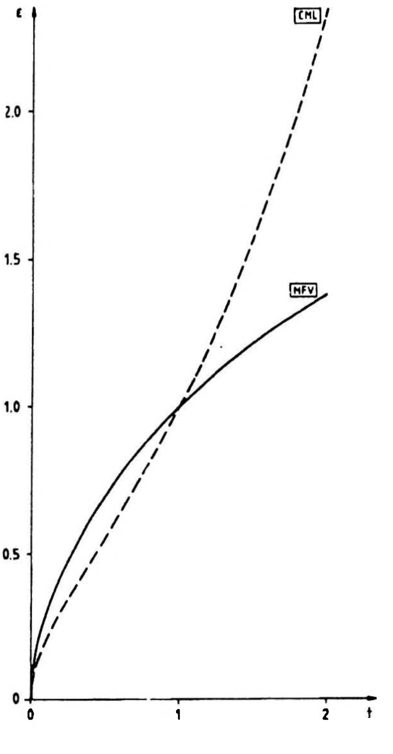


Fig. 3. Scale parameters ε_{MFV} and ε_{CML} versus type parameter t (see Eq. 5). The values equal each other only at $t=1$, i.e., at the Cauchy distribution where $\varepsilon_{MFV} = \varepsilon_{CML} = 1$ holds

3. ábra. A kétféle skálaparaméter: ε_{MFV} és ε_{CML} értékei a t típusparaméter függvényében (ld. az (5) formulát). Az értékek kizárólag $t=1$-nél, azaz a Cauchy-eloszlásnál egyeznek meg, ahol is $\varepsilon_{MFV} = \varepsilon_{CML} = 1$

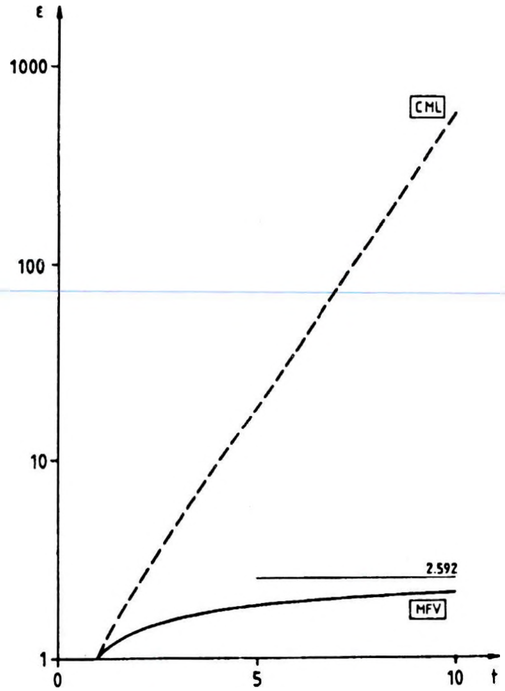


Fig. 4. Scale parameters ε_{MFV} and ε_{CML} versus type parameter t (see Eq. 5). At large value of t, i.e., in cases of heavy flanks, ε_{CML} increases exponentially as a function of t, whereas, ε_{MFV} tends to a finite value

4. ábra. A kétféle skálaparaméter: ε_{MFV} és ε_{CML} értékei a t típusparaméter függvényében (ld. az (5) formulát). A t típusparaméter nagy értékeinél, azaz súlyos szárnyak esetén, ε_{CML} exponenciálisan növekszik t-vel, míg ε_{MFV} véges értékhez tart

(e.g. for $t=0$, i.e., for the Gaussian distribution type, *CML* has 60 % efficiency and *MFV* has an efficiency of 74 %). The indices of robustness are 94 % for *MFV* and 87 % for *CML* (calculating according to $f_D(t)$, see STEINER, HAJAGOS [1993], based on the type-distributions characterized by $f_J(t)$, the results are 89 % for *MFV* and 79 % for the *CML*-estimation). The latter value differs from 100 % about twice as much as the index of robustness for *MFV*. These values and Figs. 5 and 6 clearly show that *MFV*- and *CML*-estimates differ from each other significantly — at least in respect of robustness. Paragraph 3.4. of the present article shows, however, that the same is valid concerning the resistance if there are very many far lying outliers. (Even Fig. 6. itself shows that the *MFV*-method has a much greater resistance compared to the *CML* calculations if we interpret the heavy flanks belonging to great t values as a symmetrical appearance of the outliers causing no bias but a considerable decrease in accuracy; see also STEINER [1991]. In the present article, however, the effects of outliers are treated in the overwhelming majority of cases in respect of the bias.)

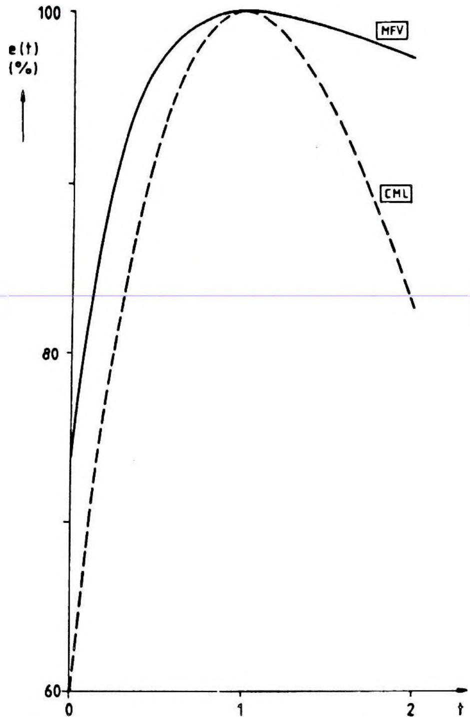


Fig. 5. Efficiency curves of *MFV*- and *CML*-estimates versus type parameter *t*. The disadvantages of the *CML*-estimate are obvious. (The calculation method is given, e.g., in STEINER, HAJAGOS [1993].)

5. ábra. Az *MFV*- és *CML*-becslések hatásfokgörbéi a *t* típusparaméter függvényében. A *CML*-becslés hátrányai nyilvánvalóak. (A számítás módjára nézve ld. pl. STEINER, HAJAGOS [1993])

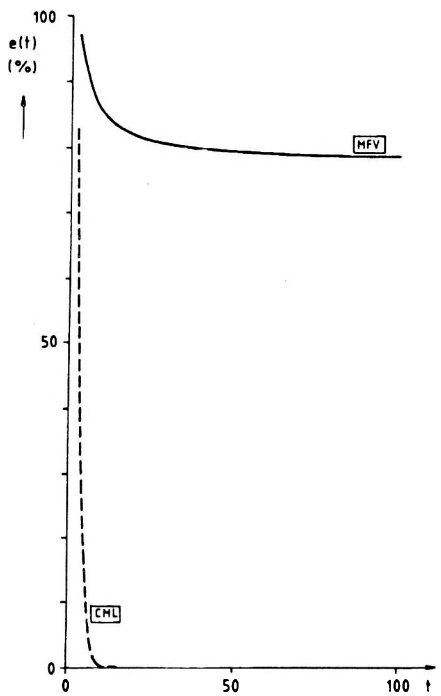


Fig. 6. Efficiency curves of *MFV*- and *CML*-estimates versus type parameter *t*. For very heavy flanks the efficiency of *CML* decreases to zero, that of the *MFV*, however, remains at a high efficiency level

6. ábra. Az *MFV*- és *CML*-becslések hatásfokgörbéi a *t* típusparaméter függvényében. Extrém mértékben súlyos szárnyak esetén a *CML* hatásfoka zérusra csökken, míg az *MFV*-é tekintélyes értékű marad

3. Three other types of outliers

3.1. Some heuristic remarks

The behaviour of the ε_{MFV}-curve in Fig. 4. is a consequence of the basic 'philosophy' of the most frequent value (*MFV*-) calculations which are also heuristically presented in some parts of STEINER [1991] and can be summarized shortly as follows: to a significant per cent of data there must be as small residuals as possible, even if some other residuals turn out to be very large. A 'built-in' resistance against outliers is therefore already ensured in the 'philosophy' of the most frequent value calculations. We can perhaps justifiably speak about a different conception of statistics as L_1-norm minimization techniques (in the simplest case: calculation of the medians) always take *all data* into account — and statistical procedures based on the L_2-norm are even more sensitive to great values of the residuals (resulting in extreme outlier-sensitivity) and therefore it is not guaranteed that, for example, the arithmetic mean characterizes the densest lying group of the data. Another question is in which sense we can justifiably speak about a 'new' or 'modern' conception of statistics as the basic idea can be found in SHORT [1763] — to say nothing about a lot of only heuristically based reweighting procedures figuring in applied disciplines in the last decades. Now, *MFV*-procedures are theoretically based on the minimization of the I-divergence (see the previously cited HAJAGOS [1991]) and their characteristic features have been investigated in detail (see the bibliography of STEINER (ed.) [1991] and the Table which is the supplement of this book).

3.2. Versions of the most frequent value calculations

The above cited Table in STEINER (ed.) [1991] shows that in the standard version of the most frequent value calculations the scale parameter $S = 2\varepsilon_{MFV}$ is used; if short flanks are guaranteed then $S = 3\varepsilon_{MFV}$ is recommended (see also STEINER, HAJAGOS [1993]).

HAJAGOS [1985] has shown that if generalized Student distributions are used as substituting distributions (the formula for the probability densities is given as Eq. 5 in STEINER, HAJAGOS [1993] applying a as the type parameter) then

$$\int_{-\infty}^{\infty} \frac{(a+1)x^2-\varepsilon_a^2}{[\varepsilon_a^2+x^2]^2} f(x)\mathrm{d}x = 0 \tag{6}$$

must hold to be sure that the minimum of the I-divergence is actually reached; ε_a represents the scale parameter if a is chosen in Eq. 6 as the value of the type parameter. For an $f(x)$ which is not symmetrical to the origin then obviously $(x-M)^2$ must figure in Eq. 6 instead of x^2 (or even $(x-M_a)^2$ can be written enhancing that we use Eq. 6; in the usual way we get M_k as the location parameter). From the point of view of the theory it seems more consistent to use ε_a fulfilling Eq. 6 with a=5 or a=9 instead of calculating with a=2 in the first step, i.e., to determine ε_{MFV} according to Eq. 3, and in the second step to multiply by k=2 or 3.

Before investigating the outlier-resistance of Eq. 6, we show the similarity (and also the differences) between the two possibilities of the most frequent value calculations. For purposes of comparison we need to calculate the quotients

$$\frac{\varepsilon_5}{2\varepsilon_{MFV}} \quad \text{and} \quad \frac{\varepsilon_9}{3\varepsilon_{MFV}}$$

as a function of t, i.e., as a function of the distribution type, see Eq. 5.

	t	$\frac{\varepsilon_5}{2\varepsilon_{MFV}}$	$\frac{\varepsilon_9}{3\varepsilon_{MFV}}$
Gaussian	0	0.9698	0.9429
	0.0156	0.9737	0.9490
	0.0312	0.9777	0.9553
	0.0625	0.9858	0.9683
	0.0125	1.0026	0.9960
Statistical	0.25	1.0378	1.0568
	0.5	1.1102	1.1936
Cauchy-type	1	1.2500	1.5000
	2	1.4883	2.1649

Table I. Comparison of two calculation methods for determining the scale parameter
I. táblázat. Kétféle skálaparaméter meghatározási módszer összehasonlítása

The results are demonstrated in *Table I.*: in broad type intervals are the values of these quotients near to unity. To check that in fact for the most frequent values similar behaviour is valid, *Fig. 7* gives the curves of the relative efficiencies e_a/e_k of the most frequent value variants M_a and M_k for the parameter pairs a=5, k=2, and a=9, k=3. In the first case, throughout the whole type interval from the Gaussian to the Cauchy-type we find greater values than 90%; in the second case the statement is only valid for 2/3 of this type interval

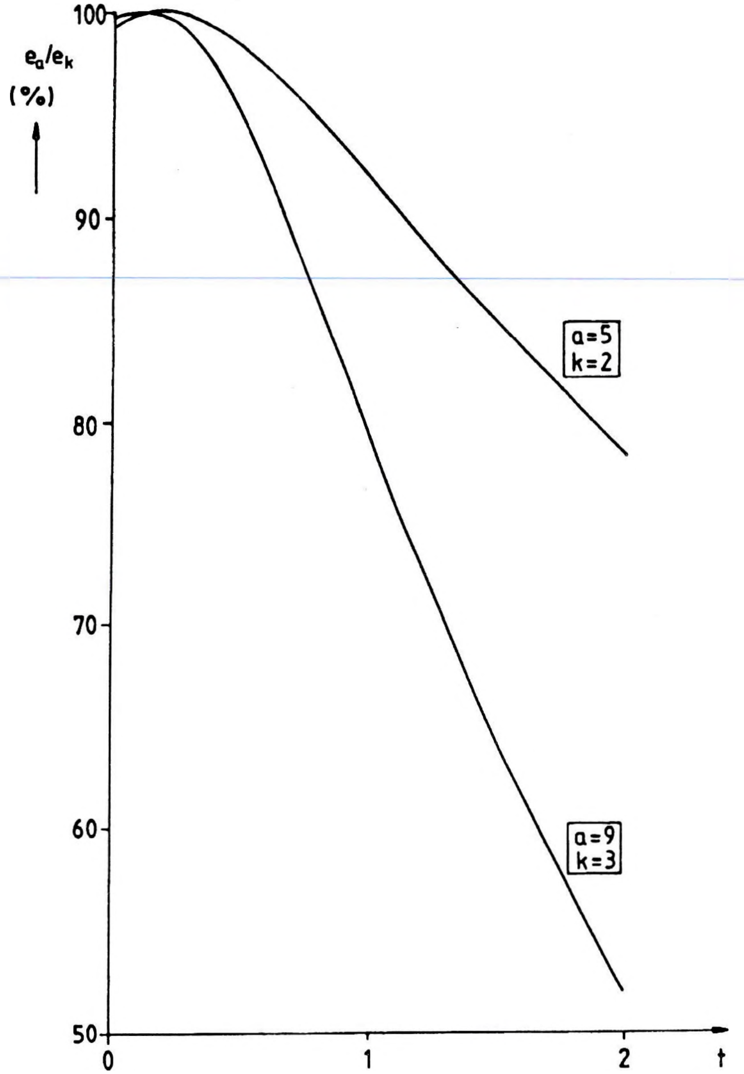


Fig. 7. Curves of relative efficiencies versus type parameter t of two pairs of location estimates: M_a $a=5$ is compared to M_k $k=2$ and M_a $a=9$ is compared to M_k $k=3$. (In calculation of M_a ε_a figures as scale parameter according to Eq. 6 instead of using $k.\ \varepsilon_{MFV}$ as for the M_k-s.) In general, the usually proposed version of the most frequent value calculations (i.e. M_k) is more advantageous in the most important type interval $0 < t < 0.5$, however, both versions behave very similarly ($e_a/e_k > 95$ %)

7. ábra. Két helyparaméter becslés-pár relatív hatásfok-görbéi a t típusparaméter függvényében: M_a a=5-öt M_k k=2-höz, míg M_a a=9-et M_k k=3-hoz hasonlítjuk. (Az M_a-k számításakor a (6) egyenlet szerinti ε_a a skálaparaméter, míg az M_k-kat $k.\varepsilon_{MFV}$-vel számítjuk.) Megállapíthatjuk, hogy a szokásosan javasolt verzió (azaz az M_k számítása) az előnyösebb, a legfontosabbnak ítélhető $0 < t < 0{,}5$ típusintervallumban azonban nagyon hasonlóan viselkedik a két verzió (a fenti tartományban a két eljárás relatív hatásfoka nagyobb 95%-nál) but this latter is also a very broad one. This means that both variants of the most frequent value calculations do not differ significantly from each other in type intervals of considerable lengths and therefore investigation results obtained for the second variant are also informative for the commonly used one.

3.3. Concentrated and dispersed outliers

Let us suppose that we calculate ε_a as the scale parameter according to Eq. 6 but there are also outliers. Two cases were investigated earlier in detail so only the results are reproduced here.

In the first case not only the outlier-free data but also the outliers occur around a fixed value. The distance between the mentioned point-groups, however, is relatively large compared with the dispersion of the values inside a group, thus modelling with two Dirac-δ-s is adequate (see *Fig. 8.*); the relative number of outliers is denoted by C for this outlier model. HAJAGOS [1988] got

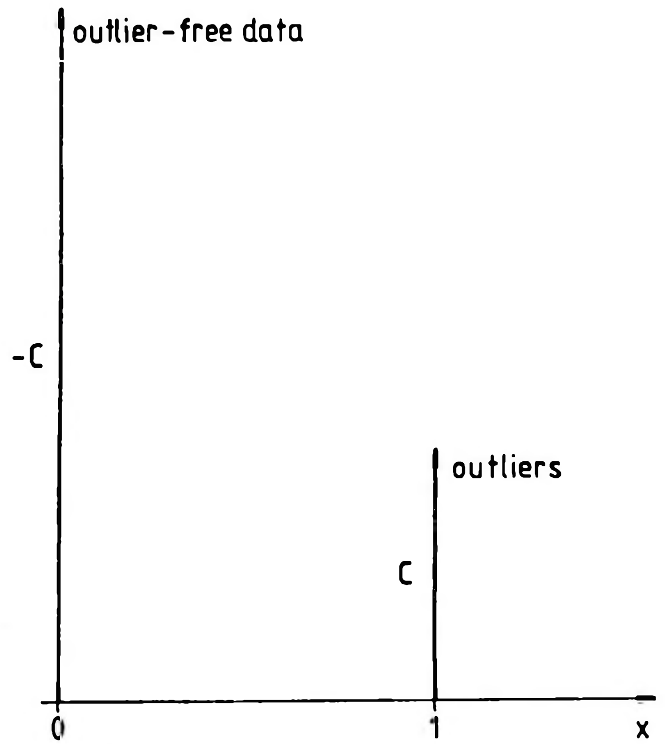


Fig. 8. Outlier-model if outliers occur very far but are relatively concentrated
8. ábra. Durvahiba-modell arra az esetre, ha a durvahibák nagyon távol, de viszonylag koncentráltan jelentkeznek

the following closed formula for the maximum tolerable relative amount (C_{max}) of such outliers as a function of a (in per cent)

$$C_{max} = 50\,[1-\{[(a^3+16a^2+63a+54)^2+4a^3(a+10)]^{1/2} - (a^3+16a^2+63a+54)\}^{1/2}/[2(a+10)]^{1/2}]\ \% \qquad (7)$$

Fig. 9 shows that this value decreases to zero if $a \to \infty$ and tends to the value of 45.68 % if $a \to 1$.

In the second case an outlier can occur anywhere but without a concentration point, therefore the 'distribution' of the outlier-free data can also be modelled here by a Dirac-δ. The permissible rate of such outliers (denoted by OUT_{max}) can be calculated in accordance with

$$OUT_{max} = 1 - \frac{(a+1)^2}{(a+1)^2+4(a+2)} \qquad (8)$$

[STEINER 1988a]. This curve is also shown in Fig. 9. The value of OUT_{max} obviously approximates 3/4 if $a \to 1$, it is more interesting, however, that OUT_{max} = 64 % holds even if a=2 is chosen in Eq. 6, i.e., if we calculate ε_{MFV}

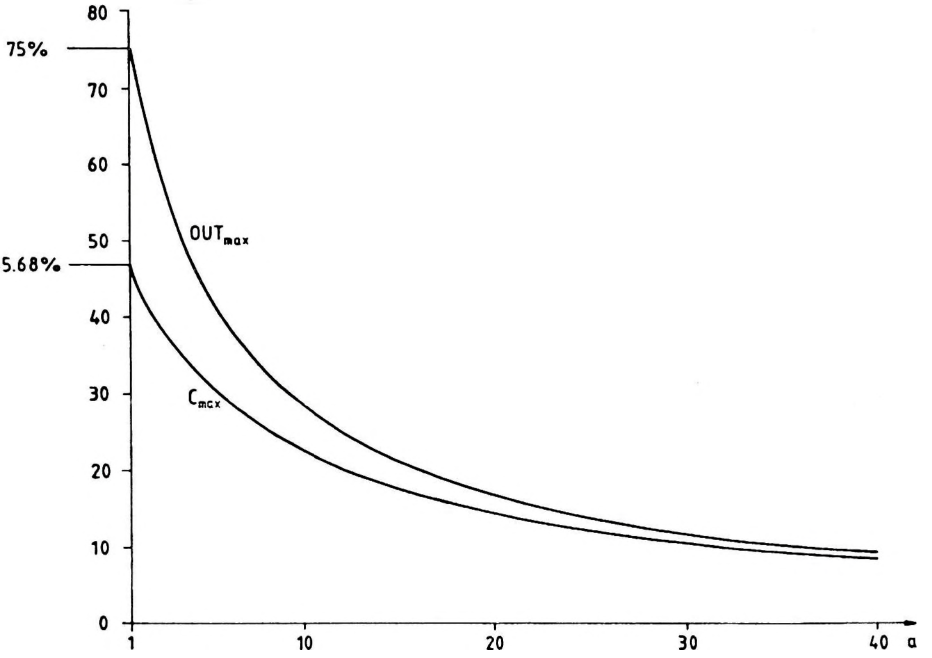


Fig. 9. C_{max} values versus a according to Eq. 7 (see also Fig. 8.); OUT_{max} curve (see Eq. 8) if outliers can occur anywhere but without any concentration point

9. ábra. A (7) egyenlet szerinti C_{max}-értékek az a függvényében (ld. a 8. ábrát), valamint a (8) egyenlet szerinti OUT_{max}-görbe arra az esetre, ha a durvahibák bárhol előfordulhatnak, de nem mutatnak koncentrálódási tendenciát according to Eq. 3 [see also CSERNYÁK, STEINER 1991, the first rows on page 92].

3.4. Classical breakdown-bound investigations

With regard to the beginnings of the systematic investigations made for robustness, resistance and breakdown bounds since 1964 one should mention the 'classical' investigations performed in Princeton and reported in ANDREWS et al. [1972].To obtain practical breakdown bounds the following samples of n elements were used: $(n-n_{out})$ data were randomly chosen from a standard Gaussian distribution, the others were 100; 200; 300; ...; $100n_{out}$. The corresponding probability density function $f_{G;out}(x)$ can be written as follows

$$f_{G;out}(x) = \frac{1-n_{out}/n}{\sqrt{2\pi}} \exp(-x^2/2) + \frac{1}{n}\sum_{i=1}^{n_{out}} \delta(x-100i) \tag{9}$$

(here, δ also represents the Dirac-δ). Standard Gaussian data occur practically only (with a probability of 0.997) in the interval (-3, +3) therefore the estimate is not accepted if it is outside this. In ANDREWS et al. [1972] the maximum n_{out}/n value was accepted as a breakdown bound for which the estimate (obtained by the investigated statistical procedure) was still less than 3.

It is useful to visualize the 'density function' $f_{G;out}(x)$ but we are forced, because of the limited graphic possibilities, to indicate also the Gaussian part of the expression in Eq. 9 with a single Dirac-δ in the origin, i.e., with the whole occurrence probability $(1-n_{out}/n)$ of the outlier-free values (see *Fig. 10* for n=100 and for n_{out} = 30). According to the heuristics given in 3.1 it is to be expected that the 'philosophy' of the most frequent value calculations results in the tolerance of a considerable rate of this sort of outlier. In fact, calculating with $f(x) = f_{G;out}(x)$ (i.e., substituting the expression of Eq. 9 as $f(x)$ into the integrand of Eq. 3 and writing simultaneously $(x-M)^2$ or $(x-MFV)^2$ instead of x^2), the curves of the standard most frequent values (M; k=2) and the MFV-s (i.e., M values for k=1) as location parameters (T-values) show equally negligible bias to a given maximum n_{out}/n value (see *Fig. 11*; the unbiased value is clearly zero in the case of Eq. 9 therefore the T-values simultanously have the meaning of bias, too). In the standard case 41% is the maximum n_{out}/n ratio, for MFV calculation it may even be 57%. (The breakdown bounds are the following for other most frequent values which are not shown in Fig. 11: 32 % for M, k=3; 31 % for M^*, k=3; 40 % for M^*, k=2 and 59 % for M^*, k=1.)

We have seen in point 2 that CML is the maximum likelihood counterpart of MFV having just the same ψ-function (see Eqs. 2a and 2b) but the scale parameter (ε_{CML}) is defined by Eq. 4. The resulting CML-curve is quite different from the MFV-curve: it seems to be 'continuous' and the breakdown bound turns out to be 50 %.

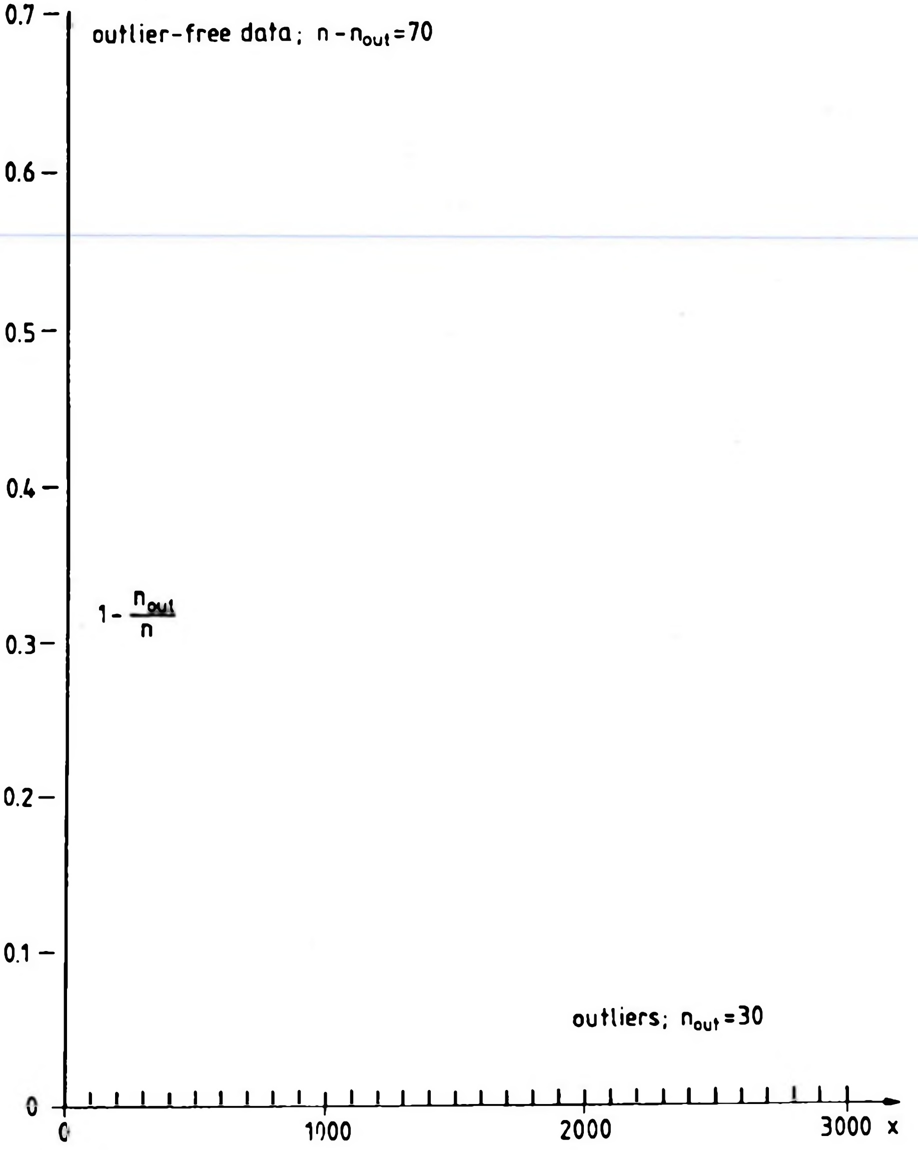


Fig. 10. Probabilites according to the breakdown bound investigations made in ANDREWS et al. [1972] (see also Eq. 9)

10. ábra. Az ANDREWS et al. [1972] szerinti „breakdown bound"-vizsgálatok valószínűségi modellje, mint Dirac-δ-k az x számegyenesen; ld. még a (9) kifejezést

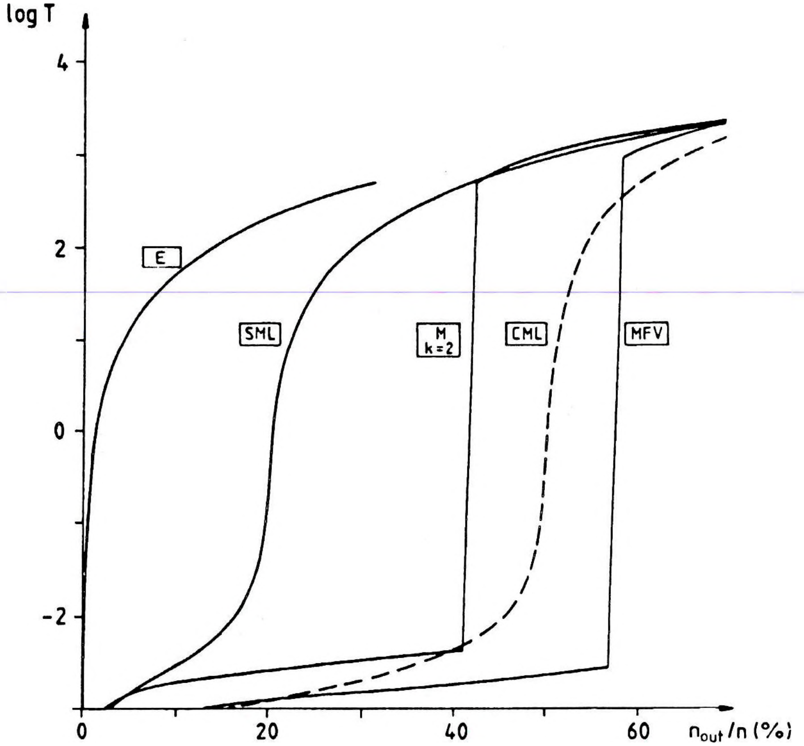


Fig. 11. Bias (T) versus n_{out}/n (see Fig. 10.) for different estimates of location. Most frequent value curves (MFV and M, k=2) show, before the jump, less bias than one per cent of the scatter characterizing the outlier-free distribution. The curve of the expected values is denoted by E. (For the SML-curve see caption of Fig. 12.)

11. ábra. Torzulások („bias", T-vel jelölve) a 10. ábra szerinti n_{out}/n függvényében, különböző helyparaméter-becslésekre. A leggyakoribb érték-görbék (MFV és M, k=2) az ugrás előtt 1%-nál kisebb torzulást mutatnak (a durvahiba-mentes eloszlás szórásához viszonyítva). A várható érték görbéjét E-vel jelöltük. (Az SML-görbével kapcsolatban ld. a 12. ábra feliratát.)

It is appropriate to show another pair of estimations, too, beyond the already known pair *CML* and *MFV*, both having for just the same distribution type optimum behaviour. However, some remarks should first be made on the importance of scale parameter determination.

The assumption of the 'a priori known type' can result in outlier-sensitive values of the scale parameter, (see the *CML*-curve in *Fig. 12* or the example given in STEINER [1988b]). The parameter of scale belonging to the distribution defined by Eq. 9 with a given (non-zero) value of n_{out} and with n=100 is denoted by ε in this figure for all investigated methods; ε_{th} is the 'theoretical value' in the sense that no outliers exist, i.e., if n_{out} = 0 holds. These latter values (characterizing the standard Gaussian distribution itself) are the following: ε_{th} = 0.6120 for *CML* and ε_{th} = 0.9254 for *MFV*. The ε_{CML}-values are seriously distorted even before the breakdown (i.e., if n_{out}/n < 50 %); on the otherhand

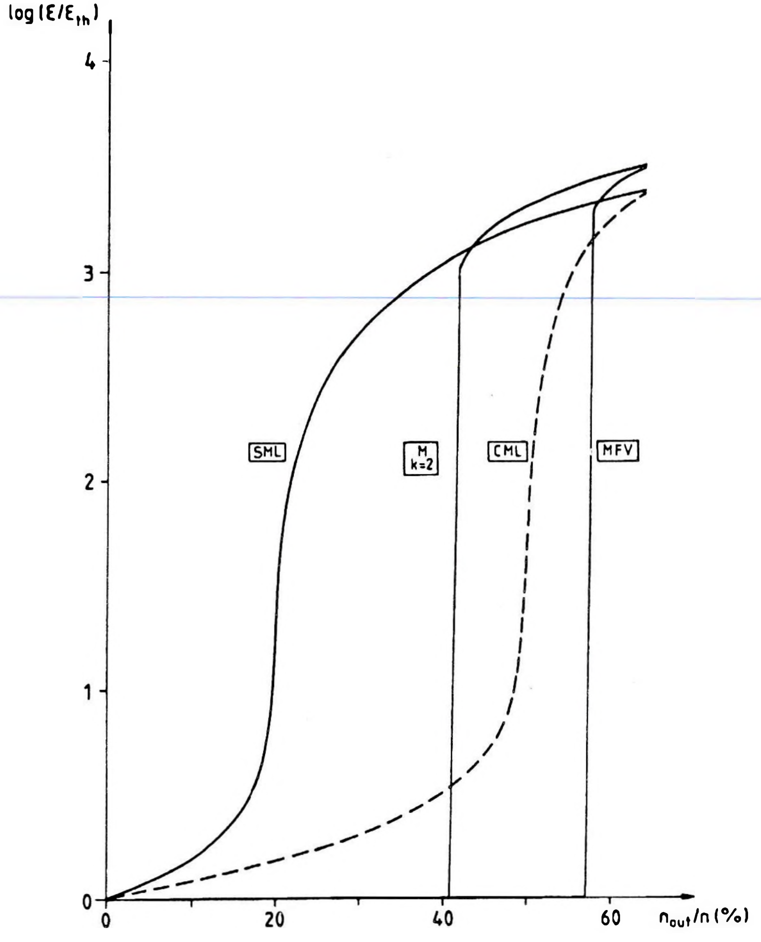


Fig. 12. Actual values compared to the theoretical ones for different definitions of the scale parameter, depending on n_{out}/n (see Fig. 10). *SML:* see Eq. 11 for ε_{SML}; *CML:* see Eq. 4 for ε_{CML}; *MFV:* see Eq. 3 for ε_{MFV} (and for M k=2 the well known $2\varepsilon_{MFV}$ is used)

12. ábra. Aktuális értékek (ε) az elméletihez (ε_{th}) viszonyítva a skálaparaméter különböző definíciói esetén, az n_{out}/n függvényében (ld. a 10. ábrát). *SML:* ε_{SML} a (11) egyenlet szerint; *CML:* ε_{CML} a (4) egyenlet szerint; *MFV:* ε_{MFV} a (3) egyenlet szerint (*M* k=2 esetén jól ismerten a $2\varepsilon_{MFV}$ skálaparamétert használjuk)

ε_{MFV}-values are practically not influenced by the outliers if $n_{out} \leq 57$ %. Comparing these log ($\varepsilon/\varepsilon_{th}$) curves with the *CML*- and *MFV*-curves in *Fig. 11* (where the logarithms of the bias are demonstrated), we can conclude that the breakdown behaviour of the location parameter estimates strongly depends on the estimation used for the scale parameter.

We now show the other pair of estimations which work optimally at the same distribution type. Both Figs. 11 and 12 show curves (marked with *M, k*=2) for the standard variant of the most frequent value calculations; we have only to define the 'maximum likelihood' counterpart of this; the only difference being in the definition of the scale parameter.

The question can be posed even more generally: Which is the counterpart of Eq. 6 if not only the I-divergence is minimized but also the scale parameter is determined according to the maximum likelihood principle? In the usual way (on the basis of Eq. 5 of STEINER, HAJAGOS [1993]) we get

$$\int_{-\infty}^{\infty} \frac{x^2-\varepsilon^2/(a-1)}{\varepsilon^2+x^2} f(x)\mathrm{d}x = 0. \tag{10}$$

(If $T \neq 0$ holds we have to substitute $(x-T)^2$ *instead of* x^2.) The formula for ε_{CML} given in Eq. 4 is obviously a special case of Eq. 10 for $a=2$, i.e., for the Cauchy type.

Standard most frequent value calculations (M, $k=2$) work optimally in the case of the geostatistical (or simply statistical) distribution type which can be characterized by the type parameter $a=5$ (see e.g. the Table at the end of STEINER [1991]; the corresponding density curve is given in STEINER, HAJAGOS [1993] Fig. 1). If this estimation method is called the *s*tatistical *m*aximum *l*ikelihood method and is denoted by *SML* (analogously to *CML* which abbreviates the expression 'Cauchy *m*aximum *l*ikelihood'), the scale parameter can be denoted by ε_{SML}. As a special case of Eq. 10,

$$\text{for } \varepsilon_{SML} \quad \int_{-\infty}^{\infty} \frac{x^2-\varepsilon^2/4}{\varepsilon^2+x^2} f(x)\mathrm{d}x = 0 \tag{11}$$

must hold. Calculating with just the same $\psi(x)=x/(1+x^2)$ known from Eqs. 2a and 2b, the corresponding parameter of location is also denoted by *SML* (as the estimation method itself). Of course $(x-SML)^2$ is to be written instead of x^2 in Eq. 11 if $f(x)$ is not symmetrical to the origin. (If the type parameter t used in Eq. 5 is equal to 1/4 we get the density function of the statistical distribution, see Table I.)

As Fig. 12 shows, ε_{SML} behaves similarly to ε_{CML} but the inflection of the curve is at n_{out}/n = 20 %. Consequently the same breakdown bound value is shown in Fig. 11., see the *SML* curve — which (interesting enough) is nearer to the curve of the expected values (E) than to its own counterpart (M, $k=2$): the latter has a breakdown bound value of n_{out}/n = 41%. This pair of estimations stresses even more that the statistically and information theoretically based choice of the scale parameter determination has significant advantages also in respect of the resistance over the 'automatic' application of the maximum likelihood principle. In general, questions of scale parameter definitions and/or determinations must not be treated as second order problems even if the goal is a possible unbiased determination of the location parameter characterized by possible minimum statistical uncertainty.

REFERENCES

ANDREWS D. F., BICKEL P. J., HAMPEL F. R., HUBER P. J., ROGERS W. H., TUKEY J.W. 1972: Robust estimates of location. Survey and Advances. Princeton University Press, Princeton, New Jersey, 373 p

CSERNYÁK L., STEINER F. 1991: Investigations concerning the practical computation of the most frequent value of data systems. *In:* The Most Frequent Value: Introduction to a modern conception of statistics. (ed. F. STEINER) Chapter VI. pp. 83-101, Akadémiai Kiadó, Budapest

HAJAGOS B. 1985: Die verallgemeinerten Studentschen t-Verteilungen und die häufigsten Werte. Publ. Techn. Univ. Heavy Ind., Series A Mining, **40**, (1-4), pp. 225-238

HAJAGOS B. 1988: Ausreisser-Insensitivität für asymmetrisch liegende Datengruppen im Falle von verallgemeinerten häufigsten Werten. Publications of the Technical University for Heavy Industry, Series A Mining. **45**, (1-4), pp. 125-132

HAJAGOS B. 1991: The most frequent value as estimation with minimum information loss. *In:* The Most Frequent Value: Introduction to a modern conception of statistics. (ed. F. STEINER) Chapter X. pp. 135-155, Akadémiai Kiadó, Budapest

HUBER P. J. 1982: Robust Statistics. Wiley, New York, 308 p.

SHORT J. 1763: Second paper concerning the parallax of the sun etc. Philos. Trans. Roy. Soc. London, **53**

STEINER F. 1988a: Konzentrationsempfindlichkeit der verallgemeinerten häufigsten Werte. Publications of the Technical University for Heavy Industry, Series A Mining. **45**, (1-4), pp. 133-150

STEINER F. 1988b: Einige Fragen der Modellierung der Wahrscheinlichkeitsverteilungstype der Geophysik auf Grund von Stichproben. Publications of the Technical University for Heavy Industry, Series A Mining. **45**, (1-4), pp. 209-228

STEINER F.(ed.) 1991: The Most Frequent Value: Introduction to a modern conception of statistics. Akadémiai Kiadó, Budapest, 315 p.

STEINER F. 1991: Distortion of error characteristics by outliers. Acta Geod., Geoph., Mont. Hung. **26**, (1-4), pp. 453-468

STEINER F., HAJAGOS B. 1993: Practical definition of robustness (present issue)

TARANTOLA A. 1987: Inverse Problem Theory. Elsevier, Amsterdam, 613 p.

REZISZTENCIA-VIZSGÁLATOK. A SKÁLAPARAMÉTER-FORMULA MEGVÁLASZTÁSÁNAK FONTOSSÁGA

HAJAGOS Béla és STEINER Ferenc

Ha csak sommás megállapításokra korlátozódunk, egyetérthetünk azzal a nézettel, hogy az „eloszlástípusra nézve robusztus" és „durva hibájú adatokkal szemben rezisztens" tulajdonságok gyakorlatilag szinonim fogalmaknak tekinthetők, noha fogalmilag persze különböznek egymástól. Ha azonban egy durvahiba-modellre vonatkozóan két statisztikai eljárás kvantitatív összehasonlítása válik szükségessé, kiderülhet, hogy a kevésbé robusztus eljárás mutat jelentősen nagyobb rezisztenciát. A „mintamedián" és az „α-levágott átlag" (α=0,1) becslés-pár szolgálhat a fentiekre példaként.

A dolgozat bemutatja ezenfelül, hogy a helyparaméter meghatározásakor a skálaparaméter-definíció helyes megválasztása kulcsfontosságú lehet. Távoli durvahibák esetén például a szokásosan *CML*-lel jelölt helyparaméter-becslés lényegesen kisebb rezisztenciájú, mint az *MFV*-vel jelölt, noha

a ψ-függvényeik azonosak. (*CML* a „Cauchy maximum likelihood"-ból képzett betűszó, mivel a *CML*-meghatározásnál a skálaparaméter-meghatározás is ezen elv alapján történik; *MFV* itt a „most frequent value" betűszava a $k=1$ variáns esetére.) Megjegyzendő még, hogy a durvahibák maximálisan elfogadható mértéke (a klasszikus „breakdown bound"-értelemben) szintén nagyobb az *MFV*-számításra, mint a *CML*-számítások esetén.

GEOPHYSICAL TRANSACTIONS 1993
Vol. 38. No. 4. pp. 231-238

COMMENT ON AN OLD DOGMA: 'THE DATA ARE NORMALLY DISTRIBUTED'

Péter SZŰCS*

Attention is called to the dangers applying the χ^2-test in normality investigations. As is well known, the χ^2-test is one of the most frequently used methods for normality investigations when the hypothetical distribution is Gaussian. The Monte-Carlo simulations carried out show that the χ^2-test at the usual significance levels find different distributions (significantly differing from the Gaussian one) from the Gaussian distribution. This situation is termed the 'trap of the χ^2-test' and it may further strengthen the lack of credibility of the predominant presence of Gaussian mother distributions.

Keywords: χ^2-test, normality investigation, significance level, probability

1. Introduction

Depending on the type of probability distribution some authors directly reject the appearance of Gaussian distributions as being mother ones [MOSTELLER, TUKEY 1977, TUKEY 1977]. For example we can read on p. 661 of TUKEY [1977]: 'When the underlying distribution, as always, is nongaussian...'.

We can use several so called normality tests to check whether a sample originates from Gaussian distribution or not. One of the most frequently used methods for normality investigations is the χ^2-test. In this we almost always utilize the sample mean and the standard deviation as parameters, i.e. we carry out the test of goodness of fit [VINCZE 1968]. The question arises whether the level of probability of the χ^2-test finds some distributions different from the normal one — as is Gaussian distribution. We performed Monte-Carlo investigations to answer the question. Taking our results into consideration we

* University of Miskolc, Department of Geophysics, H-3515 Miskolc-Egyetemváros
Manuscript received: 22 November, 1993

suggest, as a first step, another test [CSERNYÁK 1989] instead of the χ^2-test for a given distribution family.

2. Dangers of the χ^2-test

HAJAGOS [1988] carried out Monte-Carlo investigations that indicated the dangers of the χ^2-test. At that time however the investigations could not have been expanded to sufficiently great sample and repetition numbers because of the limitations of the domestic computer field. We therefore felt justified in carrying out similar investigations as the present level of computer sciences can now offer us far more scope.

What type of distributions do we submit to the χ^2-test? We investigated three different representatives of the $f_a(x)$ supermodel. We can define the supermodel in the following manner [STEINER 1990]:

$$f_a(x) = n\,(a) \cdot \frac{1}{\left(\sqrt{x^2+1}\,\right)^a} \quad (a>1)\,. \tag{1}$$

where a is the type parameter, since the tails of the distribution functions are wider when the values of a are small. When the values of a are great, the tails will be much shorter and the maximum will be flatter. It can be proved that for $a \to \infty$ the standard form approaches the Gaussian distribution function. The $n(a)$ figuring in (1) is a normalization factor and can be calculated as follows:

$$n(a) = \frac{\Gamma\left(\frac{a}{2}\right)}{\sqrt{\pi} \cdot \Gamma\left(\frac{a-1}{2}\right)}\,. \tag{2}$$

The $f_a(x)$ model-family is able to model the cases that may occur in practice. If we have no preliminary information about the type of data distribution, the application of a=5 can be offered for geostatistical tasks [STEINER 1991, page 298, fig.1]. Let us take this a=5 type as one of our investigated distributions. The a=9 distribution was named after JEFFREYS [1961]. This is a representative of the distributions with the shortest tails, which are likely to occur in the geosciences. Thus, our second investigated distribution will be the Jeffreys one. Our third distribution will be the $f_3(x)$. This represents a distribution with wide tails, but it is still not Cauchy type.

During the Monte-Carlo investigations we created samples with 100 and 400 elements from the above mentioned distributions with the aid of a random generator. We repeated the sampling a thousand times. After finishing the χ^2-tests we were able to calculate probability values to an accuracy of two decimal places for different significance levels. These values show with how much probability the χ^2-test would accept the given type and size samples as normally distributed ones at the given significance level. The thousandfold sampling proved to be reliable. When we repeated the investigations, there was only a negligible fluctuation in the third decimal figure of the probability values. We can see the detailed results of the investigation in Figs. 1 and 2. The curves have great probability values. For the samples with 100 elements (see *Fig. 1*) we accept our data originating from geostatistical (a=5) distribution as normally

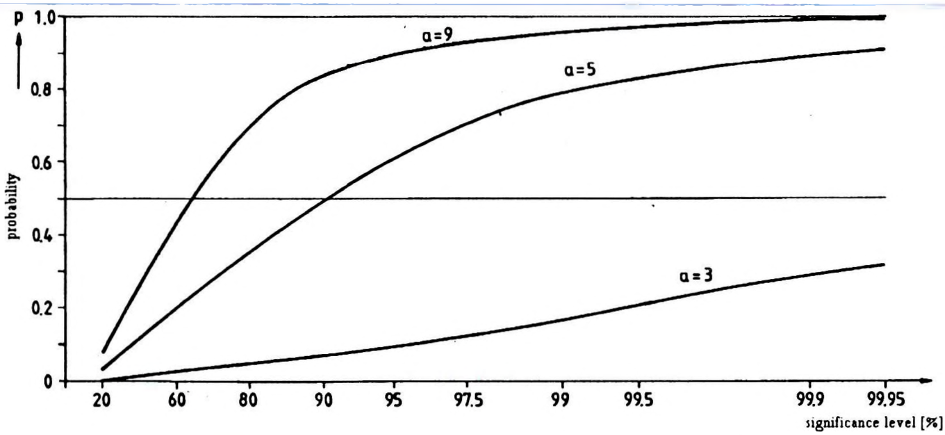


Fig. 1. Probabilities of acceptance of the Gaussian hypothesis at the given significance levels (χ^2-test, n=100)

1. ábra. A Gauss-hipotézis elfogadásának valószínűségei az adott szignifikancia szinteken (χ^2-próba, n=100)

distributed ones in half of the instances at the 90 percentile significance level. In the case of a=9 the situation is even worse: the concrete probability value is 0.842 at the 90 percentile significance level. For the samples with 400 elements the situation is slightly better although the probabilities remain high enough henceforward (*Fig. 2*). In the case of a=3 there was no 'acceptance'. Based on the χ^2-test we would even say, with high probability, that our samples with 400 elements originated from the Jeffreys distribution as normally distributed ones.

These findings can be termed the 'trap of the χ^2-test' that may further strengthen the lack of credibility of the predominance of Gaussian mother distributions. From the practical aspect this situation has a harmful effect on those users who apply the least squares method without deeper consideration and investigation. From the theoretical aspect this can lead to the general acceptance of the standard deviation as a universal uncertainty property, and we may wrongly take into account the message and the validity domain of the Heisenberg relation [CSERNYÁK, STEINER 1991].

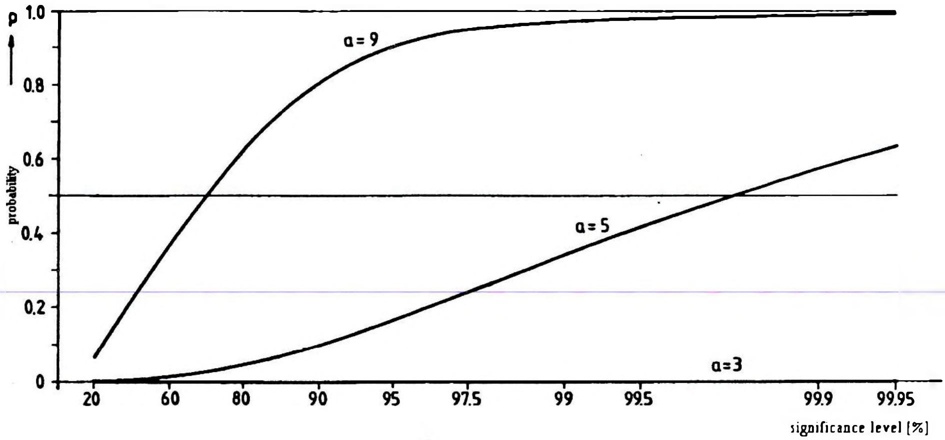


Fig. 2. Probabilities of acceptance of the Gaussian hypothesis at the given significance levels (χ^2-test, n=400)

2. ábra. A Gauss-hipotézis elfogadásának valószínűségei az adott szignifikancia szinteken (χ^2-próba, n=400)

The question may arise, with how much probability we would accept the Gaussian hypothesis for the χ^2-test if our samples originated from any member of the $f_a(x)$ supermodel. To answer the question we should know with what degree of probability the different a values in the $f_a(x)$ supermodel would occur. During our investigation we applied two different distribution functions that are able to model the occurrence probabilities [see Fig. 4, and Eqs. 10 and 11 of STEINER, HAJAGOS 1993]. These are as follows:

$$\text{I.} \qquad f_D\left(\frac{1}{a-1}\right) = \frac{16}{a-1} \cdot e^{-\frac{4}{a-1}}, \tag{3}$$

$$\text{II.} \qquad f_J\left(\frac{1}{a-1}\right) = \frac{64}{a-1} \cdot e^{-\frac{8}{a-1}}. \tag{4}$$

We summarize our results in *Table I.* Naturally the results of the table were not calculated from infinite different distributions. We obtained the numerical values in a similar way to the way in which we completed the χ^2-tests for eleven different distributions of the $f_a(x)$ supermodel, and we integrated numerically the results weighted with (3) and (4) probability distributions.

	Significance levels									+
	20%	60%	80%	90%	95%	97.5%	99%	99.5%	99.9%	99.95%
n=100 I., f_D	0.027	0.168	0.297	0.378	0.456	0.525	0.563	0.594	0.651	0.668
n=400	0.018	0.072	0.130	0.183	0.228	0.267	0.315	0.349	0.419	0.441
n=100 II., f_J	0.045	0.263	0.499	0.562	0.656	0.718	0.769	0.797	0.844	0.858
n=400	0.042	0.154	0.262	0.346	0.410	0.460	0.516	0.556	0.637	0.662

Table I. Probability values for the acceptance of the Gaussian hypothesis when using the χ^2-test at the given significance levels if our distribution originated from the $f_a(x)$ supermodel with f_D or f_J probability distributions

I. táblázat. Valószínűségek a Gauss-hipotézis elfogadására χ^2-próba alkalmazása esetén az adott szignifikanciaszinteken, ha eloszlásuk az $f_a(x)$ szupermodellből származik f_D vagy f_J valószínűségsűrűséggel

The rows belonging to I were calculated with the help of (3), the values belonging to II were calculated with the aid of (4). For (3) the geostatistic distribution (a=5) occurs with the greatest probability whereas in the case of (4) the most probable distribution is the Jeffreys one (a=9). The large probabilities we find in the table tend to underline the dangers of applying the χ^2-test. For example, even for samples with 400 elements the probabilities of acceptance of the Gaussian hypothesis are 0.315 and 0.516. These are very great probability values, especially if we take it into consideration that in the case of (3) and (4) the occurrence probability of Cauchy distribution is still not negligible.

3. The Csernyák test

It is a well known result of mathematical statistics that the distribution function of the 'extent' of the sample with n elements

$$R = X_{\max} - X_{\min} \tag{5}$$

is associated with the type of mother distribution [CRAMÉR 1946]. The sample size cannot be regarded as statistics that characterize the distribution because R is obviously proportional to the scale parameter (S) as well as to the sample size. We neglect S if we compare R to the empirical interquartile range determined from the same sample in the following manner:

$$C = \frac{R}{2q_{emp}} \tag{6}$$

We can accept this as the statistical function of the test for type determination [CSERNYÁK 1989]. This expression is suitable for normality investigations so we refer to the procedure as the Csernyák test.

On the basis of our calculations it can be stated that the Csernyák test is more reliable in the applied type range. Our results are shown in *Figs. 3* and *4*. If these figures are compared with Figs. 1 and 2 it can be realized that in case of the Csernyák test we accept the samples as Gaussian type with much less probability than in the case of the χ^2-test.

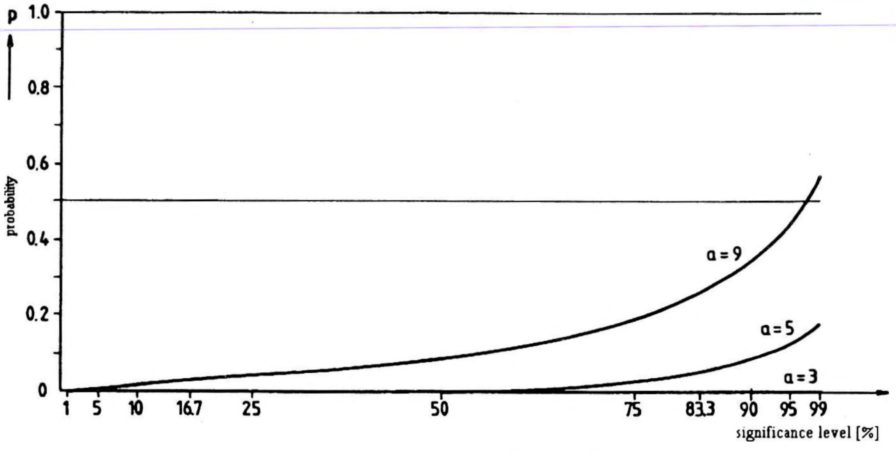


Fig. 3. Probabilities of acceptance of the Gaussian hypothesis at the given significance levels (Csernyák test, n=100)

3. ábra. A Gauss-hipotézis elfogadásának valószínűségei az adott szignifikancia szinteken (Csernyák teszt, n=100)

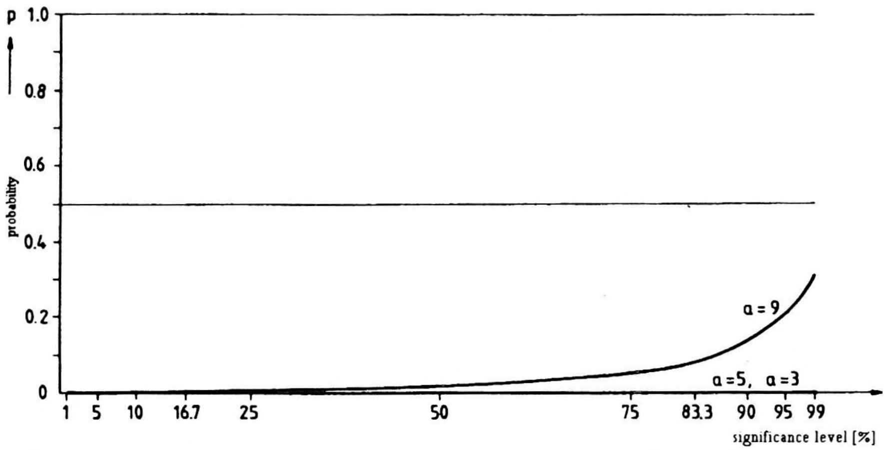


Fig. 4. Probabilities of acceptance of Gaussian hypothesis at the given significance levels (Csernyák test, n=400)

4. ábra. A Gauss-hipotézis elfogadásának valószínűségei az adott szignifikancia szinteken (Csernyák teszt, n=400)

It might well be said that the Csernyák test supposes freedom from outliers. Although this may be true, it does not alter the situation: it makes no difference whether the great value of R is caused by outlier free types with heavier tails than the tails of normal distribution, or by the appearance of outliers. The rejection of the hypothesis calls attention in both cases to the need to handle the methods of traditional statistics cautiously.

4. Conclusions

Based on the results of Monte-Carlo investigations we can establish the following facts:

— the χ^2-test cannot be recommended for the normality tests of different distributions occurring in the practice of geosciences. Even if our samples are quite different from the Gaussian distribution, the χ^2-test accepts them as normally distributed ones with large probabilities at the most frequently used significance levels;

— when applying the χ^2-test the lack of credibility of the predominant presence of Gauss mother distribution may contribute to the survival of the traditional (not robust and not resistant) statistical algorithms;

— for measured data sets we would suggest the use of the Csernyák test as a first step if our distribution originates from the f_a supermodel.

REFERENCES

CRAMÉR H. 1946: Mathematical methods of statistics. Princeton University Press, Princeton, IV. J.

CSERNYÁK L. 1989: Determination of type using sample range. Acta Geodaetica, Geophysica et Montanistica, Acad. Sci. Hung. **24**, 3-4, pp. 441-447

CSERNYÁK L., STEINER F. 1991: The inadequacy of the Heisenberg relation in generally posed questions of uncertainties. *In*: The Most Frequent Value: Introduction to a modern conception of statistics. (ed. F. STEINER) Appendix IX. pp. 271-295, Akadémiai Kiadó, Budapest

HAJAGOS B. 1988: Normalitätsuntersuchungen mit Hilfe der χ^2-Probe an Stichproben, die aus Student-schen Mutterverteilungen stammen. Publications of the Technical University for Heavy Industry, Miskolc Series A. Mining, Vol. **44**. pp. 217-230

JEFFREYS H. 1961: Theory of Probability. Clarendon Press, Oxford

MOSTELLER F., TUKEY J. W. 1977: Data analysis and regression. Addison — Wesley Reading, Mass

STEINER F. 1990: A geostatisztika alapjai. Tankönyvkiadó, Budapest, 363 p.

STEINER F. (ed.) 1991: The Most Frequent Value. Akadémiai Kiadó, Budapest (Hungary), 315 p.

STEINER F., HAJAGOS B. 1993: Practical definition of robustness. (present issue)

TUKEY J.W. 1977: Exploratory data analysis. Addison- Wesley, Reading, Mass

VINCZE I. 1968: Matematikai statisztika ipari alkalmazásokkal. Műszaki Könyvkiadó, Budapest, 352 p.

MEGJEGYZÉS EGY RÉGI DOGMÁHOZ: „AZ ADATOK GAUSS-ELOSZLÁSÚAK”

SZŰCS Péter

Ez a cikk a χ^2-próba normalitásvizsgálatbeli alkalmazásának a veszélyeire szeretné felhívni a figyelmet. Mint jól ismert, az egyik leggyakrabban alkalmazott módszer a normalitásvizsgálatra a χ^2-próba, amikor a hipotetikus eloszlás a Gauss-féle. Az elvégzett Monte-Carlo vizsgálatok azt mutatják, hogy a χ^2-próba a szokásos szignifikanciaszinteken nagy valószínűséggel Gauss-eloszlásúnak talál attól szignifikánsan különböző eloszlásokat. Ezt akár a „χ^2-próba csapdájának” is nevezhetnénk, ami tovább erősítheti a Gauss-eloszlás anyaeloszlásként való túlnyomó előfordulásának a tévhitét.

GEOPHYSICAL TRANSACTIONS 1993
Vol. 38. No. 4. pp. 239-249

COMPARISON OF THE KARHUNEN-LOÈVE STACK WITH THE CONVENTIONAL STACK

Leif BRULAND*

Several applications of the Karhunen-Loève (KL) transform to seismic data are known, among which is the use of the first principal component as an alternative stack — the KL stack. On analysing and comparing the KL stack with the conventional stack, it was found that the KL stack is more influenced by noise, especially coherent noise, than the conventional one. With approximately the same signal amplitudes from trace to trace, the conventional stack is therefore the better choice. On the other hand, if the signal amplitudes vary and the noise is uncorrelated with approximately constant energy on all traces, the KL stack should be preferred.

It has been claimed that the KL stack is relatively insensitive to small time shifts of the signals, and that correction for residual statics may be unnecessary when the KL stack is used. It is confirmed here that the KL stack generally gives the better signal-to-noise ratio in such cases. However, the time shifts may seriously distort the output signal, and the distortion is found to be very sensitive to changes in the time shifts, in view of which it is important to correct for residual statics even if the KL stack is used.

Keywords: seismic, stacking, Karhunen-Loève Transformation

1. Introduction

The Karhunen-Loève Transform (KLT) is used to represent a set of, say, M input vectors or traces by a particular set of M orthogonal vectors called principal components. The principal components are linear combinations of the input vectors constructed in such a way that most of the coherent energy is contained in the first component, or in the first few components. The KLT can therefore be used to express information in a compact way. The principal components have long been used in multivariate statistical analysis both for data reduction and in interpretation.

* Institute of Solid Earth Physics, University of Bergen, Allegaten 41, N-5007 Bergen, Norway
Manuscript received: 18 June, 1993.

Since the first principal component, which can be looked upon as a weighted stack, usually contains most of the coherent energy from the input data, it may be used as an alternative stack. This was demonstrated by HEMON and MACE [1978], who initially suggested the application of the KLT to seismic data. Several other applications of the KLT to seismic data were later presented by ULRYCH et al. [1983], LEVY et al. [1983], JONES, LEVY [1987], YEDLIN et al. [1987] and FREIRE, ULRYCH [1988].

In this paper we are mainly concerned with the use of the first principal component as an alternative stack, hereafter called a KL stack. After a short introduction to the theory of the KLT, the properties of the KL stack are explored and compared with those of the conventional stack.

2. The Karhunen-Loève Transform

Let the data be given as

$$\bar{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iN})^T, \ i = 1, 2, \ldots, M \tag{1}$$

where M is the number of traces, and N the number of samples per trace, $M<N$. All traces are assumed to have zero mean values.

We now search for a vector $\bar{y}$ as a linear combination of the $\bar{x}$'s

$$\bar{y} = \sum_{i=1}^{M} a_i \bar{x}_i = X\bar{a} \tag{2}$$

where $X = \{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_M,\}$, and $\bar{a} = (a_1, a_2, \ldots, a_M)^T$.

The energy (or variance) of $\bar{y}$ is then

$$V(\bar{y}) = \bar{y}^T \bar{y} = \bar{a}^T \ X^T X\bar{a} = \bar{a}^T C\bar{a}. \tag{3}$$

where $C = X^T X$ is the covariance matrix of the data.

The first principal component is defined as the vector $\bar{y}$ that maximizes $V(\bar{y})$ under the restriction

$$\bar{a}^T \bar{a} = \sum_{i=1}^{M} a_i^2 = 1. \tag{4}$$

Maximizing (3) subject to (4) is equivalent to maximizing

$$f(\bar{a}, \lambda) = \bar{a}^T C\bar{a} + \lambda\,(1-\bar{a}^T \bar{a}), \tag{5}$$

where λ is a Lagrange multiplier. Differentiation of $f(\bar{a}, \lambda)$ and equating the result to 0 leads to:

$$\frac{\partial f(\bar{a}, \lambda)}{\partial \bar{a}} = 2C\bar{a} - 2\lambda\bar{a} = 0$$

or

$$(C - \lambda I)\,\bar{a} = 0. \tag{6}$$

From (6) it follows that λ must be an eigenvalue and $\bar{a}$ the associated eigenvector of C. Therefore we must have

$$\bar{a}^T C\bar{a} = \bar{a}^T(\lambda\bar{a}) = \lambda,$$

and the solution to the maximization problem is the eigenvector corresponding to the largest eigenvalue of C (all eigenvalues of C are ≥ 0).

The next principal component is found from (6) when $\bar{a}$ is the eigenvector associated with the next largest eigenvalue, and so on. We can thus write

$$Y = XA \tag{7}$$

where

$$A = \{\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_{M},\}$$

Since C is symmetric, the eigenvectors are orthogonal, and $A^T = A^{-1}$. Multiplication of (7) by A^T gives

$$X = YA^T, \tag{8}$$

which is then the inverse transformation.

The variance of the i^{th} trace is

$$V(\bar{x}_i) = \bar{x}_i^T\,\bar{x}_i.$$

The variance of $\bar{y}_i$ is

$$V(\bar{y}_i) = \bar{y}_i^T\,\bar{y}_i = \bar{a}_i^T X^T X\bar{a}_i = \lambda_i\,,$$

and the eigenvalues are therefore just the energy or variance of the principal components.

The total energy of the input data is

$$\sum_i V(\bar{x}_i) = \text{Trace}\,[X^T X] = \sum_i \lambda_i = \sum_i V(\bar{y}_i)\,. \tag{9}$$

From this it follows that the total energy is invariant under the transformation.

Since $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M$, most of the energy is contained in the first principal components. We can therefore approximate X by a linear combination of the principal components with largest energy, say the first $P<M$ components:

$$X \approx Y^{N\times P} \cdot A^{P\times M} . \tag{10}$$

The amount of reconstructed energy can be calculated from

$$E(P) = \frac{\sum_{i=1}^{P} \lambda_i}{\sum_{i=1}^{M} \lambda_i} . \tag{11}$$

3. KLT and Singular Value Decomposition (SVD)

An SVD of the data matrix X also leads to the matrix of coefficients, A, and the matrix of principal components, Y. To see this we start with the matrix

$$B = \begin{bmatrix} 0 & X \\ X^T & 0 \end{bmatrix}, \; B^T = B \tag{12}$$

The eigenvalue problem for this matrix can be written

$$\begin{bmatrix} 0 & X \\ X^T & 0 \end{bmatrix} \begin{bmatrix} \bar{u} \\ \bar{v} \end{bmatrix} = l \begin{bmatrix} \bar{u} \\ \bar{v} \end{bmatrix}, \tag{13}$$

where $\bar{u}$ is an $(N \times 1)$ vector, $\bar{v}$ is an $(M \times 1)$ vector, and l is an eigenvalue. Since B is symmetric, l will be real. From (13) we get

$$X\bar{v} = l\bar{u}, \;\; X^T\bar{u} = l\bar{v} \tag{14}$$

Premultiplication of the two equations by X^T and X, respectively, gives

$$X^T X\bar{v} = l^2\bar{v}, \;\; XX^T\bar{u} = l^2\bar{u} \tag{15}$$

We thus see that $\bar{v}$ is an eigenvector of $C=X^TX$ and $\lambda=l^2$ the associated eigenvalue, while $\bar{u}$ is an eigenvector of XX^T with the eigenvalue λ.

For convenience we assume the rank of C and XX^T to be M. C is then a positive definite matrix, and therefore all eigenvalues are greater than zero. We order the eigenvalues so that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M \geq 0$, and let the corresponding eigenvectors be normalized so that $\bar{v}_i^T \bar{v}_i = 1, \; \bar{u}_i^T \bar{u}_i = 1, \; i=1,2,\ldots,M$. We then define the matrices V and U as

$$V = \{\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_M\}$$

$$U = \{\bar{u}_1, \bar{u}_2, \ldots, \bar{u}_M\}$$

and define Λ as the matrix with the eigenvalues as its diagonal elements with zeroes elsewhere.

From (14) we get

$$XV = U\Lambda \tag{16}$$

Comparing (16) with (7) we find that

$$A = V$$
$$Y = U\Lambda \tag{17}$$

Postmultiplication of (16) with V^T gives the decomposition of X

$$X = U\Lambda V^T = YA^T$$

which according to (8) is equivalent to the inverse transformation.

4. Comparison of the KL stack with the conventional stack

Some properties of the KL stack are more easily revealed by observing that the principal components can be derived in a different way.

From (17) it follows that the principal components are scaled versions of the first M eigenvectors of XX^T. These eigenvectors can also be found from a maximization problem, viz.

$$\max\,(\bar{u}^T XX^T\bar{u}) \tag{18}$$

under the restriction

$$\bar{u}^T\bar{u} = 1$$

Since this leads to exactly the same sort of problem as was defined by (5), only with $X^T X$ replaced by XX^T, $\bar{u}$ will be the eigenvector of XX^T that is associated with the largest eigenvalue λ. The first principal component is just $\bar{y} = \lambda\bar{u}$. But expression (18) can be written

$$\max\,(\bar{u}^T XX^T\bar{u}) = \max\sum_{i=1}^{M} (\bar{x}_i^T\bar{u})^2 \tag{19}$$

Thus, the normalized first principal component maximizes the sum of the square of the inner products between this component and the traces. It can be easily shown that the normalized conventional stack maximizes the sum of the inner products. In summary, The normalized first principal component, $\bar{u}$, maximizes $\sum_i (\bar{x}_i^T \bar{u})^2$

The normalized conventional stack, $\bar{s}$, maximizes $\sum_i (\bar{x}_i^T \bar{s})$

From these properties of $\bar{u}$ and $\bar{s}$ we can draw some conclusions:
If one or more traces are reversed in polarity, this will have no influence on the KL stack. This is shown by *Fig. 1.* where exactly half of the traces have been reversed in polarity so that the conventional stack becomes a zero trace.

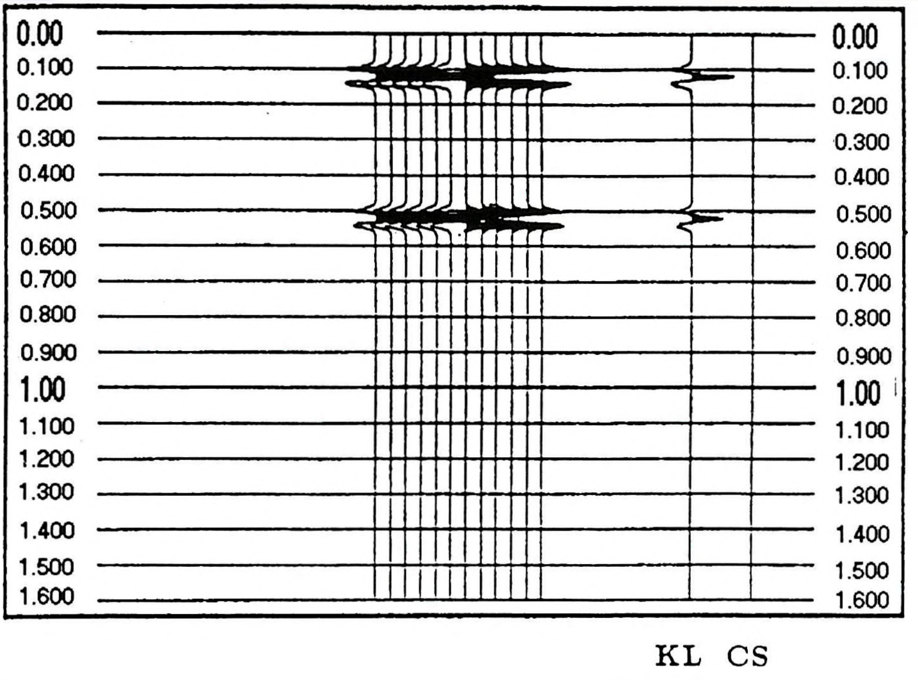


Fig. 1. Traces may be reversed in polarity without affecting the KL stack. In this example 6 out of 12 identical traces have been reversed in polarity. The comventional stack (CS) gives a zero trace, the KL stack (KL) reproduces the input trace

1. ábra. A csatornák polraitása megfordulhat anélkül, hogy a KL stacking eredményét megváltoztatná. A példán 12 azonos csatornából 6 ellentétes polraitású. A hagyományos stacking (CS) zéró csatornát eredményez, a KL stacking (KL) a bemeneti csatornát adja vissza

If the noise is uncorrelated from trace to trace, and all traces have identical signals and the same signal-to-noise ratio, the conventional stack is the optimum (weighted) stack. In this case the weights in (2) will also be equal, and therefore the KL stack is also optimum. Now, if the noise energy varies from trace to trace, the KL stack will be most influenced by the traces with highest noise energy. This is true whether the noise is correlated or not, but the effect will be more pronounced if the noise is coherent over two or more traces. This result is illustrated in *Fig. 2.*

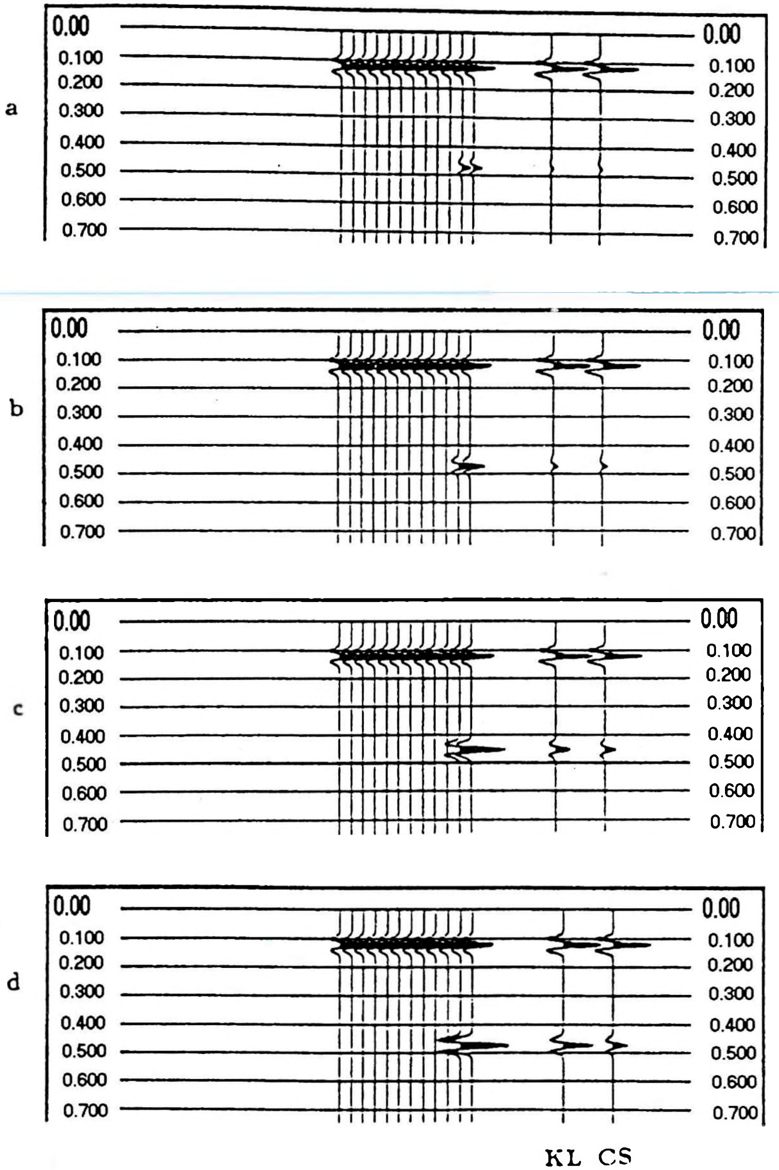


Fig. 2. Illustration of the effect of a 'noise signal' present on 2 traces (in this case). When the noise energy is small compared with the signal energy, there are no discernible differences between the KL stack (KL) and the conventional stack (CS) (a and b). With an increasing relative amount of noise energy, the differences become quite pronounced (c and d). The stacks have been scaled to equal signal amplitudes

2. ábra. Két csatornán jelen lévő „zavar jel" hatása. Ha a zaj energiája a jel energiájához képest kicsi, a KL stacking (KL) és a hagyományos stacking (CS) között nincsenek észrevehető különbségek (a és b). A zaj energiája relativ hányadának növekedésével a különbségek meglehetősen hangsúlyozottak lesznek (c és d). A stackingeket az egyenlő jel amplitudókhoz igazítottuk

The synthetic input data to the left of Fig. 2a, contain one coherent signal (identical on all traces) and a 'noise' signal present on only two out of twelve traces. The KL stack and the conventional stack are shown to the right. The stacks have been normalized to the signal amplitude. There is no visually discernable difference between these stacks but the traces with noise were given slightly higher weights than the other traces in the KL stack. If we increase the noise energy on the two input traces, these traces will be given successively higher weights in the KL stack (Fig. 2b-2d).

It may be illustrative to calculate the weights for traces with and without noise in a case like the last one.

Let the traces be given as

$$\bar{x}_i = \bar{s}, \qquad i = 1, 2, ..., m$$
$$\bar{x}_i = \bar{s} + \bar{n}, \quad i = m+1, m+2, ..., M$$

We assume $\bar{s}^T \bar{n} = 0$ (i.e., no overlap between coherent noise and signal), and denote $\bar{s}^T \bar{s} = a$, $\bar{n}^T \bar{n} = b$ and $(\bar{s}+\bar{n})^T (\bar{s}+\bar{n}) = a+b = c$. We then have

$$X^T X = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}, \tag{20}$$

where A is an $m \times m$ matrix, B is $m \times (M-m)$ and C is $(M-m) \times (M-m)$. The elements in A and B are all equal to a, and those of C are all equal to c. The eigenvalue-eigenvector problem is then

$$\begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \begin{bmatrix} \bar{u} \\ \bar{v} \end{bmatrix} = \lambda \begin{bmatrix} \bar{u} \\ \bar{v} \end{bmatrix}, \tag{21}$$

where $\bar{u}$ is an $m \times 1$ and $\bar{v}$ an $(M-m) \times 1$ vector. The eigenvector associated with the largest λ has only two different elements, since all elements in $\bar{u}$ must be equal, and so must all elements in $\bar{v}$. These values, which we denote u and v, respectively, are the weights given to traces without noise and with noise in the calculation of the KL stack.

The system is now reduced to

$$\begin{bmatrix} ma & (M-m)\,a \\ ma & (M-m)\,c \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \lambda \begin{bmatrix} u \\ v \end{bmatrix}, \tag{22}$$

Solving for λ gives

$$\lambda = \frac{1}{2} [(M-m)\,c+ma] + \frac{1}{2} \sqrt{[(M-m)\,c+ma]^2 - 4ma\,(M-m)\,(c-a)}\,, \tag{23}$$

and the ratio v/u becomes

$$\frac{v}{u} = \frac{\lambda}{\lambda - b\,(M-m)}$$

Since the total energy of the traces is $E = (M-m)\,c + ma$ and the total noise energy is $N = b\,(M-m)$, we get the following inequality for the ratio v/u

$$\frac{E}{E-2N} > \frac{v}{u} > \frac{E}{E-N} \tag{24}$$

We thus see that the noise traces will always get larger weights in the calculation of the KL stack, even if the noise is present on only one trace. It may be concluded that as long as the signal is completely coherent with constant amplitudes from trace to trace, the conventional stack should be preferred to the KL stack irrespective of the noise structure.

It has been claimed that the KL stack is relatively insensitive to small trace-to-trace time shifts of the signal, and therefore residual static correction can often be avoided when the KL stack is used [HEMON, MACE 1978, ULRICH et al. 1983]. However, this is only partly true, as can be seen from the following argument.

One choice of the weights, a_i, in equation (2) which satisfies (4) is $a_k = 1$, $a_i = 0$ when $a \neq k$. Thus the energy of the first principal component is always greater than or equal to the energy of the trace with the highest energy. This means that even if the signal is somewhat out of phase from trace to trace, the signal will not be cancelled by a KL stack as it might be by a conventional stack. With uncorrelated noise, the S/N ratio will therefore be higher in the KL stack than in the conventional stack. However, there is no guarantee that the KL stack will reproduce the signal; in fact it may be highly distorted, and the form of the signal in the KL stack is very sensitive to small changes in the statics. This is illustrated in *Fig. 3*, where quite different signals appear in the KL stack although only one trace has been changed from step to step. If signal distortion is to be avoided, it is therefore necessary to perform residual static correction even if the KL stack is to be applied.

Next we consider the case with varying signal amplitudes across the traces. If the noise is approximately uncorrelated with nearly the same energy on all traces, we can use the arguments of the last example to see that in this case the KL stack is preferable to the conventional stack. This follows from the fact that the energy in the KL stack cannot be less than the energy in the trace of maximum energy, and since the difference in trace energy is due to the difference in signal energy, the S/N ratio will always be higher in the KL stack than in the single traces. This will not always be the case for the conventional stack. If the noise varies from trace to trace, the situation becomes more obscure since the relative amount of signal energy to noise energy will affect the weights in the KL stack. With an increasing amount of coherent noise, the KL stack should again be avoided.

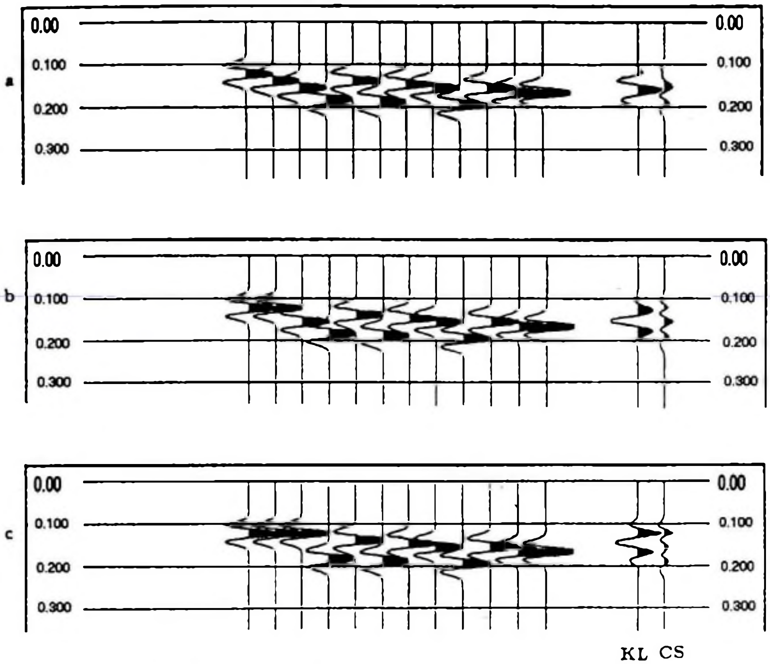


Fig. 3. KL stacks (KL) and conventional stacks (CS) of sets of traces with time shifted signals. Only one of the traces (second from the left) in Fig. 3b is different from those in Fig. 3a. In Fig. 3b and in 3c, only the third trace is different

3. ábra. KL stackingek (KL) és hagyományos stackingek (CS) csatornasorozata időben eltolt jelekkel. A 3b. ábrán csak egyetlen csatorna (balról a második) tér el a 3a. ábrán lévő csatornáktól. A 3b és a 3c ábrák között csak a harmadik csatornában van különbség

5. Conclusions

The properties of the KL stack have been analysed and compared with those of the conventional stack, and the results can be summarized as follows:

Both stacks are optimal in the case of identical signals contaminated by completely uncorrelated noise. With identical signals on all traces the conventional stack is superior to the KL stack in all other cases.

Correlated noise will always have a greater influence on the KL stack than on the conventional stack. The differences between stacking methods are small as long as the amount of noise energy is small compared with the total signal energy, but they increase rapidly with an increasing relative amount of correlated noise energy. This is true whether the noise is coherent over all traces (except for the case of identical noise 'signals' on all traces) or only a few.

If traces with residual statics are KL stacked, the S/N ratio will normally increase (and never decrease), but the signal may be highly distorted. It is therefore important to perform residual static correction even if KL stacking is to be applied.

If the signal amplitudes vary across the traces while the noise is uncorrelated and has approximately the same energy on each trace, the KL stack is a better choice than the conventional stack. This may be so even when the noise varies and/or is correlated to some extent, but it would be very difficult to prescribe which method to use in such cases.

REFERENCES

FREIRE S. L. M., ULRYCH T. J. 1988: Application of singular value decomposition to vertical seismic profiling. Geophysics **53**, 6, pp. 778-785

HEMON CH, MACE D. 1978: Essai d'une application de la transformation de Karhunen-Loève au traitement sismique. Geophysical Prospecting **26**, 3, pp. 600-626

JONES I. F., LEVY S. 1987: Signal-to-noise ratio enhancement in multichannel seismic data via the Karhunen-Loève transform. Geophysical Prospecting **35**, 1, pp. 12-32

LEVY S., ULRYCH T. J., JONES I. F., OLDENBURG D. W. 1983: Applications of complex common signal analysis in exploration seismology. Proceedings of the 53rd Annual SEG meeting, Las Vegas, S6.6

ULRYCH T. J., LEVY S., OLDENBURG D. W., JONES I. F. 1983: Applications of the Karhunen-Loève transformation in reflection seismology. Proceedings of the 53rd Annual SEG meeting, Las Vegas, S6.5

YEDLIN M. J., JONES I. F., NAROD B. B. 1987: Application of the Karhunen-Loève transform to diffraction separation. IEEE Transactions on acoustics, speech, and signal processing. ASSP-35, 1

A KARHUNEN-LOÈVE ÉS A HAGYOMÁNYOS STACKING ELJÁRÁS ÖSSZEHASONLÍTÁSA

Leif BRULAND

A szeizmikában a Karhunen-Loève (KL) transzformáció számos alkalmazása ismert, ezek közül az első főkomponensnek alternatív összegzésként való alkalmazása a KL stacking. A KL stackinget és a hagyományos stackinget elemezve és összehasonlítva, megállapítottuk, hogy a KL stackinget a zaj, különösen pedig a koherens zaj jobban befolyásolja, mint a hagyományos stackinget.

Csatornáról csatornára haladva közel azonos jelamplitúdók mellett ezért a hagyományos stacking a jobb választás. Másrészt azonban, ha a jel amplitúdója változik, és a zaj minden csatornán közel azonos energiájú és korrelálatlan, a KL stackinget kellene előnyben részesíteni.

Azt állították, hogy a KL stacking viszonylag érzéketlen a jelek kismértékű időbeli eltolódásaira, és a maradék statikus korrekció KL stacking esetében felesleges. Megerősítjük, hogy ilyen esetekben valóban a KL stacking adja a jobb jel/zaj viszonyt. Azonban, az időbeli eltolódások a kimenő jelet lényegesen torzíthatják, a torzulás nagyon érzékeny az időbeli eltolódásokra, és mindezekre való tekintettel, fontos, hogy a KL stacking esetében is végrehajtsuk a maradék statikus korrekciót.

GEOPHYSICAL TRANSACTIONS 1993
Vol. 38. No. 4 pp. 251-259

INTERCONNECTING GRAVITY MEASUREMENTS BETWEEN THE AUSTRIAN AND THE HUNGARIAN NETWORK

Géza CSAPÓ*, Bruno MEURERS**, Diethard RUESS***, Gábor SZATMÁRI*

An account is given of the comparative measurements carried out on the Hungarian and Austrian gravity base networks in the period of 1991-1993. This work includes absolute and relative gravity measurements. The absolute measurements were performed with the JILAG-6 absolute gravimeter, the relative measurements with 5 LCR gravimeters on 24 ties between selected points of the base networks along the border of the two countries.

It has been established that 40 μGal difference exists between the gravity datum of Austria and that of Hungary. To determine the source of this deviation further investigations and readjustment of the Hungarian gravity network are needed.

Keywords: gravity surveys, Austria, Hungary, network

1. Introduction

In the wake of the rapid progress in instrument design and measuring techniques the Earth sciences require the development of geodetical base networks covering as large areas as possible in order to solve the increasing numbers of theoretical and practical tasks.

For some years the gravimetric network of Austria has been connected to those of Germany, Switzerland and Italy (1985-1987), while the network of Hungary to that of former Czechoslovakia (1985-1988). Further cooperation was rendered possible by the countries of Central Eastern Europe lifting the secrecy on their base networks and striving to participate more and more intensively in joint projects initiated by international organizations. This re-

* Eötvös Loránd Geophysical Institute of Hungary, H-1145 Budapest, Kolumbusz u. 17-23
** Universität Wien, Institut für Meteorologie und Geophysik, A-1190 Wien, Hohe Warte 38
*** Bundesamt für Eich- und Vermessungswesen, A-1025 Wien, Schiffamtgasse 1-3
Manuscript received: 30 August, 1993

sulted in the conducting of connecting measurements between the Austrian and Czech, and the Slovakian gravity base networks in 1991. Similar work was performed in 1991-93 between Austria and Hungary. The framework for these projects was set up partly by bilateral agreements on scientific cooperation, partly by the 'DANREG' program started in 1989. The connecting measurements include absolute and relative gravity surveys.

2. Absolute gravity measurements

The Austrian Gravity Base Network (AGBN) contains 23 absolute stations, and at several selected points repeated determinations have been performed as well [RUESS et al. 1989]. On the basis of these measurements the AGBN point catalogue was up-dated for 1993 prior to the interconnecting measurements.

In 1989 at the time of the adjustment of Hungarian Gravity Network (HGN-80) the '*g*' values determined at five points with a GABL absolute gravimeter between 1978-80 were accepted as constraints, thus the datum level and scale of the network were determined by these values [CSAPÓ, SÁRHIDAI 1990]. In the period of 1991-93, using the JILAG-6 equipment, RUESS et al. repeated the earlier absolute measurements at four points. From data compiled in *Table I*, it is evident that the discrepancy of values determined by the two different type of instruments is substantially higher than the accuracy of absolute gravimeters [BOULANGER et al. 1991]. The examination of such conspicuous discrepancies goes beyond the scope of this paper. For the

absolute station	year	GABL JILAG-6 mGal	VG μGal/m	difference μGal	variation μGal/year
81 SIKLÓS	1978	678.291	339.4	+ 30	+ 2.3
	1991	678.321	339.4		
82 BUDAPEST	1980	824.328	250.2	- 22	- 1.9
	1991	824.307	250.0		
85 KŐSZEG	1980	784.748	267.2	- 33	- 2.5
	1993	784.715*	271.0		
86 SZERENCS	1980	872.816	290.6	- 31	- 2.2
	1993	872.785*	298.0		

* calculated with the corrected vertical gradient (VG)

Table I. Results of absolute measurements in Hungary
I. táblázat. A magyarországi abszolut mérések eredményei

comparison the Hungarian gravity data were reduced by the average of differences (15 μGal)* obtained on the four reobserved absolute stations. The repeated absolute measurements require the re-adjustment of HGN-80 as a necessity. In this respect, in 1993–94 several new absolute points have been measured in Hungary; re-adjustment of the network is due to be performed after these measurements have been completed.

The absolute measurements were processed by RUESS in the usual way, i. e. corrections with regard to systematic instrumental effects, air pressure, polar motion and height (reduction to ground level) were applied. The 'vertical gradient' measurements were performed by means of 2 LCR gravimeters with an accuracy of ±2μGal.

Using three independent sets of measurements at Kőszeg, Hungary (*Fig. 1.*) the most probable value of gravity and its error can be calculated in two different ways:

a) each drop taken as an individual measurement

b) one set (containing 1200–1800 drops) taken as one measurement.

We regarded version b as a more realistic approximation, and these values are given in Table I.

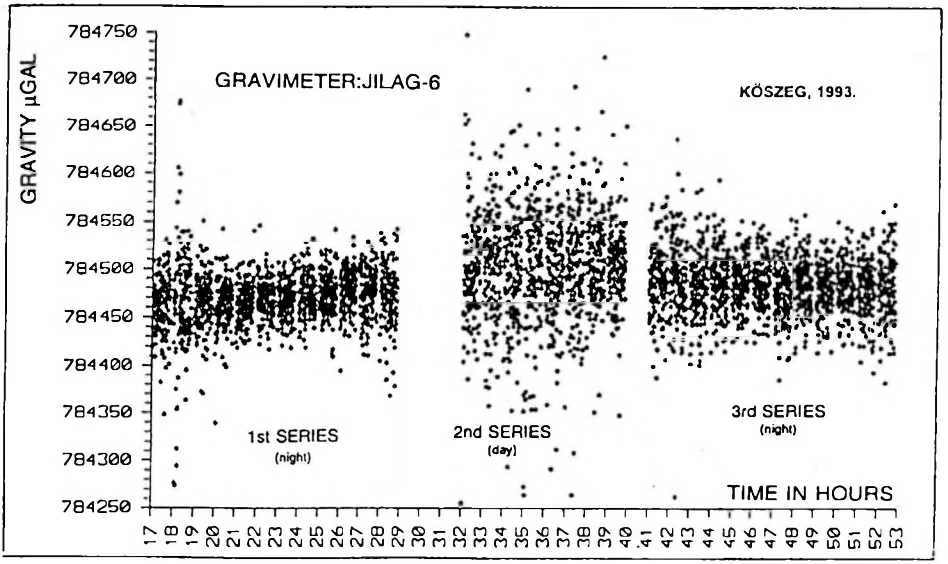


Fig. 1. Chart showing the results of absolute gravity measurements (4800 drops)

1. ábra. Abszolut módszerrel végzett nehézségi mérés eredményének diagramja (4800 ejtés)

* 1 μGal=1 × $10^{-8}ms^{-2}$

3. Relative gravity measurements

The relative gravity measurements were performed with LCR gravimeters by researchers of the institutes listed on the front page together with those of the Geophysical Department of the Mining University of Leoben on the 24 ties shown in *Fig. 2*. In *Tables II* and *III* the observed Δg values for each gravimeter and their simple arithmetic mean are compiled. The average of the latter is

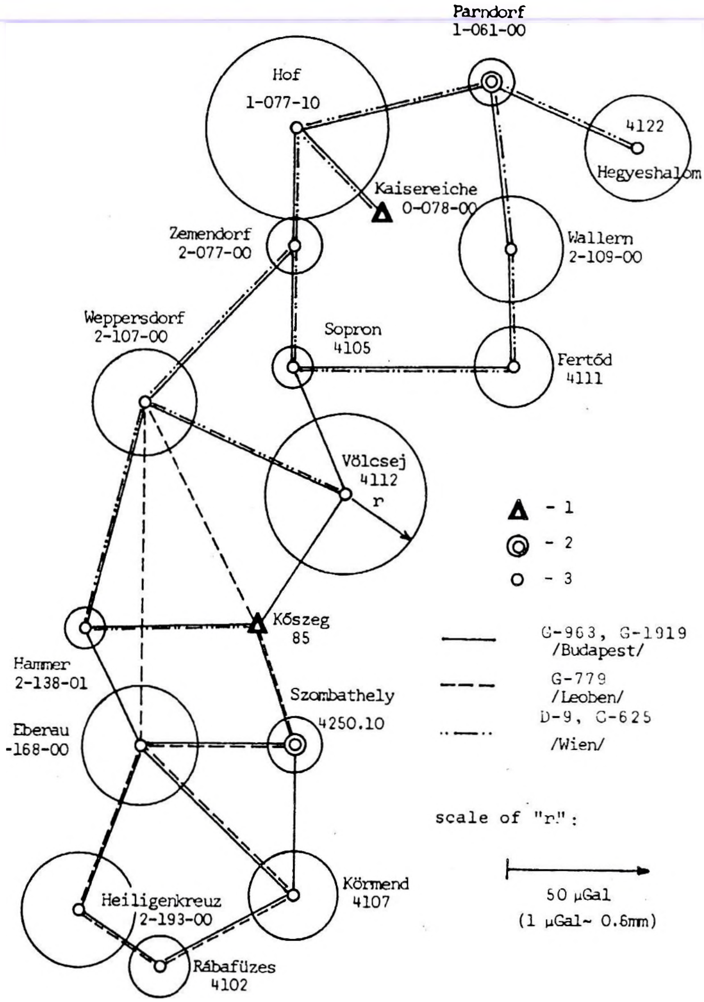


Fig. 2. Sketch of comparative measurements on the Austrian and the Hungarian gravity base networks and the 'error circles' of the measurements. 1—absolute station; 2—base point of the 1st order network; 3—base point of the 2nd order network

2. ábra. Az osztrák és a magyar gravimetriai alaphálózat összehasonlító méréseinek vázlata és a mérések „hibakörei". 1—abszolut állomás; 2—I. rendű bázispont; 3—II. rendű bázispont

TIE	GRAVIMETER				mean and error	
	D-9	G-625	G-963	G-1919	mGal	± μGal
KAISEREICHE - HOF	42.530	42.543	42.611	42.578	42.566	36
ZEMENDORF - HOF	37.173	37.196	37.196	37.179	37.186	12
HOF - PARNDORF	13.718	13.738	13.716	13.728	13.725	10
HEGYESHALOM - PARNDORF	7.223	7.251	7.240	7.264	7.245*	17
WALLERN - PARNDORF	18.463	18.482	18.440	18.440	18.456	20
FERTŐD - WALLERN	8.993	9.003	9.034	9.032	9.016	21
SOPRON - FERTŐD	15.864	15.892	15.877	15.871	15.876	12
ZEMENDORF - SOPRON	7.563	7.549	7.551	7.544	7.552	8
WEPPERSDORF - ZEMENDORF	24.519	24.512	24.545	24.527	24.526	14
WEPPERSDORF - VÖLCSEJ	25.919	25.911	25.960	25.913	25.926	23
HAMMER - WEPPERSDORF	3.915	3.906	3.895	3.920	3.909	11
HAMMER - KŐSZEG	12.341	12.333	12.313	12.326	12.328	12

* high seismic noise during the observations

Table II. Results of gravimetric measurements (northern part)
II. táblázat. A graviméteres mérések eredményei (északi rész)

±15 μGal. The ties were measured once in the order A-B-A-B-A or A-B-C-B-A-B-C-B-A (with the exception of the Sopron-Völcsej-Kőszeg part, which was observed twice with the instruments G-963 and G-1919). The readings of gravimeters G-625, G-779, G-963 and G-1919 were taken by CPI techniques, gravimeter D-9 was equipped with a feedback system. The results were reduced to the benchmark of each point and corrected for drift, Earth tides, barometric effect and scale factor. The scale factors were determined by comparison on national calibration lines. After the comparative measurements a calibration campaign was performed using the gravimeters of Wien and Budapest on the earlier established Göstling-Hochkar vertical calibration line [MEURERS, RUESS 1985]. Results of the measurements are compiled in *Table IV*, the calculated correction factors for scale constant in *Table V*.

TIE	GRAVIMETER			mean and error	
	G-779	G-969	G-1919	mGal	± μGal
EBERAU - WEPPERSDORF	11.001	—	—	11.001	
WEPPERSDORF - KŐSZEG	8.400	—	—	8.400	
SZOMBATHELY - KŐSZEG	8.565	8.545	8.561	8.557	11
KŐSZEG - VÖLCSEJ	—	17.517	17.482	17.500	25
VÖLCSEJ - SOPRON	—	6.147	6.175	6.161	20
EBERAU - SZOMBATHELY	10.860	10.831	10.844	10.845	15
EBERAU - HAMMER	—	7.076	7.067	7.072	6
KÖRMEND - EBERAU	19.532	19.578	19.548	19.553	23
KÖRMEND - SZOMBATHELY	—	30.415	30.419	30.417	2
KÖRMEND - RÁBAFÜZES	17.236	17.257	17.259	17.251	13
HEILIGENKREUZ - EBERAU	10.325	10.371	10.356	10.351	23
HEILIGENKREUZ - RÁBAFÜZES	8.030	8.056	8.055	8.047*	15

* high seismic noise during the observations

Table III. Results of gravimetric measurements (southern part)
III. táblázat. A graviméteres mérések eredményei (déli rész)

4. Adjustment of measurements

The adjustment of the network shown in Fig. 2. was carried out as a constrained network in three versions (*A, B, C*). In version '*A*' the absolute gravity value of Kaisereiche and Kőszeg measured with the JILAG-6 gravimeter was taken as a constraint. In version '*B*' in addition to the two absolute values, the points of HGN-80, while in version '*C*' apart from the absolute values, the points of AGBN were taken as constraints as well. All measurements were assumed to be of the same reliability, and the Δg values observed by each gravimeter were taken as independent measurements. The results of adjustment are contained in *Table VI.* The errors of the adjusted Δg values are 6-12 μGal, which — on account of the limited number of measurements — can be regarded as satisfactory. To give a better illustration of the quality of measurements, 'error circles' were plotted in Fig. 2. The radius (r) of each circle was calculated by the following relationship:

$$r_i = \sum v_i^2 / n_{vi}$$

where v_i — residuals belonging to point i obtained from adjustment 'A' in μGal,
n_{vi} — number of residuals belonging to point i.

TIE	known Δg (mGal)	gravimeter	observed Δg 1st day	2nd day	3rd day	mean (mGal)	error (mGal)	scale factor
HOCHKAR - AIBLBODEN	71.447	G-963	.447	.465	.443	71.451	± 11	0.999 944
		G-1919	.381*	.410	.413	71.401	23	1.000 644
HOCHKAR - LASSING	157.184		.107	.124	.129	157.120	13	1.000 407
			.136	.142	.147	157.141	12	1.000 274
AIBLBODEN - GÖSTLING	126.892		.804	.783	.803	126.797	14	1.000 749
			.863	.857	.855	126.859	5	1.000 260
LASSING - GÖSTLING	41.155		.145	.129	.121	41.132	13	1.000 559
			.112	.126	.119	41.119	8	1.000 876
HOCHKAR - GÖSTLING	198.339	**	.252	.253	.250	198.251		1.000 444
			.248	.268	.266	198.261		1.000 393

* gross error due to carelessness of the observer

** total gravity difference (calculated from the above four observed Δg)

Table IV. Results of measurements on the Göstling-Hochkar vertical calibration line
IV. táblázat. Hitelesítő mérések eredményei a Göstling-Hochkar vertikális bázison

gravimeter	correction factors for the scale constant 1	2*	3
G-963	1.001434	1.000415	1.002993
G-1919	1.000532	1.000513	1.001434

* calculated from mean value of four ties

Table V. Calibration factors of gravimeters G-963 and G-1919
1—Hungarian Gravimetric Calibration Line; 2—Göstling-Hochkar vertical calibration line; 3—adjustment, version '*A*'
V. táblázat. A G-963 és G-1919 graviméter méretaránytényezői
1— Magyar Gravimetriai Hitelesítő Alapvonal; 2—Göstling-Hochkar vertikális hitelesítő vonal; 3—„*A*" kiegyenlítési változat

AGBN and HGN-80 base points	known gravity (mGal) (-980 000)	adjusted gravity (G)			G_A-G_K	G_B-G_K	G_C-G_K
		A	B	C	(μGal)		
HOF	837.967	838.000	838.025		+ 33	+ 58	
PARNDORF	851.690	851.724	851.745		+ 34	+ 55	
ZEMENDORF	800.799	800.819	800.841		+ 20	+ 42	
WEPPERSDORF	776.279	776.296	776.298		+ 17	+ 19	
WALLERN	833.230	833.267	833.279		+ 37	+ 49	
HAMMER	772.366	772.385	772.385		+ 19	+ 19	
EBERAU	765.268	765.307	765.312		+ 39	+ 44	
HEILIGENKREUZ	754.927	754.955	754.949		+ 28	- 22	
HEGYESHALOM	844.486	844.451		844.446	- 35		- 40
FERTŐD	824.258	824.250		824.222	- 8		- 36
SOPRON	808.421	808.374		808.351	- 47		- 70
VÖLCSEJ	802.226	802.217		802.206	- 9		- 20
SZOMBATHELY	776.173	776.158		776.136	- 15		- 37
KÖRMEND	745.750	745.750		745.718	0		- 32
RÁBAFÜZES	762.995	763.001		762.971	+ 6		- 24
No of ties		24	24	24			
No of independent measurements		78	78	78			
No of unknowns		15	8	7			
M_0^* (μGal)		±17	±30	±21			

M_0^* = standard error of unit weight

Table VI. Main parameters of adjustments
VI. táblázat. A kiegyenlítések főbb paraméterei

5. Evaluation of the results

Based on the evidence of the error circles the '*g*' values of different points have different reliability. This is explained by the fact that the number of residuals changes from point to point and that from the statistical point of view the radius of the error circles is uncertain owing to the limited number of measurements. For several ties (e.g. Kaisereiche-Hof) differences exceeding the reliability of measurement were observed between the readings of different gravimeters. It can be seen from Table VI that both networks deviate to a small extent from the scale determined by the absolute measurements. This deviation is $-3.5 \times 10^{-5} \pm 1.7 \times 10^{-5}$ for the Austrian Base Network, and $2.5 \times 10^{-5} \pm 2.6 \times 10^{-5}$ for the Hungarian one. According to versions *B* and *C* the datum of HGN-80 is higher by about 40 μGal than the datum of AGBN. The difference can be explained by changing gravity or by a more trivial reason, an inaccurate g value of Sopron, which even at the 90 % probability level cannot be regarded as having the same reliability as those of the rest, i.e. the standard error of unit weight (M_0) of version B adjustment is significantly higher than those of the other two versions.

Our supposition, that the changes obtained at the reobserved absolute points are due to some other reason than instrument error, is based on the monotonously decreasing value of g obtained during reobservations carried out six times in nearly regular time intervals between 1980 and 1993 with different types of absolute gravimeters (GABL, JILAG, AXIS). The rate of decrease is 1.9 μGal/year during the investigated time interval (Table I).

To clarify the reason for the 40 μGal discrepancy, further investigations are needed.

In conclusion the following can be established:

1) From the viewpoint of plotting common gravity maps for the territories of the two countries the difference revealed has no practical importance;
2) Due to causes discussed in this paper, HGN-80 requires readjusment;
3) For joint gravity projects requiring high accuracy it is essential that the gravimeters be calibrated on the same calibration line.

REFERENCES

BOULANGER Yu. et al.1991: Results of the 3rd International Comparison of Absolute Gravimeters in Sevres 1989. BGI, Bull. d'Inf. **68**, 24-44. June 1991

CSAPÓ G., SÁRHIDAI A.: 1990: Adjustment of the New Hungarian Gravimetric Network (MGH-80). (in Hungarian) Geodézia és Kart. 1990/3. pp. 181-190, Budapest

MEURERS B., RUESS D. 1985: Errichtung einer neuen Gravimeter Eichlinie am Hochkar. ÖZfVuPh, **73**, 3, pp. 175-183

RUESS D., STEINHAUSER P., JERAM G., FALLER J. 1989: Neue Absolutschweremessungen in Österreich. Tagungsber 5. Int. Alpengravimetrie Kolloquium Graz, Österr. Beitr. zu Meteorol. und Geophysik, **2**, pp. 95-110

ÖSSZEKAPCSOLÓ GRAVITÁCIÓS MÉRÉSEK AUSZTRIA ÉS MAGYARORSZÁG GRAVIMETRIAI ALAPHÁLÓZATAI KÖZÖTT

CSAPÓ Géza, Bruno MEURERS, Diethard RUESS, SZATMÁRI Gábor

A dolgozatban a magyar és osztrák gravimetriai alaphálózatok 1991-93 között végzett összekapcsoló méréseiről számolnak be a szerzők. Ez a munka abszolut és relatív graviméteres méréseket tartalmazott. Az abszolut méréseket JILAG-6, a relatív méréseket 5 db LCR graviméterrel végezték 24 mérési kapcsolaton, a két ország teljes közös határszakaszán kiválasztott bázishálózati pontok között.

Megállapították, hogy a két ország alaphálózatának referenciaszintje 40 μGal-lal különbözik. Ezen eltérés okainak felderítése további vizsgálatokat, illetve a magyarországi alaphálózat újrakiegyenlítését igényli.

Strike oil
by advertising
with us

GEOPHYSICAL TRANSACTIONS OFFERS YOU ITS PAGES TO WIDEN THE SCOPE OF YOUR COMMERCIAL CONTACTS

Geophysical Transactions, contains indispensible information to decision makers of the geophysical industry. It is distributed to 45 countries in 5 continents.

Advertising rates (in USD)

	Page	Half page
Black and white	400/issue	250/issue
Colour	800/issue	450/issue

Series discount: 4 insertions — 20%

For further information, please contact:
Geophysical Transactions, Eötvös Loránd Geophysical Institute of Hungary

P.O.B. 35, Budapest, H—1440, Hungary
tel: (36-1) 163-2835 telex: 22-6194
fax: (36-1) 163-7256