ACTA TECHNICA

ACADEMIAE SCIENTIARUM HUNGARICAE

EDITOR: M. MAJOR

IN MEMORY OF PROF. Á. KÉZDI

VOLUME 98
NUMBERS 1—2

AKADÉMIAI KIADÓ, BUDAPEST 1985

ACTA TECHN. HUNG.

ACTA TECHNICA

A JOURNAL OF THE HUNGARIAN ACADEMY OF SCIENCES

Acta Technica publishes original papers, preliminary reports and reviews in English, which contribute to the advancement of engineering sciences.
Acta Technica is published by

AKADÉMIAI KIADÓ
Publishing House of the Hungarian Academy of Sciences
H-1450 Budapest, Alkotmány u. 21.

Subscription information

Orders should be addressed to

KULTURA Foreign Trading Company
H-1389 Budapest P.O. Box 149

or to its representatives abroad

Acta Technica is indexed in Current Contents

ACTA TECHNICA
Volume 98 Nos 1—4

CONTENTS

BOOK REVIEWS

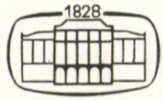
CONTENTS

PRINTED IN HUNGARY
Akadémiai Kiadó és Nyomda, Budapest

ACTA TECHNICA

ACADEMIAE SCIENTIARUM HUNGARICAE

VOLUME 98



AKADÉMIAI KIADÓ, BUDAPEST 1985

Acta Technica Academiae Scientiarum Hungaricae, 98 (1—2), pp. 5—7 (1985)

LIFE AND WORK OF PROFESSOR ÁRPÁD KÉZDI

This volume was meant to be handed over to Prof. Á. Kézdi on occasion of his 65th birthday, at the 6th Conference on Soil Mechanics and Foundation in Budapest, first week of October 1984. Unfortunately and much to the sadness of his family, friends and admirers, this touching ceremonial act could not take place. Prof. Á. Kézdi, Hungarian scientist of international reputation in technical sciences, departed in the sixty-fifth year of his life.

At the 6th Conference on Soil Mechanics and Foundation in Budapest, many foreign scientists and friends of the first rank in this special field commemorated Professor Kézdi, remembering him as an outstanding personality, a colleague who had gained distinction in his profession.

In recent years the outstanding scientific achievements and extremely rich career of Professor Kézdi have been praised in quite a number of Hungarian and foreign periodicals.

A devoted and highly efficient scientist, he paid distinctive attention to the analysis, and working out, of theoretical problems of soil mechanics significant also in respect of practical application. He never considered the results of research to be unquestionable facts, professing that any theoretical method or calculation process would lose validity beyond a certain limit. He was hundred per cent a scientist, combining superior talent with a will hard as steel, and with fascinating stamina.

Deeply absorbed in work, he systematized and developed the national and international geotechnical results, and added new results. The scope of research he was involved in was wide. He made his mark as a scientist in the fields of soil physics, earth dynamics, and pile load capacity alike. The results of his scientific work have been published in 44 books and 150 papers by both home and foreign editors.

A book titled "Soil Mechanics", a masterpiece in 4 volumes, has been a success both in Hungary and internationally. Published in Hungarian, German, English, Spanish, and Russian language between 1969 and 1979, it is still used as a university text-book in several countries. The first book to bring him international reputation was "Erddrucktheorien" published by Springer Verlag, Berlin in 1962. Another book

titled "Stabilized Earth Roads" has been published in Hungarian, German, and English while his book "Soil Physics" in German, English, and Spanish.

He never got tired of making the Hungarian soil mechanics internationally appreciated. His part in the activities of the International Soil Mechanics and Foundation Association was decisive for more than a quarter of a century. Many times he attended the conferences of the International Association as chairman, general report lecturer, or author. His opinion was considered decisive in disputes concerning terminology at the meetings of the Executive Committee of the Association. In the period between 1973 and 1977, Professor Kézdi held the office of vice-chairman of the Association. He was invited to deliver lectures in different parts of the world, and his firm knowledge earned international appreciation for Hungarian soil mechanics in many countries.

However, it was not only he who travelled throughout the world. Upon, or without invitation, visitors of a large number came to meet Professor Kézdi so that the Geotechnical Department headed by him as well as the Conferences on Soil Mechanics and Foundation also initiated by him became a meeting-place of internationally appreciated scientists in the field of soil mechanics.

The large number of invitations Professor Kézdi received as an expert from the United States, GFR, Italy, Spain, Jugoslavia and other countries reflect his international reputation. Also, he contributed to the successful solution of the geotechnical problems of almost all the most important projects in this country in the course of the past 35 years.

In addition to research work which was the basis of Professor Kézdi's international reputation, he offered maximum also in the university chair, a passionate teacher to whom university and teaching meant life itself. After his graduation in 1942, he spent his hard-working life at the Geotechnical Department of the Technical University Budapest. In 1950 he became head of the Department and stayed in this post for more than 30 years.

His ambition was to offer the best also in education, and he made enormous efforts to comply with the requirements imposed upon himself even after his health had been impaired.

In addition to education, his role in public life both at the university and in Hungarian science was significant.

Professor Kézdi was Vice-Rector of the university for two terms. He held the office of vice-chairman of the Scientific Association of Transportation, MTESZ, for years. He was office-holder of a number of international associations as well as member of several national scientific or editorial committees.

In appreciation of his outstanding work in the field of both science and education, Professor Kézdi was awarded the State Prize in 1966 and he was elected corresponding member of the Hungarian Academy of Sciences in 1970, and member in 1976.

Professor Kézdi was Honorary Doctor of Technische Universität Dresden, Hochschule für Bodenkultur Wien, and Honorary Professor of Lima and Ica universities, Peru. Holder of several orders, he was awarded the Order for Socialist Hungary on the occasion of his retirement in 1983. All this appreciation encouraged him to embark upon new and increasingly hard tasks.

Professor Árpád Kézdi was not only passionately interested in his profession but at the same time had a wide-ranging classical education and was an enthusiast of classical music.

It is not only reverence but also the respect for our devoted predecessors working with outstanding efficiency that has encouraged us to compile this memorial issue to express our gratitude for the rich professional heritage left to us, which we cherish, convey to those coming after us, and what we have learned we continue developing.

The spirit devoted to science and higher education remains our model in professional life.

G. PETRASOVITS

Acta Technica Academiae Scientiarum Hungaricae, 98 (1—2), pp. 9—21 (1985)

PROFESSOR Á. KÉZDI'S PUBLICATIONS AND THEIR CRITICAL REVIEWS

I. Books

Kézdi Árpád:
Cementtalaj utak vizsgálata és méretezése. Közlekedési Kiadó, Budapest 1951

Kézdi Árpád:
Talajmechanika I. Tankönyvkiadó, Budapest 1952

Kézdi Árpád:
Talajmechanika II. Tankönyvkiadó, Budapest 1954

Kézdi Árpád (Dr. Póczy Mihály):
Földművek I–II. (Chapters A and B in Volume I, Chapter E in Volume II.) Tankönyvkiadó, Budapest 1957

Kézdi Árpád:
Talajmechanika. Földművek (in: "Mérnöki Kézikönyv", Editor Dr. Palotás László) (Chapters II and IV in Volume 2) Műszaki Könyvkiadó, Budapest 1957

Kézdi Árpád:
Talajmechanika I. 2. Enlarged edition. Tankönyvkiadó, Budapest 1959

Kézdi Árpád:
Chapters 7. Mérnöki biológia (102–115), 12. Talajmechanikai alapfogalmak (153–164), 13. Műszaki földtani vizsgálatok menete és terjedelme (165–172), 17. Az alapozás földtana (206–228), 20. A létesítmények épségét veszélyeztető tényezők. In Mosonyi E.–Papp F.: "Műszaki Földtan", Műszaki Könyvkiadó, Budapest 1959

Kézdi Árpád—(Széchy Károly):
Alagutak, alapozás, földművek, talajmechanika. Műszaki Értelmező Szótár, 10. Terra, Budapest 1960

Kézdi Árpád:
Talajmechanikai praktikum. Tankönyvkiadó, Budapest 1961

Kézdi Árpád—(Markó Iván):
Földművek védelme és víztelenítése (Volume 1) Műszaki Könyvkiadó, Budapest 1962

Kézdi Árpád:
Talajmechanikai alapfogalmak. Műszaki földtani munkák a felszínen. In: „Bányászati Kézikönyv" (Editor-in-chief Boldizsár Tibor). Sections B and C in chapter 3 in Volume III. Műszaki Könyvkiadó, Budapest 1962

Kézdi Árpád:
Erddrucktheorien. Springer-Verlag, Berlin–Göttingen–Heidelberg 1962

Kézdi Árpád—(Markó Iván):
Földművek védelme és víztelenítése. Volume 2. Műszaki Könyvkiadó, Budapest 1964

Kézdi Árpád:
Bodenmechanik I–II. Verlag der ungarischen Akademie der Wissenschaften, Budapest, VEB Verlag für Bauwesen, Berlin 1964

Kézdi Árpád:
Stabilizált földutak. Akadémiai Kiadó, Budapest 1967

Kézdi Árpád:
Handbuch der Bodenmechanik. Bd.1, Bodenphysik. Akadémiai Kiadó, Budapest, VEB Verlag für Bauwesen, Berlin 1968

Kézdi Árpád:
Talajmechanika I. 3rd enlarged edition, Tankönyvkiadó, Budapest 1969

Kézdi Árpád—(Markó Iván):
Erdbauten. Schutz und Entwässerung. Werner–Verlag, Düsseldorf 1969

Kézdi Árpád:
Talajmechanika II. 2nd edition. Tankönyvkiadó, Budapest 1970

Kézdi Árpád:
Handbuch der Bodenmechanik Bd.2. Bodenmechanik im Erd–, Grund– und Strassenbau. Akadémiai Kiadó, Budapest, VEB Verlag für Bauwesen, Berlin 1970

Kézdi Árpád (edited by):
Proceedings of the 4th Budapest Conference on Soil Mechanics and Foundation Engineering (3rd Danube European Conference). Akadémiai Kiadó, Budapest 1971

Kézdi Árpád:
Talajmechanika I. 4th edition. Reprint of the 3rd enlarged edition. Tankönyvkiadó, Budapest 1972

Kézdi Árpád:
Handbuch der Bodenmechanik, Band 3. Bodenmechanisches Versuchwesen. Akadémiai Kiadó, Budapest, VEB Verlag für Bauwesen, Berlin 1973

Kézdi Árpád:
Stabilisierte Erdstrassen. Akadémiai Kiadó, Budapest, VEB Verlag für Bauwesen, Berlin 1973

Kézdi Árpád:
Handbook of Soil Mechanics. Vol. 1. Soil Physics. Akadémiai Kiadó, Budapest, Elsevier Scientific Publishing Co.Amsterdam 1974

Kézdi Árpád—(Markó Iván):
Földművek. Víztelenítés. Műszaki Könyvkiadó, Budapest 1974

Kézdi Árpád:
Chapter 5: Lateral earth pressure. Chapter 19: Pile foundations. In: Winterkorn, H. F., Fang, H. Y.: Foundation Engineering, Handbook Van Nostrand Reinhold Company, New York /Cincinnatti/ Toronto /London/ Melbourne 1975

Kézdi Árpád:
Manual de la Mecanica de Suelos Tomos Fisica del Suelo. Traduccion: Andres Pesti y Juan C. Hiedre López. Universidad Central de Venezuela. Ediciones de la Biblioteca, Caracas 1975

Kézdi Árpád:
Talajmechanika II. 3rd edition. Reprint of the 2nd enlarged edition. Tankönyvkiadó, Budapest 1975

Kézdi Árpád:
Fragen der Bodenphysik. Akadémiai Kiadó, Budapest, VDI–Verlag, Düsseldorf 1976

Kézdi Árpád:
Handbuch der Bodenmechanik. Band 4. Anwendung der Bodenmechanik in der Praxis. Akadémiai Kiadó, Budapest, VEB Verlag für Bauwesen, Berlin 1976

Kézdi Árpád:
Talajmechanika. Példák és esettanulmányok. Tankönyvkiadó, Budapest 1976

Kézdi Árpád—(Lazányi I.):
Proceedings of the Fifth Budapest Conference on Soil Mechanics and Foundation Engineering. Akadémiai Kiadó, Budapest 1976

Kézdi Árpád:
Talajmechanikai praktikum. 3rd enlarged edition. Tankönyvkiadó, Budapest 1976

Kézdi Árpád:
Talajmechanika I. 5th edition. Reprint of the 4th edition. Tankönyvkiadó, Budapest 1977

Kézdi Árpád:
Soil Physics. Selected Topics. Akadémiai Kiadó, Budapest 1979. (with Elsevier Scientific Publishing Company, Amsterdam)

Kézdi Árpád:
Stabilized Earth Roads. Akadémiai Kiadó, Budapest 1979. (with Elsevier Scientific Publishing Company, Amsterdam).

Kézdi Árpád:
Talajmechanika II. 4th edition. Reprint of the 3rd edition. Tankönyvkiadó, Budapest 1979

Kézdi Árpád:
Talajmechanika. Példák és esettanulmányok. 2nd edition. Reprint of the 1st edition. Tankönyvkiadó, Budapest 1979

Kézdi Árpád:
Handbook of Soil Mechanics. Volume 2, Soil Testing. Akadémiai Kiadó, Budapest 1980. (Elsevier Scientific Publishing Company Amsterdam.)

Kézdi Árpád (herausgegeben v.):
Bodenmechanik in der Sowjetunion. Akadémiai Kiadó, Budapest, VDI Verlag GmbH, Düsseldorf 1981

II. Papers

Kalcher Árpád:
Szádfalak grafikus méretezése. Vízügyi Közlemények. Budapest 1943
Kalcher Árpád:
Mérnöki biológia. Technika, Budapest 1944
Kalcher Árpád:
Közlekedési vonalak távlati ábrázolása. Technika, Budapest 1944
Kalcher Árpád:
Újabb kutatások cölöpök teherbírásának meghatározására. Magyar Technika–Általános mérnök, Vol. II, No. 1, Budapest 1947
Kézdi (Kalcher) Árpád:
A talaj elektromos jelenségei és technikai alkalmazásuk. Vízügyi Közlemények, 1947, No. 1–4 Budapest
Kézdi Árpád (Jáky József):
Az újjáépülő szegedi közúti Tisza-híd altalajvizsgálata. Magyar Technika–Általános Mérnök, Vol. III, No. 7, Budapest 1948
Kézdi Árpád:
Épületek alapozása homokcölöpökkel. Mélyépítéstudományi Szemle, Vol. 1, No. 1, Budapest 1951
Kézdi Árpád:
Homoktalajok tömörségének gyors meghatározása. Mélyépítéstudományi Szemle, Vol. 1, No. 7, Budapest 1951
Kézdi Árpád:
Einige Probleme der Spannungsverteilung im Boden. Acta Techn. Hung. Vol. II. Budapest 1951
Kézdi Árpád:
Talajmechanikai kérdések Sztálinvárosban. Építés-Építészet. Vol. III. No. 11–12, Budapest 1951
Kézdi Árpád:
A Balaton északkeleti peremén bekövetkező mozgások vizsgálata. Hidrológiai Közlöny. Vol. 32, No. 11–12, Budapest 1952
Kézdi Árpád:
Van-e zavartalan talajminta? Mélyépítéstudományi Szemle, Vol. 3, No. 1, Budapest 1953
Kézdi Árpád:
Makroporózus talajok vizsgálata roskadás szempontjából. M.Tud.Akadémia Műsz. Tud.Oszt.Közleményei. Vol. XII. No. 1–4 pp. 191–200, Budapest 1954
Kézdi Árpád:
Tömörítés = minőségi földmunka. Mélyépítéstudományi Szemle. Vol. 3, No. 11–12, Budapest 1953
Kézdi Árpád:
A feltöltési anyagok vizsgálata talajfizikai alapozási és földműépítési szempontból. Mélyépítéstudományi Szemle, Vol. 5, No. 9, Budapest 1955

Kézdi Árpád:
Über die Tragfähigkeit und Setzung von Pfahlgründungen. In: Gedenkbuch für J.Jáky. Akadémiai Kiadó, Budapest 1955

Kézdi Árpád:
Kísérleti cement–talaj utak építése és kipróbálása. Az Építőipari Műszaki Egyetem Tudományos Közleményei, Vol. I, No. 1, Budapest 1955

Kézdi Árpád:
Soil Mechanics. Feature article; Applied Mechanics Reviews, Vol 8, 1955, New York N.Y.

Kézdi Árpád:
Rézsűk állékonysága. Vízügyi Közlemények, 1956, No. 1, Budapest 1956

Kézdi Árpád:
Dovolená zatizeni a ssedáni základu. Inzenyrské Stavby, Vol. 4, No. 10, Praha 1956

Kézdi Árpád:
Bearing Capacity of Piles and Pile Groups. Proc. 4th Int.Conf.Soil Mech. Found. Engg.Vol.II. Butterworths Scientific Publications, London 1957

Kézdi Árpád:
Földművek állékonysága. Az Építőipari és Közlekedési Műszaki Egyetem 1955. évi tudományos ülésszakának előadásai. Tankönyvkiadó, Budapest 1957

Kézdi Árpád:
Erfahrungen mit der Zement-Bodenvermörtelung in Ungarn. Strassen- und Tiefbau. Vol. 9, No. 9, Heidelberg 1957

Kézdi Árpád:
Cementtalajutak tartóssága. Mélyépítéstudományi Szemle, Vol. 7, No. 7–8, Budapest 1957

Кезди, А.:
Несущая способность свай. Основания и фундаменты, Москва 1957

Kézdi Árpád:
Earth Pressure on Stiff Retaining Wall, Tilting about the Toe. Brussels Conference 58 on Earth Pressure Problems; Proceedings, Vol. I, Bruxelles 1958

Kézdi Árpád:
Cinq ans de mécanique du sol en Hongrie. Annales de l'Institut Technique du Bâtiment et des Travaux Publics. Juillet-Août, 1958. Onzième année, No. 127–128

Kézdi Árpád:
Vplyvy posobiace na stabilitu svahov. Stavebnicky Časopis, Vol. VI., No. 1. Slovenska Akademia Vied, Bratislava 1958

Kézdi Árpád:
Einiges über Rutschungen im Strassenbau. Strassen- und Tiefbau, Vol. 10, No. 3, Heidelberg 1958

Kézdi Árpád:
Beiträge zur Berechnung der Spannungsverteilung im Boden. Der Bauingenieur, Vol. 33(1958). No. 2

Kézdi Árpád:
Megjegyzések rézsűk állékonyságának vizsgálatához. Építés- és Közlekedéstudományi Közlemények. No. 3–4, Budapest 1959

Kézdi Árpád:
Cölöpök és cölöpcsoportok teherbírása. Építőipari és Közlekedési Műszaki Egyetem Tud.Közl. Vol. IV, No. 3, Budapest 1959

Kézdi Árpád:
Earth Pressure on Retaining Wall Tilting about the Toe. Acta Techn. Hung. Tom. XXV, No. 3–4, Budapest 1959

Kézdi Árpád—(Sándor I.):
Lösung der Differentialgleichung der eindimensionalen Konsolidation mittels Matrissenkalküls. Acta Techn. Hung. Tom. XXVII, No. 3–4, Budapest 1960

Kézdi Árpád:
Contributions to the bearing capacity of piles. Acta Techn. Hung. Tom. XXIX, No. 3–4, Budapest 1960

Kézdi Árpád:
Nekolika pitanja, praktične mehanike tla i fundiranja. Technika, Vol. XV, No. 12, Nase Gradevinarstvo, Beograd 1960

Kézdi Árpád:
Bemerkungen zur Frage der Tragfähigkeit von Pfahlgruppen. Symposium on Pile Foundations. Stockholm 1960

Kézdi Árpád:
Untersuchung einiger Grundbruchfälle. Vortäge der Baugrundtagung 1960 in Frankfurt am Main. Deutsche Gesellschaft für Erd- u. Grundbau, Hamburg 1961

Кезди, А.:
Опыт Строительства на лессовых грунтах в Венгрии. В "Вопросы строительства на лессовых грунтах", Воронеж 1961

Kézdi Árpád:
A talajmechanika alkalmazásai a mérnöki gyakorlatban. Közlekedéstudományi Egyesület, Budapest 1961

Kézdi Árpád:
Cölöpök teherbírása. Közlekedéstudományi Egyesület, Budapest 1961

Kézdi Árpád:
The Effect of Inclined Loads on the Stability of a Fundation. Proc. 5th Int. Conf. Soil Mech. Found. Engg. Vol. 1, Paris 1961

Kézdi Árpád:
Útburkolatok alatti szűrőrétegek viselkedésének vizsgálata. Építés- és Közlekedéstudományi Közlemények, Vol. 3. No. 4, Budapest 1962

Kézdi Árpád:
Einige Betrachtungen zur Untersuchung der Standsicherheit von Böschungen. Bauplanung–Bautechnik. Vol. 17, No. 2, 1963

Kézdi Árpád:
Scherverformungen von Sand. Az Építőipari és Közlekedési Műszaki Egyetem Tudományos Közleményei Vol. XI No. 5, Budapest 1963

Kézdi Árpád:
Semleges feszültség és áramlási nyomás. Vízügyi Közlemények, Vol. 1963, No. 1, Budapest 1963

Kézdi Árpád:
Setzungen im Löss infolge der Erhöhung des Grundwasserspiegels. Proceedings, "Europäische Baugrundtagung", Wiesbaden 1963

Kézdi Árpád:
Über Bodenstabilisierung im Strassenbau. Die Strasse, Vol. 1963, No. 3, Berlin 1963

Kézdi Árpád:
Diskussionsbeitrag anlässlich des Seminars über Bodenstabilisierung in Linz, 1963. Mitteilungsblatt der Forschungsgesellschaft für das Strassenwesen im Ö.I.A.V. Österreichische Ingenieur-Zeitschrift, Vol. 7, 1964

Kézdi Árpád:
Lectures on Soil Mechanics. Publication of the School of Engineering, Princeton University, Princeton, N. J 1964

Kézdi Árpád:
Some properties of packings. In: "Mechanical and Physico-Chemical Properties of Soils." Highway Research Record, No. 52. Highway Research Board publication 1177. Washington, D. C 1964

Kézdi Árpád:
Earth Pressure Theories. Soil Mechanics Lecture Series: Design of Structure to Resist Earth Pressures. Sponsored by the Soil Mechanics and Foundation Division, Illinois Section, Am. Soc. Civ. Engrs. and Civ. Engg. Dept. Ill. Inst. of Technology. Chicago, Illinois 1964

Kézdi Árpád:
Egy új talajfizika alapjai. Budapest 1964

Kézdi Árpád:
Stresses around Wellbores. Lab. Mem. No. LM 65. Socony Mobil Co. Field Research Station, Dallas, Texas 1965

Kézdi Árpád:
Lectures in Soil Mechanics Lab. Mem. No. LM 65. Socony Mobil Co. Field Research Station. Dallas, Texas 1965

Kézdi Árpád:
Problems of Sand Control in the Oil Industry. Lab. Mem. No. LM 65. Socony Mobil Co. Field Research Station. Dallas, Texas 1965

Kézdi Árpád:
On some properties of soil mixtures. Proceedings, American Society of Civil Engineers, Journal of the Soil Mechanics and Foundations Division. Vol. 91, No. SM 4, New York 1965

Кезди, А.:
Некоторые вопросы исследования устойчивости склонов при разработке угля. Известия Академии Наук Армянской СССР. Ереван 1965

Kézdi Árpád:
Az Oroville-gát. Vízügyi Közlemények, Vol. 1966, No. 2, Budapest 1966

Kézdi Árpád:
Soil Mechanics. In: Applied Mechanics Surveys; (Edited by H. Norman Abramson, Harold Liebowitz, John M. Crowley, Stephen Juhász. Spartan Books.) Washington, D. C 1966

Kézdi Árpád—(Brahma, S. P.):
Strength of Soil-Cement. Paper presented in Madras, India. Published in Indsearch, Madras 1966

Kézdi Árpád:
Új eredmények a talajfizikában. Mélyépítéstudományi Szemle, Vol. XVI, No. 6, Budapest 1966

Kézdi Árpád:
Szemcsés talajok nyírószilárdsága. Mélyépítéstudományi Szemle. Vol. 16, No. 8, Budapest 1966

Kézdi Árpád:
Contributions to the Investigations of Granular Systems. In: Kravtchenko, J.–Sirieys, P. M.(Editors): Rheology and Soil Mechanics. Symposium Grenoble, 1964. International Union of Theoretical and Applied Mechanics, Springer–Verlag, Berlin/Heidelberg/New York 1966

Kézdi Árpád:
Egy újfajta földmegtámasztásról. Mélyépítéstudományi Szemle. Vol. 16, No. 12, Budapest 1966

Kézdi Árpád:
Gyorsvasút a San Francisco-i öböl körül. Közlekedéstudományi Szemle. Vol. XVI, No. 4, Budapest 1966

Kézdi Árpád:
Grundlagen einer allgemeinen Bodenphysik. VDI-Zeitschrift, Düsseldorf 1966

Kézdi Árpád:
The phenomena of suffusion and their application to well hydraulic. Transactions, American Geophysical Union, 1966

Kézdi Árpád:
A talajstabilizáció néhány fizikai és kémiai vonatkozása. Építés- és Közlekedéstudományi Közlemények. Vol. XI, No. 2, Budapest 1967

Kézdi Árpád:
Kohéziós talajok nyírószilárdsága. Mélyépítéstudományi Szemle, Vol. 17, No. 1, Budapest 1967

Kézdi Árpád:
Bodenphysikalische Untersuchungen VDI-Zeitschrift Vol. 109, No. 23, Düsseldorf 1967

Kézdi Árpád:
Új könyvek földnyomásról, támfalakról. Mélyépítéstudományi Szemle, Vol. 17, No. 6, Budapest 1967

Kézdi Árpád:
Plastizitätslehre von körnigen Materialen. Acta Techn. Hung. Vol. 59, No. 1–2, Budapest 1967

Kézdi Árpád:
Külszíni szénfejtés hányóinak állékonysága. Földmunkák gépesítése. 6th International Conference on Mechanization of Earthwork. Contributions, Vol. II, MNK-4. Budapest 1967

Kézdi Árpád:
Statecznosé zwalowisk. IV. Ogolnopolska Konferencja Mechaniki Gruntów; Fundamentowania. Naczelna Organizacja Techniczna w Polsce; Komitet Geotechniki i Robot Podziemnych. Wrocław 1967

Kézdi Árpád:
Külszíni szénfejtések talajmechanikai vonatkozásai. Az Építőipari és Közlekedési Műszaki Egyetem Tudományos KözlÓeményei Vol. XIII, No. 5, Budapest 1967

Kézdi Árpád:
Festigkeit von stabilisierten Erdstoffen. Donau-Europäsche Konferenz "Bodenmechanik im Strassenbau". Wien, 8 bis 10. Mai. 1968. Eigenverlag d.Österreichischen Ingenieur– und Architekten–Vereines, Wien 1968

Kézdi Árpád:
Bodenmechanische Probleme im Tagebau. Teil 1. Bodenphysikalische Untersuchungen für Tagebaue. Bergbautechnik. Vol. 18, No. 6, Juni, 1968 Leipzig

Кезди, А.:
О характеристике физического состояния грунтов. Основания, фундаменты и механика грунтов, Москва 1968

Kézdi Árpád:
Bodenmechanische Probleme im Tagebau. Teil 2. Standsicherheitsuntersuchungen an Kippen. Bergbautechnik. Vol. 18, No. 8, Leipzig 1968

Kézdi Árpád:
Nuevos adelantos en la fisica del suelo. Boletin Sociedad Venezolana de Mecanica del Suelo e Ingenieria de Fundaciones. Diciembre 1967, Marzo 1968, No. 25–26

Kézdi Árpád:
Distributions of grains and voids according to their volume. Acta Techn. Hung. Vol. 63, No. 4

Kézdi Árpád—(Nagyváti B.):
La resistencia de los suelos estabilizados. Boletin Sociedad Venezolana de Mecanica del Suelo e Ingenieria de Fundaciones Julio–Diciembre 1968. No. 28–29

Kézdi Árpád—(Nagyváti B.):
Strength of stabilized Soils. Acta Techn. Hung. Vol. 62, No. 1–2, Budapest 1968

Kézdi Árpád:
Gátépítés az Eider-folyó torkolatában. Vízügyi Közlemények, Vol. 1969, No. 2

Kézdi Árpád:
Landslide in Loess along the Bank of the Danube. Proceedings, 7th Int. Conf. Soil Mech. Found. Engg. Vol. 2. Mexico City 1969

Kézdi Árpád:
Homokrézsűk állékonysága. Bányászati Lapok, Vol. 101, No. 5 1969

Kézdi Árpád:
Recent Developments in Filtration Calculation. Mobil, Research and Development Corporation, Spec. Report; Dallas, Texas 1969

Kézdi Árpád:
A rugalmasságtan alkalmazása kétfázisú közegre. In: "Szilárdságtani Kollokvium". A MTA Műsz. Tud. Oszt. Elméleti Mechanikai Bizottsága és Tartószerkezetek Mechanikája Bizottsága, Budapest 1969

Kézdi Árpád:
Stability and some Control Problems. Mobil Research and Development Corporation, Spec. Report, Dallas, Texas 1969

Kézdi Árpád:
Increasing the Bearing Capacity of Floating Piles. Speciality Session 8, 7th Int. Conf. Soil Mech. and Found Engg. Buenos Aires 1969

Kézdi Árpád:
Szemcsés közegek fizikájának szerepe az építőmérnöki mechanikában. In: Budapesti Műszaki Egyetem Jubileumi Tudományos Ülésszak, 1970, Építőmérnöki Kar, Budapest 1970

Kézdi Árpád:
Aumento de la capacidad de carga de pilotes flotantes. Facultad de Ingenieria, Universidad Nacional de Buenos Aires. Agosto 1969. Spec. Session 8, 7th Int. Conf. Soil Mech. Found. Engg.

Kézdi Árpád:
Increasing the bearing capacity of floating piles: settlement observations on a large silo on piled foundations. Proceedings, Behaviour of Piles Conference, The Institution of Civil Engineers, London 1970

Kézdi Árpád—(Paál Tamás—Pálfy Lajos):
Az Apostol utcai suvadás vizsgálata és helyreállítása. Mélyépítéstudományi Szemle, Vol. XX, No. 1, 1970

Kézdi Árpád:
A dunaújvárosi partrogyás. Mélyépítéstudományi Szemle. Vol. XX, No. 7, 1970

Kézdi Árpád—(Nagyváti B.):
Einfluss von Zusatzmitteln auf die Eigenschaften von stabilisierten Böden. Acta Techn. Hung. Vol. 68

Kézdi Árpád:
Spannungen in Zweiphasensystemen. Acta Techn. Hung. Vol. 69, 1970

Kézdi Árpád:
Néhány szó a Hold talajáról. Mélyépítéstudományi Szemle. Vol. 20, 1970

Kézdi Árpád:
Earth pressure measurements. In: Nové poznathy v mechanice zemin. (New advences in Soil Mechanics.) Vol. I. Československá Védecko technická Společnost Stavebni, Praha 1971.

Kézdi Árpád—(Varga L.—Timár A.):
Strength of transition soils. Proc. 4th Budapest Conference on Soil Mechanics and Foundation Engineering. Akadémiai Kiadó, Budapest 1971

Kézdi Árpád—(Marczal L.—Biczók E.—Kabai I.):
Behaviour of transition soils under the effect of water. Proc. 4th Budapest Conference on Soil Mechanics and Foundation Engineering. Akadémiai Kiadó, Budapest 1971

Kézdi Árpád—(Lazányi I.—Kabai I.):
Compaction of transition soils. Proc. 4th Budapest Conference on Soil Mechanics and Foundation Engineering. Akadémiai Kiadó, Budapest 1971

Kézdi Árpád:
Várható fejlődés az alapozások elméleti kérdéseiben. ÉTE Mérnöki létesítményi és Közműépítési Szakosztály, "Alapozás" 1971

Kézdi Árpád:
Phasenzusammensetzung bei stabilisierten Erdstoffen. Bauplanung-Bautechnik, Berlin. Vol. 26, 1972. (Vorabdruck aus Handbuch der Bodenmechanik Bd. 3.)

Kézdi Árpád:
Egy siló alapozásának története. Mélyépítéstudományi Szemle. Vol. XXII, No. 3, Budapest 1972

Kézdi Árpád:
Earth pressure measurements. Proceedings 5th European Conference on Soil Mechanics and Foundation Engineering, Vol. I, Madrid 1972

Kézdi Árpád—(Bögös Mária):
Szivárgás ellen védő müanyag fólia védőrétegének vizsgálata. Mélyépítéstudományi Szemle. Vol. 22, No. 8, 1972

Kézdi Árpád—(Horváth Gy.):
Kötött talajok húzó és hajlító szilárdsága. Mélyépítéstudományi Szemle, Vol. 22, No. 7, 1972

Kézdi Árpád:
Tensile and flexural strength of earth dam materials. Comptes Rendus, Onzième Congres des Grands Barrages, Q. 42. R. 10. Vol. 2, Madrid 1973

Kézdi Árpád:
Die Bedeutung der bodenphysikalischen Forschung für den Erdbau. 100-Jahrfeier Hochschule für Bodenkultur; Vorträge der Studienrichtung Kulturtechnik und Wasserwirtschaft, 18 und 19. Oktober 1972. Herausgegeben von der Studienrichtung Kulturtechnik und Wasserwirtschaft an der Hochschule für Bodenkultur in Wien. Band V, Teil 2. Wien 1973

Kézdi Árpád—(Horváth György):
Tensile and flexural strength of cohesive soils. Acta Techn. Hung. Vol. 74, Budapest 1973

Kézdi Árpád:
Bestimmung des Durchlässigkeitsbeiwertes und der kapillaren Steighöhe in einem Versuch. Bauplanung-Bautechnik Vol. 27, No. 2, 1973. Berlin, Verlag für Bauwesen. (Vorabdruck, Hdb. der Bodenmechanik Vol. 3.)

Kézdi Árpád—(Kabai I.—Biczók E.):
Mintavétel kötött talajokból. Mélyépítéstudományi Szemle. Vol. XXIII, No. 3, 1973

Kézdi Árpád—(Vidacs L.—Jancsecz S.):
Lágy talajon épült hazai töltések süllyedése. Mélyépítéstudományi Szemle, Vol. 24, No. 4, 1974

Kézdi Árpád—(Biczók E.—Horváth Gy.):
Vízmozgás homokban. A MTA Műszaki Mechanikai Tanszéki Munkaközösségének tudományos ülésszaka. Budapest, 1974. IV. Szemcsés közegek mechanikája

Kézdi Árpád—(Kabai I.—Biczók E.—Marczal L.):
Sampling cohesive soils. Periodica Polytechnica Civil Engineering. Vol. 18, No. 4, 1974

Kézdi Árpád—(Biczók E.—Nagy A.):
Lejtőkúszás mérése. Mélyépítéstudományi Szemle, Vol. 25, No. 6, 1975

Kézdi Árpád:
A talajmechanika szerepe az építőipari feladatok megoldásában. Műszaki Tervezés, Vol. XV, No. 7, Budapest 1975

Kézdi Árpád:
Vote of thanks, to Prof. Kérisel, after having given the 15th Rankine lecture. Geotechnique, Vol. XXV, No. 3, 1975

Kézdi Árpád—(Farkas J.—Kabai I.):
Csúszás a salgótarjáni Pécskődombon. Mélyépítéstudományi Szemle, Vol. 26, No. 3, 1976

Kézdi Árpád—(Horváth Gy.):
Shear strength, shear strain and volume change of sand. Proceedings of the Fifth Budapest Conference on Soil Mechanics and Foundation Engineering. Akadémiai Kiadó, Budapest 1976

Kézdi Árpád—(Lőrincz J.):
Validity of Stokes' law in the range of coarse particles. Proceedings of the Fifth Budapest Conference on Soil Mechanics and Foundation Engineering. Akadémiai Kiadó, Budapest 1976

Kézdi Árpád—(Sándor I.):
One-dimensional consolidation in stratified soil. Proceedings of the Fifth Budapest Conference on Soil Mechanics and Foundation Engineering. Akadémiai Kiadó, Budapest 1976

Kézdi Árpád—(Marczal L.—Farkas J.)
Measurement of skin-friction and point resistance of Benoto-piles. Proceedings of the Fifth Budapest Conference on Soil Mechanics and Foundation Engineering. Akadémiai Kiadó, Budapest 1976

Kézdi Árpád—(Marczal L.—Jancsecz S.):
Settlement of a tall, tower-like building subjected to time-dependent load. Proceedings of the Fifth Budapest Conference on Soil Mechanics and Foundation Engineering. Akadémiai Kiadó, Budapest 1976

Kézdi Árpád—(Biczók E.):
Stability and movements of a slope. Proceedings of the Fifth Budapest Conference on Soil Mechanics and Foundation Engineering. Akadémiai Kiadó, Budapest 1976

Kézdi Árpád—(Farkas J.—Kabai I.):
Landslide on a road in a residential area. Proceedings of the Fifth Budapest Conference on Soil Mechanics and Foundation Engineering. Akadémiai Kiadó, Budapest 1976

Kézdi Árpád:
Process of hydraulic soil failure. Acta Techn. Hung. Tom. 1976

Kézdi Árpád:
Problems in Soil Physics. Problemas de Fisica del Suelo. Conferencia dictada en la Universidad. National Autonoma de Mexico. Sociedad Mexicana de Mecanica de del Suelos, Mexico 1976

Kézdi Árpád:
Philosophy of Deep Foundations. Filosofia de las cimentaciones profundas. Third Nabor Carrillo Lecture. Presented at the Eighth National Meeting of Mexican Society of Soil Mechanics. Sociedad Mexicana de Mecanica de Suelos, Mexico 1976

Kézdi Árpád:
Kritische Dichte von Sand. In: Beiträge zur Bodenmechanik und zum Grundbau aus dem In- und Ausland. Dr. Ing. Heinz Muhs zum 65. Geburtstag. Mitteilungen der Deutschen Forschungsgesellschaft für Bodenmechanik an der TU Berlin, Berlin 1976

Kézdi Árpád—(Horváth Gy.):
Stresses and Strains in Sand in Axially Symmetrical Case. Proceedings, Ninth International Conference on Soil Mechanics and Foundation Engineering. Tokyo 1977

Kézdi Árpád—(Kabai I.—Biczók E.):
Sampling macroporous soils. Soil Sampling. Papers presented at the Specialty Session 2. Ninth International Conference on Soil Mechanics and Foundation Engineering. Tokyo 1977

Kézdi Árpád:
Bemerkungen zur Anwendung der Pfahlgründungen. In: Vertikale und kombinierte Tragfähigkeit von Schlitzpfeilern und SOB-Pfählen. II. Internationales Symposium des DDR-Komitees für Bodenmechanik und Grundbau, Weimar, 1976. Bauakademie der DDR. Bauinformation. Berlin 1977

Kézdi Árpád:
Nyilatkozat a tokiói nemzetközi talajmechanikai konferenciáról. (In Japanese). Soils and Foundation, ("Tsuchi-tokiso"). The Japanese Society of Soil Mechanics and Foundation Engineering. Vol. 27, No. 3, Tokyo 1978

Kézdi Árpád:
Old Errors—New Principles. Ground Engineering, July 1978. Vol. 11, No. 5,

Kézdi Árpád:
Surface subsidence due to open-pit coal mining. Periodica Polytechnica- Civil Engineering. Vol. 23, No. 1, 1979. Technical University, Budapest

Kézdi Árpád:
Új elvek és módszerek a geotechnikában. Műszaki Tudomány, Vol. 54 No. 3–4

Kézdi Árpád—(Marczal L.—Jancsecz S.):
A szegedi nagypaneles lakóépületek süllyedése. Mélyépítéstudományi Szemle. Vol. 29, No. 1, 1979

Kézdi Árpád—(Marczal L.)
Földmegtámasztások vasalt talajjal. Mélyépítéstudományi Szemle, Vol. 29, No. 5, 1979

Kézdi Árpád:
Safety factors for different types of failure. In: Vol. 1. Proceedings of the Seventh European Conference on Soil Mechanics and Foundation Engineering. Brighton, England 1979

Kézdi Árpád:
Theorie der Spannungen und Setzungen. In: Bauwerkssetzungen. Entstehung — Berechnung–Messung. Lehrunterlage für eine Veranstaltung der Technischen Akademie Wuppertal. Wuppertal 1979

Kézdi Árpád:
Spannungen und Deformationen in Erd- und Steindämmen, Böschungen. 5th Danube European Conference on Soil Mechanics and Foundation Engineering ČSSR, Bratislava 1977. Proceedings, Vol. IV, 1979

Kézdi Árpád:
A szondázási módszerek szabványosítása. Mélyépítéstudományi Szemle, Vol. XXIX, No. 11, 1979

Kézdi Árpád:
Zur Problematik bei der Ermittlung von Scherfestigkeitsparametern. In: Ergebnisse einer 60jährigen Entwicklung in der Bodenmechanik. Vorträge zum Berg- und Hüttenmännischen Tag 1978. in Freiberg. Ehrenkolloquium für Prof. Dr. Ing. Franz Kögler. Freiberger Forschungshefte A 617

Kézdi Árpád—(Horváth György):
A pórusokban uralkodó víz- és légnyomás mérése a triaxiális nyomókísérletekben. Műszaki Tudomány, Vol. 56, No. 1–2, Akadémiai Kiadó, Budapest 1978

Kézdi Árpád:
Thermodynamics and Soil Physics. Speculations in Science and Technology. Vol. 3, No. 3. Elsevier Sequonia S. A.

Kézdi Árpád—(Otterbein K.):
Volumenänderung von Sand-Ton Gemischen. Sechste Donau Europäische Konferenz für Bodenmechanik und Grundbau. Sektion II, Várna 1980

Kézdi Árpád:
Generalbericht, Sektion II. Sechste Donau-Europäische Konferenz für Bodenmechanik und Grundbau. Bd. Várna 1980

Kézdi Árpád—(Koós–Hutás E.):
Comparative investigation of the grain shape. Scientific Papers of the Institute of Geotechnical Engineering of Wrocław Technical University. No. 32, Wroclaw 1980

Kézdi Árpád—(Mlynarek Zb.):
Static Penetration test results with soils having slight or medium cohesion. Acta Techn. Hung. Vol: 90, 1980

Kézdi Árpád—(Koós–Hutás E.):
A szemcsék alakjának összehasonlító vizsgálata. Műszaki Tudomány, Vol. 57, 1979

Kézdi Árpád—(Lőrincz J.):
Talajfizika és termodinamika. Műszaki Tudomány. Vol. 57, No. 1–2, 1979

Kézdi Árpád—(Biczók E.):
Újabb talajstabilizációs kísérletek a mezőgazdasági útépítésben. Mélyépítéstudományi Szemle Vol. 1983, No. 7

Kézdi Árpád—(Biczók E.):
Some new results in stabilizing agricultural roads. Acta Techn. Hung. 1983
Kézdi Árpád—(Horváth Gy.):
Measurement of pore air and pore water pressures in triaxial testing of soil. Acta Techn. Hung. 1983
Kézdi Árpád—(Horváth Gy.):
Triaxial pressure cell for precise testing. Acta Techn. Hung. 1983

III. Critical reviews

ERDDRUCKTHEORIEN

Beton-und Stahlbetonbau 7/1964 (W. Berlin)

Kézdi, A.: **Erddrucktheorien.** Berlin/Göttingen/Heidelberg: Springer-Verlag 1962. VIII, 319 Seiten mit 275 Abb., Gr. — 8°. Gzl. 58,50 DM.

Trotz der Fülle von Einzelaufsätzen über den Erddruck und Erdwiderstand und die hiermit zusammenhängenden Probleme in den deutschen Fachzeitschriften ist seit dem lange zurückliegenden Erscheinen der beiden klassischen Erddruckbücher in deutscher Sprache von Müller-Breslau (1906) und von Krey (1912; 3. stark erweiterte Auflage und 1. Ausgabe des bekannten Buches 1926) und dem im Rahmen des „Handbuchs für Eisenbetonbau" auch bereits 1936 veröffentlichten Buch von Mund) kein umfassendes Werk mehr herausgekommen, das diese für die Bemessung der mit immer größeren Abmessungen entstehenden Grundbauwerke wichtigsten Größen zum Inhalt hat. Der Grund hierfür mag vielleicht darin liegen, daß der für eine geschlossene Behandlung dieses Themas erforderliche Stoff durch die neueren erdstatischen Theorien einerseits und durch die Erkenntnisse der theoretischen und experimentellen Bodenmechanik über die wirksamen und neutralen Spannungen sowie die Scherfestigkeit der Böden anderseits so umfangreich und schwierig geworden ist, daß eine zusammenfassende Bearbeitung wenig reizvoll erscheint. Um so mehr ist es zu begrüßen, daß von dem Ordinarius für Tunnelbau, Erdbau und Bodenmechanik an der Technischen Universität Budapest, Professor Dr. techn. Kézdi, ein neues Erddruckbuch in deutscher Sprache geschrieben und vom Springer-Verlag herausgegeben worden ist. Es soll die Erddrucktheorien, denen der Ingenieur heute begegnet, kritisch zusammenfassen, ihre Grundlagen darlegen und die Anwendungsbereiche abgrenzen. Unter dieser Zielsetzung muß das Buch natürlich die theoretische Seite der behandelten Probleme betonen; es erfordert dadurch vom Leser beim Studium ein erhebliches Maß von Mitarbeit und Konzentration.

In den ersten Kapiteln werden die heute als Grundlage einer Untersuchung von Erddruckaufgaben anzusehenden Fragen der wirksamen und neutralen Spannungen im Boden, des Ruhedrucks und der Scherfestigkeit der bindigen und nichtbindigen Böden behandelt. Es folgen die Untersuchung der Grundlagen der Grenzgleichgewichtslehre im Erdreich (Gleitflächentheorien) und die damit mögliche strenge Lösung der beiden Sonderfälle des Grenzgleichgewichts im

schwerelosen und im reibungsfreien (vollplastischen) Bodenkörper sowie die Betrachtung der Grenzspannungszustände im unendlichen Halbraum. Erst dann werden die auf einer Grenzwertbestimmung fußenden bekannten eigentlichen Erddruckverfahren mit Verwendung der ungünstigsten ebenen oder gekrümmten Gleitfläche (Coulomb, Fellenius, Rendulic) und anschließend verschiedene Theorien, die auf der Plastizitätslehre aufbauen und vor allem für die Bestimmung des Erdwiderstandes von Bedeutung sind (Boussinesq—Résal—Caquot, Sokolowski) sowie die Gleichgewichtsmethode von Brinch Hansen beschrieben. In einem abschließenden Kapitel wird dann noch auf einige Sonderfälle (Erddruck zwischen parallelen Wänden, in Silos, auf Rohrleitungen und auf Schächte oder Brunnen — räumlicher Fall — sowie bei Verankerungen) eingegangen.

Das trotz der Fülle der behandelten Probleme knapp gefaßte Buch ist ein Versuch, die von verschiedenen Grundlagen ausgehenden unterschiedlichen Erddrucktheorien zusammengefaßt darzustellen und den einzelnen Verfahren den ihnen nach der jeweiligen Betrachtungsweise zukommenden Platz zuzuweisen. Dies ist dem Verfasser ausgezeichnet und mit Verarbeitung von viel eigenem Gedankengut gelungen. Das Buch von Kézdi wird sich deshalb seinen Platz als deutsches Standardwerk über die Grundlagen der Erddrucktheorien erobern und über lange Jahre für den, der sich hierüber unterrichten will, unentbehrlich sein.

H. Muhs

Technisch Tijdschrift 3/1962 (Belgium)

Erddrucktheorien. — par A. Kezdi. - 319 pages - 8° - 275 figures. - Springer Verlag Berlin 1962. - Prix: 58,50 DM.

Née il y a plus de 200 ans, la théorie des poussées des terres n'a cessé de retenir l'attention des ingénieurs qui au cours de ces deux siècles ont établi de multiples théories permettant de résoudre les nombreux problèmes qui se sont multipliées depuis, dans l'art de la construction.

Ces théories ont dépassé depuis longtemps les simples études des calculs de stabilité des murs de soutènement ou les projets de fondations.

Au stade actuel de cette science l'auteur du présent ouvrage a jugé utile d'en faire le point et de mettre à la disposition des ingénieurs qui désirent l'approfondir, les bases théoriques des différentes théories anciennes et modernes qui ont été établies.

Les démonstrations et explications sont faites avec un souci de clarté et de compréhension. Dans certains cas ces démonstrations sont accompagnées de tableaux ou diagrammes.

Les différentes théories sont groupées sous plusieurs chapitres dont les principaux sont les suivants:

— les définitions et calculs des tensions internes,

— résistances aux glissements,

— équations générales des tensions limites d'équilibre,
— déterminations des valeurs limites des poussées,
— solutions d'après la théorie de la plasticité,
— méthodes de détermination des tensions-limites d'équilibre,
— cas particuliers: poussées sur parois verticales, poussées sur canalisations circulaires, calculs de silos, poussées sur parois circulaires, ancrages.

Ces différents chapitres comprennent une abondante littérature bibliographique.

Ouvrage très bien présenté qui sera consulté avec intérêt.

F. Thonnard.

Teknisk Ukeblad 42/1962 (Norway)

A. Kézdi: **Erddrucktheorien.** Springer-Verlag, Berlin, Heidelberg, Göttingen 1962. 319 s. 275 ill. Pris DM 58.50. (2347).

Forfatteren til Springer-forlagets nye bok om jordtrykksteorier er den ungarske professor Arpád Kézdi som er kjent for sine arbeider innen den teoretiske geoteknikk. Kézdi har tatt det løft å sortere den altfor rikholdige litteratur om jordtrykk og stille sammen de teoretiske arbeider som har interesse for vår tids geoteknikere. Og sett ut fra denne målsetning er oppgaven lykkes. Kézdi har omhyggelig skilt ut de vesentligste teorier fra Coulomb til Brinch Hansen og fremstillet dem klart, eksakt og allikevel kortfattet.

Dette har imidlertid bare lykkes ved — noe brutalt — å skille ut empiriske metoder og resultater av modellforsøk og målinger i marken. Sett fra en praktikers synspunkt er dette naturligvis en mangel ved boken, idet det sterkt begrenser muligheten for å vurdere de forskjellige teoriers gyldighet og derved også anvendelsen av boken i praksis. Det er således ikke lett å fatte at man kan skrive en bok om jordtrykk uten å komme inn på så viktige problemer som krefter i avstivninger i byggegroper eller forankrede spuntveggers dimensjonering. Dette siste problem kunne utmerket godt ha vært behandlet, idet Brinch Hansen jo har gitt en ren teoretisk løsning basert på sin bruddlinjeteori, som Kézdi forøvrig har støttet seg sterkt til ved opplegget for boken.

For en geotekniker vil boken sikkert være til megen nytte. Den teoretiske geoteknikk er en av hovedhjørnestenene innen det geotekniske fagområdet, idet man ved den teoretiske behandling kan undersøke hvordan et materiale med idealiserte egenskaper oppfører seg når det belastes eller deformeres, og man kan få en kvalitativ vurdering av hvilke faktorer som er avgjørende for deformasjoner, jordtrykk osv. At geoteknikeren ved siden av må være fortrolig med geologien, med jordartenes virkelige materialegenskaper og dertil ha en sum av praktisk erfaring for å kunne anvende den teoretiske geoteknikk i praksis, skal bare medtas for å stille verdien av et arbeide som det foreliggende i den riktige belysning.

Boken er forøvrig velskrevet og illustrert og utstyrt med vanlig Springer-kvalitet.

Laurits Bjerrum.

Ingeniøren 22/1962 (Denmark)

Erddrucktheorien. Af *A. Kézdi.* Springer Verlag, Berlin 1962. VIII + 319 s., 275 fig. DM 58,50.

Professor Kézdi (fra det tekniske Universitet i Budapest) gennemgår i denne bog meget grundigt de eksisterende jordtryksteorier. Der er herved givet en fremragende oversigt over de indtil nu foreliggende løsningsmetoder, samt resultaterne af en række — mere eller mindre tilnærmede — løsninger. Ifølge sagens natur omhandler bogen hovedsagelig normale jordtryksproblemer i plan deformationstilstand, men i et afsluttende kapitel er der dog også behandlet mere specielle emner som f. eks. silotilstand, jordtryk på ledninger og rumlige (aksialsymmetriske) jordtryk. De grundlæggende emner som jordens styrkelære, beregning af totale og effektive spændinger, elasticitetsteori og plasticitetsteori er behandlet i det nødvendige omfang for forståelsen af metoderne i bogen.

Som titlen antyder, er fremstillingen begrænset til den mere teoretiske del af jordtryksberegningerne. Jordtryksteorien for en væg med vilkårligt omdrejningspunkt er således fyldestgørende referet, men dens anvendelse på spunsvægsberegninger er ikke nærmere omtalt. Jordtryk på bøjelige vægge er i det hele taget ikke behandlet, formentlig fordi der udover Brinch Hansens ikke findes nogen egentlig teori, men kun ingeniørmæssige, mere eller mindre empiriske metoder.

Med denne begrænsning er bogen imidlertid meget fuldstændig, og de enkelte metoder er behandlet så indgående, at man får et virkeligt godt overblik over forudsætninger, beregningsgang og vigtige resultater.

På den anden side er det kun de eksisterende metoder, der er givet. Bogen er således væsentligst refererende, og der er ikke givet nogen ny, samlet behandling af jordtryksproblemet. Bortset fra hviletryk, som er et elasticitetsteoretisk fænomen, er jordtryksteorierne derfor opdelt på den klassiske måde:

1. Ekstremummetoderne, der omfatter Coulombs teori med rette brudlinier, Fellenius' teori med cirkulære brudlinier (i ler), og Rendulics teori med logaritmiske spiraler (i sand).
2. Plasticitetsteorien (traditionelt opfattet som kun vedrørende rene zonebrud): Rankines teori med et net af rette brudlinier, samt Jelineks udvidelse til det generelle Coulombske halvrum; Soholovskis teori, der også omfatter radialzoner, og Boussinesq-Resal-Caquot's specialmetode for brudfigurer, ligedannet om et punkt ved ubelastet jordoverflade (identisk med Karman's).
3. Ligevægtsmetoden, d. v. s. Brinch Hansens teori med mere generelle brudfigurer, herunder liniebrud, samt den specielle tilnærmede beregningsmetode for zonebrud.

Man kunne i denne forbindelse ønske en understregning af, at når et tilstrækkeligt sæt ideale forudsætninger er opstillet for jordens plastiske egenskaber, vil der være en og kun en matematisk korrekt løsning til ethvert givet

jordtryksproblem. Derfor bør de forskellige jordtryksteorier helst betragtes under et enhedssynspunkt, idet de enten vil være specialmetoder til bestemmelse af den korrekte løsning for hver sin bestemte type grænsebetingelser, eller er tilnærmelsesmetoder, hvis nøjagtighed igen må afhænge af grænsebetingelsernes form. Det havde været af betydelig værdi, hvis disse forhold var trukket klarere frem, eventuelt direkte gennem bogens disposition.

Det er beklageligt, men i og for sig naturligt, at i et refererende værk af denne art vil fejl i de angivne metoder ofte ikke blive rettet eller blot kommenteret. Forfatteren har således ukritisk overtaget metoden med diskontinuitetslinier i brudzoner fra Prager, Josselin de Jong og Soholovski. Herved kan man formelt opnå løsninger i tilfælde, hvor brudfiguren faktisk er kinematisk umulig (f. eks. under jordoverflader med et udadgående knækpunkt, d. v. s. hvor vinklen i jord er mindre end 180°). Som et kuriosum kan det nævnes, at man ved at gennemføre denne betragtning konsekvent når til, at der ikke kan findes brudzoner under et indadgående knækpunkt i en jordoverflade (vinklen gennem jord større end 180°). I virkeligheden findes der en statisk bestemt, og også matematisk korrekt, brudzone i dette tilfælde, som indeholder to radialzoner. Løsningen er åbenbart ikke angivet i den af forfatteren kendte litteratur, og derfor ikke refereret.

Det kan også volde misforståelser, når forfatteren beklager, at Boussinesq-Resal-Caquot's metode giver løsninger for enhver mulig vægfriktionsvinkel δ, og således ikke kan siges at være entydig. For Coulombs jordtryksteori (for aktivt jordtryk) klares dette problem efter Rebhann ved en noget kompliceret betragtning, som kort refereret går ud på, at man for væghældninger mellem to bestemte værdier kan fastsætte δ, således at jordtrykket bliver minimum. Uden for intervallet må man enten bruge en værdi fundet ved Rankine-zonen eller må sætte δ lig med den faktiske vægruhed.

Det skal hertil bemærkes, at ved et rent zonebrud må der altid ske bevægelse mellem jord og væg, som går i samme retning i hele væggens højde. δ er altså altid lig med vægruheden. Noget andet er, at zonebrud kun er mulige for et vist område af de indgående parametre (væghældning, vægruhed, bevægelse etc.). Uden for dette område har man et stift legeme langs i hvert fald en del af væggen, og her vil den fulde vægruhed muligvis ikke blive mobiliseret.

Fejl af denne art gør, at bogen ikke kan anvendes ukritisk som opslagsbog, hvis man ønsker den eksakte løsning på et givet problem. Den kan imidlertid være en udmærket hjælp til at finde en tilnærmet løsning, og som en dybtgående orientering om jordtryksteoriernes nuværende stade er den fremragende. Der savnes stærkt et sagsregister, selv om navneregistret vil være en hjælp for de læsere, der i forvejen er orienteret i den geotekniske litteratur.

Bent Hansen.

Основания фундаменты и механика грунтов 4/1962 (Sovietunion)

Арпад Кезди: **Теория давления грунтов.** Берлин — Гёттинген — Гейдельберг, 1962, 319 стр. (на немецком языке)*

В книге известного венгерского ученого А. Кезди рассмотрены основные вопросы теории давления грунта, которые автор во введении делит на три группы: 1) проблемы, связанные с определением напряженного состояния естественного массива грунта под нагрузкой; 2) задачи давления грунта на ограждения, которые могут перемещаться — на подпорные и щпунтовые стенки, а также на стенки силосов; 3) вопросы давления засыпки силоса на его днище и грунта на подземные сооружения.

Методы определения давления грунта подразделены на четыре группы. К первой группе отнесены методы, основанные на теории упругости, ко второй — на теории пластичности, к третьей — на законах кинематики и к четвертой — на вариационном принципе экстремальности. В книге отражены все эти методы.

В первой главе изложена теория напряжений в грунтах. Здесь рассмотрены гидростатическое и капиллярное давления грунтовой воды; зависимость между главными напряжениями и напряжениями по наклонным площадкам; напряженное состояние грунта при помощи кривых, кругов Мора, эллипса полных напряжений, овала нормальных напряжений и четырехлистника касательных напряжений.

Во второй главе подробно освещены вопросы давления грунта и грунтовой воды в естественном массиве.

В третьей главе рассмотрено сопротивление грунта сдвигу как в качестве замкнутой, так и в качестве открытой системы с учетом и без учета сцепления. Показана связь между сдвигающими напряжениями и перемещениями сдвига и влияние продолжительности нагружения. Эта глава заканчивается данными об объемных весах грунтов в зависимости от их плотности и влажности.

Глава четвертая посвящена общим уравнениям теории предельного равновесия грунта. В зависимости от того, какие из трех механических констант грунта (объемный вес, угол внутреннего трения, удельное сцепление) принимаются равными нулю, автор различает восемь родов задач, среди которых находят место задачи теории пластичности и теории сыпучей среды.

В дальнейшем изложении, касающемся связи между характеристиками дифференциальных уравнений и линиями скольжения сыпучего тела, а также вывода уравнений Кёттера, автор в основном придерживается методики изложения, принятой В. В. Соколовским. Большой интерес представляет последний параграф этой главы, в котором рассматривается

влияние поворота стенки на вид поверхности скольжения, построение круга Мора для деформаций сыпучего тела и возможные формы образования областей предельного равновесия за ограждением по Хансену.

В пятой главе даются уравнения предельного равновесия в полярных координатах для невесомого тела, а в шестой — для тела, лишенного внутреннего трения, т. е. для идеально пластического тела. Здесь приведены решения для предельной центральной нагрузки на заглубленный фундамент, для предельной нагрузки на дно скважины, для активного и пассивного давления грунта на плоскую стенку, для предельной высоты плоского откоса и для давления грунта на круговую обделку тоннеля.

В главе седьмой описывается предельное напряженное состояние наклонной полуплоскости и применение теории Ренкина к определению напряжений, действующих по контуру закрепленного кругового отверстия. Далее дано решение В. В. Соколовского для сыпучего клина.

В восьмой главе приведены решения теории предельного равновесия, основанные на принципе экстремальности. Здесь рассмотрено: определение активного давления идеального сыпучего тела на подпорную стенку по методу Кулона, вывод теорем Ребхана, различные графические построения для определения равнодействующей активного давления грунта на подпорную стенку, давление на нее от равномерной и сосредоточенной нагрузок и способ учета сцепления грунта. Приведены подробные таблицы коэффициентов активного давления грунта.

В этой же главе освещены методы Феллениуса и Рендулика, учитывающие кривизну поверхности скольжения при определении активного давления грунта на подпорную стенку путем принятия производящей этой поверхности в качестве дуги окружности или логарифмической спирали.

В главе девятой изложены общие теории предельного напряженного состояния, относящиеся к определению давления грунта на подпорные стенки: Буссинеска—Резаля—Како и В. В. Соколовского.

В главе десятой рассмотрено определение активного давления грунта на подпорную стенку по методу Хансена, являющемуся по существу применением уравнения Кёттера к расчетной схеме скольжения засыпки за подпорной стенкой по круглоцилиндрической поверхности.

Для облегчения расчетов по этому довольно сложному методу приводятся вспомогательные таблицы коэффициентов, относящихся к напряжениям, силам и моментам.

В этой же главе рассматривается давление грунта на гибкие и шероховатые жесткие шпунтовые стенки, могущие испытывать перемещения в грунте.

Последняя 11-я глава посвящена особым, но связанным между собой случаям давления земли. Здесь прежде всего рассмотрено определение давления земли, заключенной между параллельными стенками по методу

Янсена, дополненному учетом сцепления по Терцаги. Представляет интерес приведенная в книге эпюра вертикальных напряжений, имеющая максимум посредине высоты стенок.

Во втором параграфе этой главы показано определение по методу А. Фёльми давление на трубы в траншеях и насыпях, а в третьем — давление сыпучего тела на дно и стенки силоса по Янсену — Кёнену и по Како. При этом приведены формулы и графики, отражающие увеличение давления при опорожнении силоса. В четвертом параграфе рассмотрены основные решения, полученные В. Г. Березанцевым для осесимметричных задач. Наконец, в последнем параграфе рассмотрено сопротивление грунта у анкерных плит.

Таким образом, в книге Арпада Кезди изложен обширный круг вопросов, относящихся к давлению грунтов, рассматриваемых в качестве сыпучих тел. Некоторые из этих вопросов, а также трактовка других из них представляют новизну для советского читателя. Это — диаграммы связи между напряжениями и перемещениями грунта при сдвиге, классификация задач предельного равновесия и методов их решения, формы областей предельного равновесия грунта за стенкой в зависимости от вида ее перемещения, давление грунта на тоннель, проведенный на косогоре, определение давления грунта на подпорную стенку по методу Хансена, учет сосредоточенной нагрузки на поверхности грунта за подпорной стеной и учет сил сцепления.

Вместе с тем из поля зрения автора выпали многие вопросы, тесно связанные с взятой им темой, которые бы могли значительно обогатить содержание его книги и приблизить ее к требованиям практики. Это — определение давления грунта на ломаную и на пологую подпорную стенки, а также на стенки с разгрузочными площадками и углового профиля, определение давления грунта на подпорную стенку ограниченной длины в условиях пространственной задачи и определение динамического давления грунта.

Что касается вопросов кинематики, то они в книге затронуты, но их изложение не доведечо до возможности практического использования для расчета давления сыпучего тела с учетом величины перемещения подпорной стенки.

Не получили также освещения в книге графические методы определения давления грунта на подпорные стенки, предоженные С. С. Голушкевичем.

Вопросы, связанные с давлением грунта на тонкие (шпунтовые) стенки, на подземные сооружения, вопросы прочности оснований и устойчивости откосов освещены недостаточно. Их либо не следовало касаться совсем, считая, что они выходят из круга вопросов, рассматриваемых в книге, дибо их нужно было рассмотреть более полно. Однако последнее привело бы к сильному увеличению объема книги.

Автор привел с надлежащими ссылками большие выдержки из работ В. В. Соколовского и В. Г. Березанцева. Ряд работ других советских ученых, в которых теория давления грунтов и других сыпучих тел получила значительное развитие во многих направлениях, в книге не нашел своего отражения.

К достоинствам книги следует отнести стройность ее построения, строгость выводов, ясность изложения и четкость иллюстраций.

Несмотря на имеющиеся недостатки, книга проф. Арпада Кезди заслуживает положительной оценки и представляет интерес для советского читателя.

Г. К. Клейн

Schweizerische Bauzeitung 38/1962 (Switzerland)

Erddrucktheorien. Von *Arpád Kézdi,* Professor an der Technischen Universität Budapest. 319 Seiten, 275 Abb. Berlin 1962, Springer-Verlag. Preis DM 58.50.

Die Frage nach dem auf eine Stützmauer wirkenden Erddruck hat seit Coulomb die Wissenschaftler und die Ingenieure beschäftigt. Bald zeigte sich, daß die Fragestellung erweitert werden mußte auf die Untersuchung des Grenzgleichgewichtes im Boden, wie sie sich bei vielen Aufgaben des Tiefbaues stellt. Viele Theorien wurden entwickelt, die teils in Vergessenheit gerieten und erst in neuerer Zeit wieder ihrer Bedeutung gemäß Beachtung fanden (z.B. F. Kötter 1893), teils unbekannt blieben, teils aber zum Gemeingut des Bauingenieurs wurden. Das Bedürfnis nach einer zusammenfassenden Darstellung dieser Theorien unter dem Gesichtspunkt des Bodenmechanikers bestand schon lange. Kézdi hat sich dieser Aufgabe durch stark konzentrierte Darstellung des großen Gebietes mit Erfolg unterzogen. Nach einführenden Kapiteln über den Spannungszustand im Boden, den Ruhedruck und einer übersichtlichen Darstellung der Scherfestigkeit von Böden werden die allgemeinen Gleichungen des Grenzgleichgewichtes eingehend abgeleitet und angewandt auf den schwerelosen und den reibungsfreien Körper. Die Behandlung der plastischen Grenzzustände im unendlichen Halbraum leitet über zu den Erddruckproblemen im engeren Sinne: Bestimmung der Grenzwerte des Erddruckes und Erdwiderstandes nach den verschiedenen Methoden (Coulomb, Rankine, Fellenius, Rendulic, Caquot, Sokolowski, Brinch Hansen). Die Behandlung einiger Sonderfälle beschließt das Werk. Druck und Ausstattung sind vorzüglich. Bei einer Neuausgabe sind neben der Ausmerzung der unvermeidlichen Druckfehler doch hier und dort gewisse ergänzende Bemerkungen oder Hervorhebungen im Texte, zur Erleichterung des Studiums, anzuraten. Das Buch ist dem fortgeschrittenen Studenten, aber vor allem dem praktisch tätigen Ingenieur, der seinen Berechnungsmethoden kritisch gegenübersteht und in die Grundlagen der von ihm angewandten Methoden Einblick gewinnen will, sehr zu empfehlen.

Prof. G. Schnitter, ETH, Zürich

HANDBUCH DER BODENMECHANIK

Bauplanung-Bautechnik 10/1969 (German Dem. Rep.)

Handbuch für Bodenmechanik. Band I. Bodenphysik. Von *A. Kézdi.* Übersetzung aus dem Ungarischen. VEB Verlag für Bauwesen, Berlin 1969. 21 cm · 29,5 cm, 260 Seiten, 400 Bilder und 37 Tabellen. Leinen 44,— M.

Mit der geplanten Herausgabe des vierbändigen Handbuches der Bodenmechanik wird erstmalig in deutscher Sprache eine nahezu vollständige Zusammenfassung über den gegenwärtigen Wissensstand in der theoretischen, experimentellen und angewandten Bodenmechanik vorgelegt.

In den neun Kapiteln des ersten Bandes „Bodenphysik" werden Fragen der Zusammensetzung, der Klassifikation und Struktur der Böden, Fragen der Wasserbewegung im Untergrund und schließlich Festigkeits- und Formänderungseigenschaften der Erdstoffe sowie Stabilitätsprobleme im Erdreich behandelt.

Im Vordergrund des ersten Bandes stehen solche Probleme, die das Verhalten des Untergrundes und die Eigenschaften der Erdstoffe infolge der eigenen Druckhaftigkeit im unbelasteten und belasteten Zustand bestimmen. Hierbei sind besonders die neuesten Erkenntnisse der zeitabhängigen Vorgänge in der Bodenmechanik einschließlich der rheologischen Eigenschaften der Erdstoffe berücksichtigt. Zahlreiche durchgerechnete Beispiele und eine Fülle physikalischer sowie bodenmechanischer Kennzahlen vergrößern das Verständnis für das disperse Dreistoffsystem.

Besonders ansprechend sind die Darstellungen der physikalischen Kennzahlen in Dreiecksnetzen.

Der Autor versteht es, die zum Teil sehr komplizierte Thematik der Konsolidierungstheorie, der Festigkeitslehre oder der Bruchtheorie systematisch und verständlich darzustellen. Er bedient sich hierzu einer klaren mathematischen Formulierung und sprachlichen Ausdrucksform. Insofern bereitet das Studium der „Bodenphysik" dem interessierten Leser ein großes Vergnügen. Es bildet gleichermaßen für Studenten und Fachleute ein ausgezeichnetes Lehr- und Nachschlagewerk. Aus ihm können wesentliche Erkenntnisse zur Lösung praktischer Aufgaben gewonnen werden.

Auf die in Vorbereitung befindlichen drei Bände dürfen die Leser schon heute gespannt sein.

Band II Bodenmechanik im Erd-, Grund- und Straßenbau
Band III Bodenmechanisches Versuchswesen
Band IV Anwendung der Bodenmechanik in der Praxis

Ewert

Teknisk Tidskrift 16/1969 (Sweden)

Handbuch der Bodenmechanik, bd 1: *Bodenphysik,* av *Arpád Kézdi.* VEB Verlag für Bauwesen, Berlin 1969. 260 s., 400 fig., 37 tab. 44 DM.

Ytterligare ett tillskott till den senaste tidens flora av geotekniska handböcker är denna bok. Mera ovanligt torde vara att den kommer från Östeuropa — en östtysk omarbetning av ett äldre ungerskt verk. Den tyska bearbetningen har medfört en märkbar utvidgning och mera tidsenlig internationell anpassning. Särskilt värdefullt är anvisningarna till östeuropeiska arbeten, även om litteraturförteckningarna som förekommer inte är särskilt omfattande.

Handboken kommer i färdigt skick att omfatta fyra band och att behandla förutom jordfysik (bd 1) speciell och allmän tillämpning av geotekniken (bd 2 och 4) samt mättekniska spörsmål på laboratorium och i fält (bd 3). Bd 1 ger ett trevligt intryck såväl dispositionsmässigt som typografiskt med pedagogiskt riktig omväxling mellan rubriker, text och figurer. Texten är tydlig och utrymmet för bilder och tabeller väl avvägt. Det är heller inget tungt svårhanterligt band utan det synes verkligen kunna tjäna som en praktisk handbok.

Erik Danfors

Bau + Bauindustrie 7/1971 (German Fed. Rep.)

Handbuch der Bodenmechanik. Band II: Bodenmechanik im Erd-, Grund- und Straßenbau. Von Prof. Dr. techn. Arpád Kézdi. Deutsche Bearbeitung: Prof. Dipl.-Ing. Walter Kinze, Dresden. Gemeinschaftsauflage des VEB Verlages für Bauwesen, Berlin und des Verlages der Ungarischen Akademie der Wissenschaften, Budapest. Erscheinungsjahr: 1969. Umfang, Format, Ausstattung: 309 Seiten, 553 Abbildungen, 50 Tabellen; DIN A 4 Hochformat; Leineneinband.

Im Anschluß an den Band I seines Handbuches, in dem die Grundlagen der Bodenphysik (Struktur und Klassifikation der Erdstoffe, Spannungen im Boden, Wasser im Untergrund Festigkeit und Formänderung von Erdstoffen etc.) dargestellt sind, behandelt nun der Verfasser, einer der führenden Fachleute auf dem Gebiet der Bodenmechanik, im vorliegenden Band II die Bodenmechanik des Erd-, Grund- und Straßenbaues, die Verbesserung der physikalischen Eigenschaften von Erdstoffen und die Bodendynamik, also den Einfluß von Schwingungen auf Erdmassen.

Die theoretischen Zusammenhänge werden anhand mathematischer Ableitungen und ausführlicher Erläuterungen dargestellt. Da die Probleme meist von den grundlegenden Ausgangsgleichungen — z. B. der Elastizitätstheorie — her angepackt und ihre Lösungen in knapper, doch anschaulicher Weise vorgeführt werden, nimmt dieses Werk den Rang eines hervorragenden Lehrbuches ein. Die Diskussion und Beurteilung verschiedener Verfahren oder Ansätze verschiedener Forscher dienen nicht nur dem Studierenden oder dem wissenschaftlich Tätigen

bei der kritischen Auseinandersetzung mit dem Stoff, sie helfen auch dem Praktiker bei der Entscheidung, welche Methode für „seinen Fall" die zweckmäßigste ist. Durch diese breite Diskussion der Verfahren vermittelt das Werk einen Überblick über den internationalen Stand der Forschung auf den behandelten Gebieten. Der praktisch tätige Ingenieur wird aber vor allem in der Fülle der gebrauchsfertigen Anwendungsformeln mit den entsprechenden Aufbereitungen in Form von graphischen Darstellungen und Tabellen eine wirksame Hilfe bei allen wesentlichen Problemen finden.

Die Zahl der Beispiele ist in dem vorliegenden Band ebenso wie Literaturangaben mit Absicht beschränkt. Der Verfasser beabsichtigt nämlich, im Band III, der das bodenmechanische Versuchswesen behandeln wird, ein für die ersten drei Bände geltendes ausführliches Quellenverzeichnis zu liefern, während der geplante Band IV, der der Anwendung der Bodenmechanik in der Praxis gewidmet sein wird, mit reichhaltigen Beispielen ausgestattet werden soll.

Zusammenfassend darf gesagt werden, daß der vorliegende Band II dieses Handbuches der Bodenmechanik eine sehr wertvolle Bereicherung der Literatur dieser Disziplin darstellt. Hinsichtlich der deutschen Bearbeitung durch Prof. Kinze bleiben keine Wünsche offen. Das Werk, gleichermaßen für Studium und Praxis geeignet, verdient uneingeschränktes Lob.

Dr.-Ing. E. Grasser

Bauplanung-Bautechnik 4/1971 (German Dem. Rep.)

Handbuch der Bodenmechanik. 1. Auflage. Band II Bodenmechanik im Erd-, Grund- und Straßenbau. Von *A. Kézdi,* VEB Verlag für Bauwesen, Berlin 1970. 21 cm · 29,7 cm, 312 Seiten, 553 Bilder und 48 Tabellen. Leinen 52,— M.

Der vorliegende zweite Band des Handbuches der Bodenmechanik — Bodenmechanik im Erd-, Grund- und Straßenbau — folgt im wesentlichen dem Arbeitstitel und der Gliederung, wie sie als Anhang im ersten Band angegeben wurden. Der Verfasser hat sich in diesem Buch die Aufgabe gestellt, den gegenwärtigen internationalen Stand der Kenntnisse auf dem Gebiet der angewandten Bodenmechanik aufzuzeigen, die durch seine eigenen Veröffentlichungen und Forschungsergebnisse wesentlich bereichert wurden. Die tiefgründige Aussage der behandelten Problematik in der Theorie und in ihrer Anwendung basiert auf einer umfassenden Einschätzung der bodenphysikalischen und bodenmechanischen Grundlagen der Erdstoffe in der Wechselwirkung zwischen Baugrund und Bauwerk. Dabei kommt den analytischen Zusammenhängen hinsichtlich Stabilität sowie Spannungs- und Verformungseigenschaften des Erdstoffes als Bauwerk und Baugrund eine vorrangige Bedeutung zu.

Die Aufdeckung der spannungs- und verformungstheoretischen Zusammenhänge nimmt einen breiten Raum ein und stellt besonders an die mathemati-

schen Kenntnisse des Nutzers hohe Anforderungen. In dieser Hinsicht wurde keine Vereinfachung im Sinne einer Verwässerung zugelassen.

Der Verfasser ist dieser Schwierigkeit begegnet durch eine sehr übersichtliche und logische Gliederung des Stoffes und durch das ständige Bemühen, die Entwicklungsschritte und Ergebnisse umfassend und in ausgezeichneter Qualität in Bildern und grafischen Darstellungen zu veranschaulichen.

Zum Verständnis der Problematik trägt entscheidend das kritische Werturteil des international anerkannten Fachmannes bei, der eine ständige Aufgabe darin sieht, die strengen theoretischen Lösungen selbst und mit der Wirklichkeit zu vergleichen und sorgfältig auszuwerten. Auf diese Weise wird systematisch die Aufmerksamkeit auf die zusammenhängenden grundsätzlichen Randbedingungen bei der praktischen Anwendung der Ergebnisse gelenkt und darüber hinaus das aufgezeigt, was in der weiteren Forschung und Überprüfung durch die Praxis einer Lösung bedarf.

Im Abschnitt Bodenmechanik der Erdbauten stehen die bekannten Verfahren zur Berechnung der Standsicherheit der Böschungen, der Spannungen in Dämmen sowie der Sohlspannungen und Grundbrüche unter Dämmen im Vordergrund. Besondere Bedeutung findet dabei die Einschätzung der Faktoren, die die Standsicherheit der Böschungen beeinflussen, insbesondere die Faktoren Zeit und Wasser, und das Problem der Definition der Standsicherheit.

Ausgehend von der gründlichen Analyse des Lastsetzungs- und Zeitsetzungsdiagramms des Baugrundes befaßt sich der Abschnitt Bodenmechanik des Grundbaus mit den Problemen der Grenztragfähigkeit von Flachfundamenten und Pfählen sowie mit der Halbraumtheorie der Spannungsverteilung und mit den Verformungseigenschaften, insbesondere den Setzungen des Baugrundes.

Der Verfasser analysiert eingehend die Wechselwirkung zwischen Baugrund und Bauwerk sowie die Problematik der mangelhaften Übereinstimmung der errechneten mit den in der Praxis gemessenen Setzungen.

Der Abschnitt Bodenmechanik des Straßenbaus beschäftigt sich eingehend mit dem Aufbau und der Statik der Deckenkonstruktionen, die die Spannungen und Verformungen aus der Wechselwirkung zwischen Decke und Untergrund infolge Verkehrslast aufnehmen müssen. Die bekannten Verfahren werden auf ihre Anwendbarkeit hin einer gründlichen Prüfung unterzogen. Eine Vielzahl von Kennzahlen erleichtert die unmittelbar praktische Anwendung der Verfahren.

Zur Verbesserung der physikalischen Eigenschaften von Erdstoffen werden in einem gesonderten Abschnitt die bekannten Stabilisierungsverfahren eingeschätzt mit dem Ziel, Erdstoffe mit ungünstigen Eigenschaften hinsichtlich Verdichtbarkeit und Durchlässigkeit zu verbessern und wirtschaftlich einzusetzen.

Der letzte Abschnitt beschäftigt sich im wesentlichen mit dem Einfluß von Schwingungen auf Erdmassen.

Der vorliegende 2. Band ist ein würdiges Glied der Gesamtkonzeption des Verfassers, ein Standardwerk der gesamten Bodenmechanik zu schaffen, das auf

Bekanntem und Bewährtem aufbaut und die neuesten wissenschaftlichen Ergebnisse berücksichtigt.

Die ausgezeichnete Ausstattung des Werkes und die Sorgfalt bei der Drucklegung tragen wesentlich mit zum Gelingen dieses Vorhabens bei.

Das Buch kann jedem, der sich mit der Bodenmechanik beschäftigt, bestens empfohlen werden.

Ewert

Bauplanung-Bautechnik 10/1973 (German Dem. Rep.)

Handbuch der Bodenmechanik. 1. Auflage. Band III Bodenmechanisches Versuchswesen. Von *A. Kézdi*. VEB Verlag für Bauwesen, Berlin, und Verlag der Ungarischen Akademie der Wissenschaften, Budapest 1973. 21 cm × 29,5 cm, 284 Seiten, 345 Bilder, 33 Tafeln, 30 Formblätter und 3 Anlagen. Leinen 47,— M.

Der III. Band gliedert sich in fünf Abschnitte
Erkundung und Aufschluß des Baugrundes
Untersuchungen im Laboratorium
Grundwasseruntersuchungen
Untersuchung der Tragfähigkeit des Baugrundes
Untersuchung von Erdbauten

Bei der Abhandlung des Stoffgebietes ging der Autor davon aus, die physikalischen und mechanischen Eigenschaften nicht als Selbstzweck zu bestimmen und zu bewerten, sondern die sie charakterisierenden Kennzahlen für die erd- und grundbaupraktische Anwendung zu nutzen zur

— Charakterisierung, Beschreibung und Klassifizierung von Lockergesteinen
— Durchführung von Berechnungen in der Grund- und Erdbaumechanik
— Qualitätsvorgabe und Qualitätskontrolle im Erdbau.

Es wäre allerdings zu wünschen, daß die Fragen der Anzahl und des Standortes der Aufschlüsse und ihre Repräsentanz für die anstehenden Baugrundverhältnisse des Makro- und Mikrostandortes sowie die Probleme der Spezifizierung von Erdstoffproben am Bohrort zur Erhöhung der Abbildgenauigkeit des Baugrundmodells breiter und differenzierter behandelt werden.

Die Gliederung ist übersichtlich und logisch aufgebaut, so daß der interessierte Fachmann und Studierende sich schnell einen Überblick über den Inhalt verschaffen kann, aber auch rasch das Detail findet.

Die rd. 400 Bilder, Tafeln und Übersichten sind sorgfältig ausgewählt und informativ im Inhalt. Sie tragen wesentlich zum Verständnis der Problematik und zur Einprägsamkeit des zum Teil schwierigen Stoffes bei.

Den Schluß des III. Bandes bildet ein umfangreiches Quellenverzeichnis von rd. 600 Literaturangaben namhafter Autoren, die auf diesen Gebiet im internationalen Maßstab ihre Ergebnisse aus Forschung und Praxis veröffentlicht haben.

Der Verfasser hat mit dem vorliegenden III. Band einen weiteren erfolgreichen Schritt getan, die gegenwärtigen internationalen Erkenntnisse auf dem Gebiet des bodenmechanischen Versuchswesens zusammenzufassen und im gewissen Sinne zu normieren. Das Handbuch wird nicht nur der weiteren Entwicklung und Vervollkommnung der Theorie und Praxis des bodenmechanischen Versuchswesens neue wirksame Impulse geben, sondern auch zur weiteren Standardisierung im nationalen wie internationalen Maßstab, besonders innerhalb des RGW, beitragen.

Ewert

Tiefbau-Ingenieurbau-Strassenbau 4/1973 (German Fed. Rep.)

Arpád Kézdi: **Handbuch der Bodenmechanik** Band I: Bodenphysik 1968; Band II: Bodenmechanik im Erd-, Grund- und Straßenbau 1969, Berlin—Budapest. Gemeinschaftsauflage VEB-Verlag für Bauwesen Berlin und Verlag der Ungarischen Akademie der Wissenschaften, Budapest, 1968/69. Band I 258 Seiten, 395 Bilder, 36 Tabellen. Band II 309 Seiten, 553 Bilder, 52 Tabellen, Ganzleinen

Der Verfasser, international bekannt durch seine zahlreichen wichtigen bodenmechanischen und bodenphysikalischen Abhandlungen, Verfasser des in deutscher Sprache 1964 in denselben Verlagen erschienenen zweibändigen Werkes »Bodenmechanik«, legt nach einer fünfjährigen Zeitspanne das als umgestaltetes und erweitertes Werk betitelte »Handbuch der Bodenmechanik« im Großformat vor.

Dem Vorwort von Walter Kinze in Dresden — fast gleichlautend wie das aus dem Jahre 1964 — folgt in Anlehnung an die Ausgabe des Jahres 1964 in neun Abschnitten unterteilt die Darstellung der Bodenphysik: 1. Ursprung der Boden und bodenphysikalische Kennziffern; 2. Bestandteile des Bodens; 3. Klassifikation der Erdstoffe; 4. Struktur der Erdstoffe; 5. Spannungen im Boden; 6. Bewegung des Wassers im Untergrund; 7. Festigkeit der Erdstoffe; 8. Formänderung von Erdstoffen; 9. Bruchzustände im Erdreich und zusätzlich ein Abschnitt über Formelzeichen. Der Verfasser betont die Unabhängigkeit von seinem früheren Werk »Bodenmechanik«. Er ist dazu berechtigt, denn er hat eine wesentliche Erweiterung des Stoffgebietes unter Gliederung in neun statt ursprünglich fünf Kapiteln gebracht.

Er ist Bodenmechaniker und Ingenieur. Dadurch wird es verständlich, daß er das Schwergewicht auf die bodenphysikalischen und mechanischen Grundlagen und Kennziffern legt, weniger auf die genetisch begründeten, erst die unterschiedlichen bodenphysikalischen Kennwerte und Verhalten begründenden physikalischen Eigenschaften. Es wäre unbedingt im Rahmen eines Handbuches erwünscht, den in zahlreichen Publikationen international manifestierten Grundlagen der »festen« und »veränderlichfesten« (festen und pseudofesten) Erd- und Felsarten die gebührende Beachtung zu schenken und vor allem die dafür

maßgebende Literatur anzuführen, z. B. K. Keil: Geotechnik und »Der Dammbau«, die dem Autor zweifelsfrei bekannt sind und die vor allem dabei in die Praxis führen. Insofern vermißt der Ref. die Wechselbeziehungen zwischen geologischer, also erdgeschichtlicher und bodenphysikalischer Grundlage.

Seine Ausführungen über die Arten der Wasserbindung wären im Sinne der Geotechnik, Band III, Seite 163 ff., zu ergänzen. Besonders ist aber das Wesen und der Unterschied der festen und veränderlichfesten Gesteine (Bodenarten und Felsarten) zu berücksichtigen.

In der Disposition der Kapitel und deren Bezeichnung ist der Wechsel der Bezeichnungen für Erdart — als nach »Ohde« sinnvollsten — bemerkenswert, es wäre zweckmäßig, hier eine einheitliche Bezeichnung, z. B. Erdart oder, wie im deutschen Normenwesen, »Bodenart« im Gegensatz zu Felsart, zu wählen. Es ist — wie bereits früher betont — bemerkenswert, daß Handbücher dieser Art nicht von den hierzu berufenen Ordinarien der deutschen Hochschulen, sondern von ausländischer Seite verfaßt werden. Dies ist wohl ein einzigartiger Vorgang im Vergleich zu sämtlichen anderen naturwissenschaftlichen, medizinischen und technischen Disziplinen.

Band II. Bodenmechanik im Erd-, Grund- und Straßenbau, dem noch Band III: Bodenmechanisches Versuchswesen und Band IV: Anwendung der Bodenmechanik in der Praxis folgen sollen, bringt in fünf Abschnitten im Rahmen eines erweiterten Handbuches drei Kapitel des Bandes II aus dem Jahre 1964. Es enthält zusätzlich als Abschnitt vier: »Die Verbesserung der physikalischen Eigenschaften von Erdstoffen« und als fünften Abschnitt »Bodendynamik von Schwingungen auf Erdmassen«, während die Bodenerkundung von früher weggefallen ist.

Auch hier hat der Verfasser unter Bereicherung seiner Ausführungen und Gedankengänge durch zahlreiche instruktive Abbildungen und mathematische Ableitungen — sein besonderes Steckenpferd — durch zahlreiche Tabellen und Berücksichtigung auch ausländischer Erfahrungen ein umfassendes im wahrsten Sinne des Wortes gültiges Handbuch für den Theoretiker und Praktiker verfaßt, wofür man ihm volle Anerkennung und Dank zollen muß. Dies schließt nicht aus, daß auch hier einige Ergänzungen wünschenswert erscheinen. Z.B. ist das Frostkriterium der »festen« und »veränderlichfesten« Felsarten und die besondere Frostgefährlichkeit gerade dieser Gesteine als Sammelbegriff für Erd- und Felsarten nicht berücksichtigt worden, obwohl im Standardwerk von Ruckli ausdrücklich dieses Kriterium als entscheidend aufgeführt ist, während wiederum die des dafür zuständigen bekannten Autors fehlt. Auch die Frage der Frostsicherung durch Wasserabwehr, nämlich durch Stabilisierung und Dichtung frostempfindlicher Bodenarten, ist nicht berücksichtigt, während in die Normen neben den Erdarten auch die Felsarten — in der BRD seit 1968 und bereits vor 15 Jahren auf Initiative des Ref. in der DDR — aufgenommen wurden.

Diese wenigen Beispiele mögen genügen, das Interesse des Autors auf weniger beachtete und übersehene Tatsachen zu lenken, um seinem Handbuch

noch mehr Inhalt, noch mehr Vollständigkeit und damit noch größeres Gewicht zu verleihen.

Der VEB-Verlag hat auf holzfreiem Papier das Werk dem internationalen Standard angepaßt und dadurch im Gegensatz zu früher die Voraussetzung für den inneren Wert auch in der vortrefflichen Ausstattung in Wort und Bild gegeben. Als Ratgeber zum Studium und zur Anwendung in der Praxis verdient das Werk Beachtung und Verbreitung.

Bauingenieur 53/1978 (German Fed. Rep.)

Kézdi, Á.: **Handbuch der Bodenmechanik.** Bd. IV: Anwendung der Bodenmechanik in der Praxis. 295 S., zahlr. Abb. Berlin: VEB Verlag für Bauwesen 1976. Geb. ca. DM 63,30.

Nach 3 Bänden über Bodenphysik, Bodenmechanik im Erd- und Straßenbau sowie über Bodenmechanisches Versuchswesen, die sich als Handbücher und Nachschlagewerke sowohl bei den Studenten wie auch bei den im Beruf stehenden Ingenieuren einer großen Beliebtheit erfreuen, ist als Abschluß dieser Reihe der 4. Band erschienen, der sich mit der Anwendung der Bodenmechanik in der Praxis befaßt. An 28 Beispielen wird die Vorgehensweise zur Lösung geotechnischer Aufgaben vorgeführt. Ausgehend von den Feld- und Bodenuntersuchungen wird der Leser über die Versuche und Berechnungen nahtlos zu den sich hieraus ergebenden Entwurfskriterien und Sanierungsvorschlägen geführt. Die Beispiele sind breit gefächert ausgesucht und erstrecken sich vom Entwurf von Erdbauten, Bauwerksgründungen und Stützkonstruktionen bis zur Sanierung von Rutschungen und zu anderweitigen Schadensfällen. Mit den Abhandlungen gibt der Autor einen Einblick in seine langjährige und vielseitige Erfahrung sowie über seine Arbeitsweise als international geschätzter Gutachter. Damit empfiehlt sich der Abschlußband für Studenten als ein anschauungsreiches Lehrbuch und dem Ingenieur, der sich mit den Arbeitsweisen der angewandten Bodenmechanik vertraut machen will, als ein hilfreiches Nachschlagewerk. Gut ausgewählte Zeichnungen, Diagramme und Literaturhinweise ergänzen den Text und erleichtern den Gebrauch dieses empfehlenswerten Anleitungsbandes.

H. Breth, Darmstadt

Zeitschrift für Angewandte Geologie 6/1977 (German Dem. Rep.)

Kezdi, A.: **Handbuch der Bodenmechanik.** Band IV — Anwendung der Bodenmechanik in der Praxis. — VEB Verlag für Bauwesen Berlin/Verlag der Ungarischen Akademie der Wissenschaften Budapest 1976. 292 S., zahlr. Abb., M 48, —

Der vierte Band des Handbuchs der Bodenmechanik bietet nach den in den Bänden 1 bis 3 behandelten Grundlagen eine Vielzahl von Beispielen für

Begutachtungen aus dem Gesamtbereich des Bauwesens. Ein Beispiel ist Problemen gewidmet, die im Zusammenhang mit der Gewinnung von Braunkohle im Tagebau auftreten.

Mit diesem 4. Band wird ein Standardwerk abgeschlossen, zu dem man dem Verfasser und den Bearbeitern gratulieren kann. Es bietet Geotechnikern, Bauingenieuren und Bergleuten des Tagebaus das nötige Grundwissen, das sie zur Beherrschung ihres Fachgebietes benötigen. Als Nachschlagewerk, wie als Lehrbuch ist es hervorragend geeignet.

HANDBOOK OF SOIL MECHANICS

Новые книги за рубежом серия б 2/1975 (Soviet Union)

Kézdi A.: **Handbook of soil mechanics.** Soil Physics. Vol. 1. **(Справочное пособие по механике грунтов)** Том 1. Физика грунтов. Budapest, Akadémiai Kiadó, 1974, 294 p.

Рецензируемое справочное пособие состоит из четырех томов:

Том 1. Физика грунтов.

Том 2. Механика грунтов при земляных работах, основаниях и строительстве дорог.

Том 3. Лабораторные и полевые испытания грунтов.

Том. 4. Приложения механики грунтов в практике; примеры и описания аварий.

Настоящему изданию на английском языке предшествовали три венгерских и два немецких издания спраяочника.

Том 1 «Физика грунтов» содержит введение и девять глав.

В введении определено понятие термина «грунт» и указана область применения механики грунтов как науки, рассматривающей взаимодействие между грунтами и различного рода сооружениями (подпорными стенами, фундаментами, покрытием дорог и пр.). Следует отметить, что грунт в техническом смысле автор определяет как «верхние слои земной коры, которые поддерживают инженерные сооружения и которые являются их основанием, влияющим на их поведение». Это определение не выделяет в особую группу скальные породы, что противоречит как официальной

«Венгерской стандартой классификации геологических отложений MSZ 14045» (табл. 16), так и практике большинства зарубежных и советских специалистов.

Далее рассматриваются проблемы механики грунтов, к главнейшим из которых автор относит: проблему устойчивости, проблему деформаций оснований и проблему фазовых перемещений в грунтах.

В гл. 1 «Происхождение грунтов и их свойства» показано формирование грунтов как рыхлых горных пород (нескальных) в холмистой местности, седиментация грунтовых частиц под действием ветра (лёссы) и зависимость содержания глины в грунтах, образовавшихся из вулканических пород выветриванием, от средней годовой температуры пород. В табл. 2 свойства грунтов объединены три группы: I — размеры частиц и их свойства; II — фазовый состав грунтов; III (a, b, c)— прочностные, деформативные свойства и гидравлические характеристики грунтов. Во второй половине этой таблицы отмечены области инженерного использования отдельных свойств грунтов.

В гл. 2 рассматриваются составные части грунтов: твердая часть, вода и воздух, соотношение фаз и пределы консистенции грунтов. Следует заметить, что желательно было бы выделить из этой главы самостоятельный раздел, посвященный весьма важным для предварительной оценки строительных свойств грунтов характеристикам физического состояния грунтов: относительной плотности сыпучих грунтов и консистенции глинистых грунтов.

Далее охарактеризованы свойства воды в грунтах (молекулярная структура воды и льда, объем воды как функция температуры, ее кислотность, содержание солей, капиллярность с учетом изменения вязкости воды с изменением температуры и пр.), существенно влияющие на свойства грунтов.

Представлены также важные сведения о сжимаемости смеси воздух — вода (фиг. 46), соотношении фаз в грунте, а именно: об относительном объеме составных частей, природном содержании влаги в грунтах (рассмотрено и сезонное изменение влажности в грунтах до глубины 180 см), водонасыщенности, пористости и коэффициенте пористости, относительной плотности и степени уплотнения несвязных грунтов, имеющих важное значение для оценки физического состояния грунтов.

Последний раздел главы посвящен оценке консистенции глинистых грунтов: определению предела текучести (на приборе А. Казагранде), предела пластичности (элементарным методом почвоведов) и предела усадки. Даны многочисленные примеры и отмечена связь консистенции грунтов с их деформативными свойствами.

В гл. 3 «Структура грунтов» подробно описано строение сыпучих (гравий, галька, песок), связных (глинистых) и органических грунтов и

приведены важные данные о микроструктуре связных грунтов и образующих их минералов (строение атомной решетки грунтовых частиц, поверхностная активность минералов, форма частиц в зависимости от их минералогического состава и пр.). В конце главы показана роль органической составляющей в грунтах и вредных примесей в поровой воде.

В гл. 4 «Классификация грунтов» подробно рассмотрены: общая классификация Бюро общественных дорог США; классификация по крупности частиц, по пределам консистенции, стандартные классификации (Венгрии, США и др.). Важно отметить, что стандартная классификация США относится только к грунтам как рыхлым горным породам верхних слоев Земли; скальные же породы в ней не рассматриваются.

В гл. 5 «Напряжения в грунтах» внимание уделено эффективным и нейтральным напряжениям в грунте при установившемся движении воды (при просачивании), напряжениям при изменении объема грунта (компрессии с дренированием и без дренирования, при трехосном сжатии с учетом коэффициентов порового давления по Скемптону).

В дл. 6 «Движение воды в грунтах» обстоятельно описаны основные виды движения воды в грунтах: гравитационное движение, двумерное потенциальное течение, капиллярное движение воды, дренаж водонасыщенных грунтов, движение воды при консолидации грунтов и под действием электрического тока.

Детально рассмотрено гравитационное движение воды сквозь грунт: ламинарное (по закону Дарси) и турбулентное. Изложены методы определения коэффициента водопроницаемости (фильтрации) в лабораторных и полевых условиях и построение гидродинамической сетки движения воды в грунтах (при ограждении котлованов шпунтовыми стенками, под плотинами, в дамбах и пр.).

Приведены подробный вывод дифференциального уравнения одномерной консолидации (движения поровой воды под действием внешнего давления) и его решение в общем виде для случая равномерного, треугольного и параболического распределения уплотняющих давлений.

Далее в этой главе представлен вывод дифференциального уравнения трехмерной консолидации и дана ссылка на решение Карлсроу — Баррона и чихленный метод Скотта. Автор отмечает, что применение теории консолидации к определению скорости осадки сооружений будет рассмотрено в томе 2 настоящего справочного пособия.

Здесь уместно отметить, что применение метода конечных разностей к задаче трехмерной консолидации грунтов было разработано еще в 1937 г. и с успехом применялось советским ученым проф. В. А. Флориным.

Эта глава так же, как и последняя гл. 9 «Вопросы давления земли», изложены наиболее полно и интересно. В них приведены важные для практики расчетные показатели.

Гл. 7 «Прочность грунтов» в основном содержит теорию прочности Кулона — Мора, учитывается и нелинейность диаграммы сдвига связных грунтов; подчеркнута также необходимость учета эффективных напряжений. При рассмотрении сопротивления сдвигу связных грунтов показана нелинейная зависимость сопротивления сдвигу от тотальных и эффективных напряжений грунта. Сообщается также о применении (упорядочении) структуры глины в процессе длительного сдвига.

Следует отметить, что в следующем издании справочного пособия желательно было бы изложить и определение параметров сопротивления сдвигу связных грунтов по испытанию единичного образца грунта.

В гл. 8 «Деформации грунтов» включены краткие сведения о деформациях при ограниченном сжатии, реологических свойствах грунтов, сжимаемости при повторных нагрузках и внезапном сжатии (просадке, структурно-неустойчивых грунтов, например лёссовых). Интересные данные, подтверждающие порядок величин, полученных советскими специалистами, приведены для коэффициента Пуассона глинистных грунтов, который не является для них величиной постоянной, а изменяется в зависимости от влажности глин от 0,2 до 0,4 и в случае пластичных глин (деформирующихся без изменения объема) — до 0,5. Просадки грунтов рассматриваются в зависимости от структуры грунта и его состава, от величины действующего давления и относительной плотности грунта.

Гл. 9 «Вопросы давления земли» наиболее полно освещает рассматриваемую проблему на базе теории предельного равновесия сыпучих тел, использует классическое решение Кулона и его развитие Понселе, Ребханом, Жаки и др. На той же теоретической основе проанализировано давление при частично заданной криволинейности поверхности скольжения.

Далее указаны области использования теории давления земли в инженерном деле, хорошо изложено и понятие об активном, пассином и в состоянии покоя давлении земли, что иллюстрируется (гиг. 359) зависимостью между перемещениями подпорной стенки и соответствующим давлением земли. Составлена таблица значений коэффициента давления грунтов в состоянии покоя. Подробно изложено определение давления сыпучих и связных грунтов на наклонные подпорные стенки. Однако, по мнению рецензента, в формуле (274) для активного и пассивного давления не хватает третьего слагаемого, зависящего только от сцепления грунта и высоты подпорной стенки.

В конце этой главы помещена табл. 36 значений вертикальных и горизонтальных составляющих активного и пассивного давления земли по строгому решению проф. В. В. Соколовского. Таблица дана для трех значений угла наклона подпорной стенки к горизонту.

Рецензируемое справочное пособие представляет собой очень полное, весьма квалифицированное изложение физики грунтов применительно к проблемам механики грунтов и является хорошим пособием для студентов, изучающих механику грунтов, и инженеров, применяющих ее на практике. Перевод его на русский язык будет весьма полезным.

Чл.-корр, АН СССР
Н. А. Цытович

Книга получена редакцией журнала

Новые книги за рубежом
Серия б 10/1980 (Soviet Union)

Kézdi Á.: **Handbook of soil mechanics.** Vol. 2. Soil Testing. **Руководство по механике грунтов.** Том 2. Определение свойств грунтов. Budapest: Akadémiai Kiadó, 1980, 258 p.

Известный венгерский ученый проф. А. Кезди издал обстоятельное руководство по механике грунтов в четырех томах. Сначала оно вышло в свет на немецком языке, а сейчас переводится на английский. Английский вариант включает: том 1 «Физика грунтов» (1974 г.), том 2 «Полевые и лабораторные исследования свойств грунтов» (1980 г.), том 3 «Механика грунтов при устройстве выемок и оснований в дорожном строительстве», том 4 «Применение механики грунтов в практике строительства»[1].

Рецензируемый том 2 состоит из введения и пяти глав. Во введении указаны задачи, связанные с исследованиями грунтов, описано, как следует проводить испытания грунтов в лабораторных и полевых условиях с целью определения многочисленных показателей, характеризующих их свойства. Отмечено, что для решения инженерных проблем, связанных с грунтами оснований и грунтами, служащими материалом для сооружений, необходимы изучение грунтовых условий, отбор образцов и, наконец, лабораторные и полевые определения физических свойств и их характеристик, позволяющие предопределить поведение грунтов в тех или иных новых условиях, в которых они будут находиться. Именно последней задаче и посвящен настоящий том 2 руководства.

В гл. 1 «Разведочные изыскания» содержатся полезные таблицы, в которых указано, когда, где и какие определения следует делать, методы разведочных изысканий разделены на косвенные (геофизические, зондирование), полупрямые (бурение малым диаметром, отбор образцов и обобщение разрозненных материалов) и прямые (испытания непосредственно в котлованах и шурфах). Описаны применяемое буровое оборудование, грунтоносы, установки для статического и динамического зондирования.

В самой большой в книге гл. 2 «Лабораторные исследования» рассмотрены используемые аппаратура и оборудование, изложена методика определения различных свойств, а также обработки результатов опытов. Обстоятельно описаны следующие испытания: определение влажности, объемного и удельного весов, ситовой анализ и ареометрический способ определения гранулометрического состава мелкодисперсных грунтов, определение пределов Аттерберга с анализированием возможных погрешностей, определение водонасыщенности, консистенции, водопроницаемости.

В гл. 3 «Гидрогеологические исследования» внимание также уделено опытным откачкам из совершенных и несовершенных скважин. Приведены формулы для расчета при неоднородных толшах. В гл. 4 «Определение несущей способности при помощи пробной нагрузки» описаны испытания в полевых условиях пробной нагрузкой металлическими штампами, устанавливаемыми в котлованах и скважинах, а также статические испытания свай. Заключительнав гл. 5 «Исследования на стройплощадках» посвящена контролю за уплотнением грунтов, радиометрическим испытаниям, определению порогового давления дистанционными поропьезометрами и др.

В книге обобщено значительное количество работ по определению свойств грунтов, выполненных самим автором и учеными других стран. Она содержит много фактического и иллюстративного матернала. Книга проф. А. Кезди представит несомненный интерес для наших читателей, и ее можно рекомендовать для перевода на русский язык.

Д-р техн, наук, проф.
М. В. Малюшев

Книга получена редакцией журнала

The structural engineer 8/1982 (England)

Handbook of soil mechanics. Vol. 2. Soil testing. *Árpád Kézdi (Amsterdam: Elsevier, 1980) 258pp, $70.75, Dfl. 145. ISBN 0 444 99778 4. Soil testing* is the second of four volumes that will comprise Professor Kézdi's *Handbook of soil mechanics.* The first volume dealt with soil physics.

This second volume covers almost the entire field of soil investigations and is intended for use as a textbook or manual for laboratory and field testing of soils for civil engineering purposes. In the first part of the book, the author describes laboratory tests based on internationally accepted methods. Particularly detailed descriptions of the testing methods are given, as well as an analysis of possible sources of error, as reliability of laboratory work depends on painstaking and precise performance of individual investigations. The second part of the book covers the most important field investigations where, depending on the specific objectives, more freedom in the performance of the tests is permissible.

For the benefit of laboratory personnel, every aspect of each individual test is described, i.e. definition, test equipment, preparation and performance, data processing, numerical example, process error. The text is supplemented throughout by numerous diagrams and tables and an added facility is the provision of laboratory sheets for measured data, furnished with examples. Soil engineers and technicians, engineering geologists and civil engineers will all find this book an invaluable guide.

Geotechnique 2/1981 (England)

Handbook of soil mechanics, Volume 2: soil testing. Árpád Kézdi. Published jointly by Elsevier Scientific Publishing Company, Amsterdam, and Akadémiai Kiadó, Budapest. 1980. 206 pages. 308 line drawings, 37 half tones, 34 tables. $70.75.

The speed with which geotechnology spread throughout the world undoubtedly contributed greatly to the universal acceptance of a similar procedure for most soil tests on which this book is based. There are sound arguments for the International Society to make this more secure following the lead it has given with the proposed European standards for penetration testing.

Professor Kézdi's handbook will be in four interrelated volumes, which commence with a volume on soil physics where the theory for the tests is given. This second one, translated by P. Szőke, represents a revision of Volume III of the German edition published in 1973 (reviewed in *Géotechnique* **30,** 469). Detailed descriptions are given of some 60 different laboratory and field tests, together with an account of boring and sampling techniques. Particular attention is given to the test procedure, without reference to Standards, and the precautions to be taken, often with an analysis of possible errors. The Author is also not afraid to point out the disadvantages in particular tests and suggests permissible numeric tolerances in the results of parallel tests. Worked examples with fullpage test sheets occur frequently.

In the main part on laboratory work, the tests are dealt with under the following headings: solids tests (density and distribution), phase composition, Atterberg limits, behaviour against water, chemical tests, effect of moving water and force effects. The importance given to phase composition, i.e. the volumetric proportions in a representative sample of solid, liquid and gas, for classification rather than the liquid and plastic limits, makes the handbook especially useful when it is necessary to interpret partially saturated soils. For this purpose separate tests are given, such as water absorption capacity, slaking, capillary rise and specific collapse coefficient (due to flooding), in addition to examination of the results. Examples describe the Loess soil. A separate, concise, yet thorough description is given by Dr L. Varga of the triaxial compression test as developed at

Imperial College, consisting of load frame, cell, compensated mercury pots for pressure stabilization, manual volume change and mechanical measurement. Detailed diagrams clearly illustrate the significant individual parts and how they are assembled. Testing procedures are given for all the established effective strength tests, as well as an outline of stress path testing and the tensile strength of cohesive soils.

Testing in the field is sub-divided into the investigation of groundwater, load-bearing capacity and earthworks. The first named is a selfcontained section that describes both the tests and their analysis with interesting illustrated case histories of long term hydrological conditions in Hungary. Plate loading and pile tests follow the traditional methods. Earthworks include measurements by radioactive isotopes and touch upon the instrumentation of earthworks.

Soil and tillage research 2/1982 (Netherlands)

Handbook of Soil Mechanics. Vol. 2 Soil Testing. Árpád Kézdi. Elsevier Scientific Publishing Company, Amsterdam/Oxford/New York, jointly with Akadémiai Kiadó, Budapest, Hungary, 1980. 258 pp., 345 figs., 34 tables, US$61.75/ Dfl.145.00, ISBN 0-444-99778-4.

Árpád Kézdi's well-known handbook on soil mechanics is being published in the English language. Recently, the second volume dealing with soil testing appeared. The other volumes are:

Vol. 1 Soil Physics;

Vol. 3 Soil Mechanics of Earthworks, Foundations and Highway Engineering;

Vol. 4 Application of Soil Mechanics in Practice, Examples and Case Histories.

The volume on soil testing is a translation and revision of a German version of a Hungarian book on soil mechanics practices. It describes the classical soil measurements used in civil engineering as routine tests. Many measurements in agricultural soil mechanics relate to or have been derived from such methods. Details of the civil engineering methods presented vary slightly between countries and regions but the book refers mainly to standards accepted in Eastern Europe. In general, however, regional variations are small enough for the world-wide significance of this book to be recognized.

The methods are presented in five chapters. "Soil exploration" comprises exploration by pits and shafts and by drilling and sounding. Taking undisturbed and disturbed samples is included. The reader will find valuable general remarks on soil exploration. Kézdi's classification of constructions into seven categories, and of subsoil conditions into four categories is interesting. For each construction and subsoil condition category the desired extent and character of soil exploration is proposed in a table. The chapter on "Laboratory investigation" forms the main part

of thee book. First, the character, order and extent of the required laboratory tests are discussed, together with their application. This discussion is based on the construction and subsoil classification mentioned. Then the following tests are described: measurements concerning density, moisture content and phase distribution; grain size distribution; determination of the Atterberg limits; compactibility; water permeability and capillary rise; compressibility; direct shear, unconfined compression and tri-axial tests as well as tensile strength measurements; determination of organic matter, lime and soluble sulfate contents. In the chapter on "Investigation of groundwater" measuring methods for pressure, level and flow of groundwater are found and field determination of the coefficient of permeability is described. Load tests in the field are discussed in a chapter on "Investigation of the load bearing capacity". The final chapter deals with "Earth work investigations" and describes the most important routine tests for embankment and earth dam constructions. These include the determination of phase distribution, compactness, shear strength, and pore-water pressure. Subsequently, some experiments suitable for the evaluation of the soil as a road-foundation material are explained, for instance the CBR test. Results of earthwork quality control methods should be available for application as soon as possible. Therefore, mobile laboratories are also considered in the book.

In general, the book presents a test as follows. At first, a definition of the test and a short theoretical background are given. Then, the necessary equipment and aids are listed and extensively described. Most attention is paid to the test preparation and test performance. Finally, processing of the results is dicussed. In fact, the book shows the reader how to do the test. This is greatly enhanced by the presentation of a numerical example on the usual data sheets. If possible, the order of magnitude and range of the measuring data to be expected are given, and the sources of error are discussed. Correlations with other mechanical or physical properties are included.

The descriptions are very extensive and complete. The book is copiously illustrated with drawings, graphs and photographs. A limited amount of literature is listed. Text parts not necessary for the "main line" are printed in small type. The book may be considered as an integral part of Kézdi's standard work on soil mechanics but it may also be used as an independent work. It can well be recommended when an extensive and complete documentation on soil testing is desired or when the aim is to introduce or to improve testing where experienced laboratory personnel is not available.

A.J. KOOLEN

Bauplanung – Bautechnik 2/1981 (German Dem. Rep.)

Handbook of Soil Mechanics. Band 2 Soil Testing. Von *A. Kézdi.* Akadémiai Kiadó, Budapest 1980. 20,5 cm × 29,5 cm, 260 Seiten, 345 Bilder und 34 Tabellen. Gebunden 600,— Ft.

1973 erschien zum ersten Mal das Buch „Bodenmechanisches Versuchswesen" von *A. Kézdi* als Band 3 seines „Handbuches der Bodenmechanik". Diese erste Ausgabe, eine Gemeinschaftsausgabe des Akadémiai Kiadó, Budapest, und des VEB Verlag für Bauwesen, Berlin, erschien in deutscher Sprache und war durch die Bearbeitung von Prof. *W. Kinze,* Dresden, den Bedingungen in der DDR durch die Einbeziehung der geltenden Standards und Vorschriften weitgehend angepaßt. So ist es nicht verwunderlich, daß das Buch allen auf diesen Fachgebiet arbeitenden Kollegen ein unentbehrlicher Ratgeber sowohl für die Durchführung von Routineuntersuchungen und das ständige Bemühen um die Erhöhung ihrer Aussagekraft als auch für die Planung, Vorbereitung und Realisierung seltener angewandter Versuche, insbesondere der Feldversuche, geworden ist.

Sieben Jahre später legt der ungarische Verlag dieses Buch als Band 2 des „Handbook of Soil Mechanics", dieses Mal in Kooperation mit dem in Fachkreisen bekannten Elsevier Scientific Publishing Company, Amsterdam, vor.

Die Neuauflage hat die Gliederung des Originals

Abschnitt 1: Bodenuntersuchungen,

Abschnitt 2: Laboruntersuchungen,

Abschnitt 3: Untersuchungen des Grundwassers,

Abschnitt 4: Tragfähigkeitsuntersuchungen,

Abschnitt 5: Untersuchungen für Erdarbeiten

beibehalten. Damit werden im 1. Teil die Grundlagen und die Ausführung von Laboruntersuchungen, im 2. Teil die Feldversuche behandelt.

Der Verfasser hat sich dabei unter Mitwirkung bekannter ungarischer Fachkollegen erfolgreich bemüht, die doch recht umfassenden, durch neue theoretische Erkenntnisse und technisch-technologische Weiterentwicklungen bedingten Fortschritte der Versuchsdurchführung und auswertung zu erfassen und einzuarbeiten. So kann das Werk gleichzeitig als eine Dokumentation der Entwicklung des bodenmechanischen Versuchswesens in den letzten Jahren gewertet werden.

Der Verfasser, allgemein bekannt für sein Bestreben, die Probleme durch die Analyse der Grundlagen und iöhres jeweiligen Zusammenwirkens zu klären, legt in seinen Ausführungen besonderen Wert auf die Darstellung des Einflusses einer verantwortungsbewußten Arbeit aller an der Versuchsdurchführung und -auswertung Beteiligten auf die Lösung der Probleme, die aus der Inhomogenität des Bodens und den Schwierigkeiten der Probenentnahme resultieren.

Die neue Auflage des „Bodenmechanischen Versuchswesens" innerhalb des „Handbuches der Bodenmechanik" von *A. Kézdi* ist also in jeder Beziehung den Fachkollegen sowie den an dieser Thematik interessierten Studierenden wärmstens zu empfehlen. Ihr Studium setzt jedoch die Kenntnis der englischen Sprache voraus.

Welzien

Inżynieria i Budownictwo 8/1980 (Poland)

Kézdi A.: **Handbook of Soil Mechanics.** Vol. 2. **Soil Testing (Podręcznik mechaniki gruntów.** Tom. 2. **Badanie gruntów).** Wyd. Akadémiai Kiadó, Budapest; Elsevier Scientific Publishing Co., Amsterdam, str. 258, 345 rys., wykresów i fotografii, 34 tabl., 30 wzorów metryk i formularzy, 1 nomogram poza tekstem.

Czterotomowy podręcznik mechaniki gruntów jest jednym z wielu dzieł znanego specjalisty w tej dziedzinie Arpada Kézdi, Profesora Politechniki Budapeszteńskiej. Podręcznik ten (oprócz wydań węgierskich) został wydany w języku niemieckim. Obecnie ukazał się drugi tom w języku angielskim (tom 1 był recenzowany w *Inżynierii i Bud.* w nr 12/74), przy czym kolejność tomów zostala zmieniona — odpowiada on trzeciemu tomowi wydania niemieckiego z 1973 r. W porównaniu do poprzednich wydań tekst uległ modernizacji i rozszerzeniu: m.in. wprowadzono opisy nowych metod badań, rozbudowano rozdział poświęcony badaniu wód gruntowych itd.

Pierwszy rozdział omawia terenowe metody badań gruntów, używany przy tym sprzęt (jak świdry, sondy itp.), sposoby oceny wyników sondowań. Najobszerniejszy jest rozdział drugi, poświęcony zagadnieniom związanym z laboratoryjnymi badaniami gruntów, a przede wszystkim możliwościom, zaletom i wadom tych badań. W rozdziale trzecim rozpatrzono badania warunków wodnych w podłożu gruntowym, a więc rodzajów wód gruntowych, ich poziomów i napięć, przepływów, przepuszczalności warstw róznych gruntów itd. Rozdzial czwarty omawia próbne obciążenia podłoża gruntowego in situ: powierzchniowe, w otworach badawczych, za pośrednictwem pali. Rozdział piąty poświęcono badaniom budowli ziemnych (w tym nasypów): ich zagęszczenia, przepuszczalności itp.

Przydatność praktyczną książki dla wszystkich zajmujących się badaniami gruntów lub korzystających z wyników takich badań należy ocenić bardzo wysoko. Daje ona szeroką panoramę metod stosowanych w różnych krajach, pozwala na krytyczną ocenę wyników badań. Pewną wadę stanowi częsty brak wyraźnego podkreślenia różnic między normami, obowiązującymi w różnych krajach oraz skutków tych różnic. Część zagadnień tego rodzaju zostala wyjaśniona w tomie I, który nie jest jednak dostępny dla wielu użytkowników tomu II, między ich wydaniem upłynęło 6 lat. Układ merytoryczny książki jest przejrzysty, język prosty i jasny.

Podsumowując należy stwierdzić, że recenzowana książka stanowi cenną pozycję również dla czytelnika polskiego, stanowiąc prawdziwe kompendium wiedzy na temat badań podłoza gruntowego dla budownictwa.

Szata edytorska książki jest staranna, zbliżona do wydania niemieckiego. Jedyną wadą jest zmiana rodzaju oprawy (m.in. w wyniku usunięcia obwoluty) w stosunku do tomu I.

Dr inż. Roman Czarnota-Bojarski

Pozemni stavby 9/1980 (Czechoslovakia)

Přiručka mechaniky zemin (Handbook of Soil Mechanics). Akadémiai Kiadó, Budapest 1980, 2. dil (Soil Testing) 253 stran, 30 tabulek, 345 obrázků. V angličtině. Lze objednat v prodejně knih SNTL — Nakladatelstvi technické literatury, Praha 1. Spálená 51.

Člen korespondent maďarské akademie věd, prof. Árpád Kézdi, předni evropský odbornik v oboru mechaniky zemin, jehož „Přiručka mechaniky zemin" je v Československu známá z předchozic vydáni zejména v němčině, vydal v tomto roce ve společném nákladu Akadémiai Kiadó Budapest a Elsevier Scientific Publishing Company Amsterodam 2. díl této příručky v angličtině. Proti poslednímu vydání v roce 1973 je zejména druhá a třetí část tohoto dílu příručky doplněna o zkoušky zemin v trojosém přistroji včetně popisu vyhodnocení zkoušek a o některé problémy průzkumu podzemních vod.

Publikace je rozdělena do pěti částí. V prvni části je kniha zaměřena na metody používané při polních zkouškách, způsoby odběru neporušených vzorků a metody používané při penetračních zkouškách.

Těžiště knihy je ve druhé části, která je i obsahově nejrozsáhlejší. Je zaměřena na podrobný popis mezinárodně používaných laboratorních zkoušek zemin, jejichž přesným prováděním a vyhodnocením jednotnou metodiku lze pro potřebu inženýrské praxe charakterizovat individuální vlastnosti zemin. Právě s ohledem na praktické potřeby se věnuje autor v této části velmi podrobně popisu zkušebních metod a zejména rozboru zdrojů možných chyb. Tato část obsahuje kromě popisu postupu při zkouškách porušených i neporušených vzorků nutných ke stanovení základních fyzikálních vlastností zemin jako je měrná i objemová hmotnost, pórovitost, zrnitost, vhlkost obsah organických substanci atd., komplexní popis všech laboratorních zkoušek v oboru mechaniky zemin používaných.

Ve třetí části knihy se autor věnuje problematice hydrogeologického průzkumu, tlakovým problémům podzemnich vod, popisu čerpacích zkoušek včetně úprav studní vhodných pro čerpací pokusy, podrobně je popsáno polní stanovení koeficientu propustnosti.

Čtvrtá část publikace obsahuje popis a hodnoceni metod používaných při zkouškách únosnosti zemin. Rozsáhlý popis zkoušek pravoúhlými a kruhovými zatěžovacími deskami se stanovením modulů stlačitelnosti je doplněn stručným popisem zatěžovacích zkoušek pilot.

Ve všech dílech knihy jsou uváděny příklady záznamů výsledků zkoušek ve zkušebnich protokolech, číselné hodnoty uváděné v příkladech, v tabulkách a obrázcích jsou již výlučně v jednotkách SI. Těžište práce je ve druhé části, tj. v laboratorní technice zkoušek mechanicko-fyzikálních parametrů zemin a jejich vyhodnocení. Zejména tato část bude významnou pomůckou pracovištím, která se geologickým průzkumem nebo hodnocením výsledků průzkumu zabývají, tím

spíše, že je postup vyhodnocování zkoušek téměř ve všech případech demonstrován na příkladech.

Precizní zpracování tematiky, úplnost a šíře popisované problematiky z knihy vytváří nejen „Průvodce" a učebnici, ale nezbytnou součást knihovny všech odborných pracovišť. Mimořádně kvalitní tisk knihy dává záruky dlouholetého průvodce každého odborníka z oboru mechaniky zemin.

Ing. Jiří Chlost, CSc.

Inżynieria Budownictwo 1/1980 (Poland)

Kézdi A.: **Soil Physics. Selected Topics.** Str. 160, rys. 215. Akadémiai Kiadó, Budapest 1979.

Książka omawia wybrane zagadnienia gruntoznawstwa inżynieryjnenego, związane na ogół z tematyką dotychczasowych prac Autora i niektórych innych autorów węgierskich. Zainteresowania Autora określają również zakres omawiania poszczególnych zagadnień, nie zawsze zgodny z ich naukowym czy też praktycznym znaczeniem.

We wstępie (str. 11 ÷ 25) poruszone zostały bardzo różnorodne i mało ze sobą związane zagadnienia, od problemów klasyfikacji gruntów do ogólnego omówienia zagadnień ich reologii. W części drugiej (str. 26 ÷ 61) Autor omawia skład gruntów pod względem objętości ziaren i cząstek o różnych wymiarach, rozkład porów w zależności od ich rozmiarów oraz wpływ uziarnienia (p. 2.3 oraz 2.5) na niektóra właściwości gruntów, m.in. na ich zagęszczalność; omówione zostały również niektóre właściwości 2 ÷ 4-składnikowych mieszanin gruntów (p. 2.4).

Rozdział 3 (str. 62 ÷ 95) poświęcony został wybranym problemom wytrzymałości, zarówno gruntów niespoistych (p. 3.1 i 3.2), jak i gruntów spoistych (p. 3.3 oraz 3.5); omówiony został przy tym pokrótce również problem wytrzymałości gruntów na rozciąganie.

Najobszerniejszy jest rozdział 4 (str. 96 ÷ 153) omawiający przepływ w gruntach wody i powietrza, z uwzględnieniem pewnych zjawisk o znaczeniu praktycznym (sufozja i erozja, str. 119 ÷ 137). Omówiony został również przepływ w gruntach o stopniu wilgotności mniejszym id jedności; przemieszczaniu się nastomiast wody w gruntach spoistych poświęcono jedynie 3 strony (w tym 5 rysunków).

Trudno oprzeć się wrażeniu, że Autor przy opracowywaniu omawianej książki w dość specyficzny sposób wykorzystał dorobek światowej literatury geotechnicznej. Na ogólem 88 pozycji wykazu bibliografii (str. 155 ÷ 158) tylko 2 pozycje są w języku rosyjskim, z literatury zaś francuskiej uwzględniony został jedynie podręcznik mechaniki gruntów A. Caquota i J. Kérisela; znaczna część ponadto pozycji to nie prace badawcze, lecz opracowania o charakterze

podręczników lub monografii. Z polskich prac nie została uwzględniona ani jedna, mimo istotnóenego dorobku naukowego w zakresie niektórych zagadnień, np. w zakresie reologii gruntów.

Powyższe m.in. względy powodują, iż mimo przystępnego charakteru książka może być polecana raczej dla osób mających już pewne przygotowanie w zakresie geotechniki; w dużo mniejszym natomiast stopniu może być przydatna dla geotechników początkujących. Mniej zorientowany czytelnik mógłby na przyklad odnieść wrażenie, że pomiędzy niektórymi paramerami gruntów istnieją ścisłe zależności, podczas gdy w rzeczywistości mają one charakter jedynie luźnych często korelacji. Wiele interesującego materialu zawartego w omawianej książce powoduje jednak, że może być ona cenną pomocą dla osób specjalizujących się w odnośnych dziedzinach geotechniki.

Prof. dr hab. Antoni Piaskowski

Inženýrské stavby 11/1979 (Czechoslovakia)

Árpád Kézdi, DrSc.: **Soil Physics** (Fyzika zemín.) Výdalo nakladatelstvo Akadémiai Kiadó, Budapest 1979, 160 str., 215 obr., 6 tab.

Kniha patrí k najzaujímavejším a najoriginálnejším teoretickým geotechnickým publikáciám, ktoré sa dostali na európsky knižný trh v tomto roku. Nejde o prácu príliš rozsiahlu, ale mimoriadne pozoruhodnú svojím obsahom.

V štyroch kapitolách autor prináša nové poznatky z fyziky zemín a spresňuje naše poznatky o zeminách ako o trojfázovom systéme, v ktorom tzv. zákony: Darcyho, Hooka i Mohrova-Coulombova teória porušenia len aproximatívne vystihujú ich správanie sa — závislé v značnej miere od vzájomného pomeru fáz. Podrobná analýza v prvej kapitole ukazuje, že pri posudzovaní stavu a vlastností pórovitého prostredia obsahujúceho vodu a vzduch treba skúmať viac charakteristík, ako obvykle uvádzame pri pôdomechanickom — fyzikálnom rozbore. Účelné je rozšíriť ich základné rozdelenie a opatrne narábať s tzv. modelmi zemín. Odporúča sa skúmať „multilaterálne" vzťahy a premietať ich do závislosti, ktoré používame pri výpočte pretvárania únosnosti, priepustnosti a konsolidácie, a pomocou nich vysvetľovať správanie sa zemín pri ich rôznom zaťažení. Tento nový prístup dokumentuje na odvodení deformačnej rovnice, do ktorej zavádza stlačiteľnosť tak fázy kvapalnej, ako aj tuhej. Stlačenie nerobí teda závislým iba od zmeny objemu pórov, ako je to pri konvencionálnych výpočtoch obvyklé. Pracuje aj s novými pojmami — ako je napríklad potenciál pohybu, ktorý má lepšie vystihnúť špeciálne fyzikálne vlastnosti trojfázového systému.

V druhej kapitole autor podrobne rezoberá krivky zrnitosti zemín, všíma si tvar ich zŕn, ich špecifický povrch, ako aj pomer jemných zŕn a skeletu. Pri nesúdržných zeminách berie do úvahy ich zhutniteľnosť a stabilitu za priesaku. Pri jemnozrnných (súdržných) materiáloch venuje pozornosť predovšetkým pórovým tlakom, stupňu nasýtenia kapilárnej zóny. Dokazuje, že pri týchto zeminách práve

uvádzané charakteristiky spolu s pórovitosťou rozhodujú aj o tixotropických vlastnostiach trojfázového systému (o schopnostiach vytvárat suspenzie), čo ovplyvňuje aj intenzitu sedimentácie zŕn rôzneho špecifického povrchu.

Problémy napätí — je názov kapitoly 3, venovanej problémom porušenia a zemným tlakom z nových pohľadov na zeminu ako trojfázový systém, ktorého napätosť a pevnostné charakteristiky závisia od pórovitosti a od zastúpení jednotlivých fáz (tuhej, kvapalnej, a plynnej), ovplyvňujúcich súdržnosť (c), ako aj uhol vnútorného trenia (ϕ), pre ktoré sú uvedené viaceré vzťahy a diagramy osvetľujúce túto závislosť. Podmienky napätosti sa rozoberajú so zreteľom na vplyv dilatancie a veľkosť deviátora napätia, ktorý ovplyvňuje nielen stabilitu zemín, ale aj ich spracovateľnosť.

Kapitola 4 sa zapodieva pohybmi fáz, vychádzajúc zo zákonitosti kvapalnej fázy v súhlase s Darcyho zákonom s uvážením hodnoty Reynoldsovho čísla a veľkosti súčiniteľa hydraulického trenia. Pohybu vody v pieskoch, ktorých stabilita je týmto pohybom obvykle najväčšmi ohrozená, je venovaná osobitná časť tejto kapitoly. V nej sa vysvetľuje podstata a príčiny rozdielov priepustnosti, stability zemín, vzniku výverov, tvorenie bublín v súvislosti so zrnitosťou, pórovitosťou a podielom jednotlivých fáz zeminy. Nemalý význam má obsah ílovitých zemín a ich mineralogické zloženie. Pritom však hlavnou charakteristikou ovplyvňujúcou stabilitu zemín za priesaku zostáva ich zloženie, pórovitosť a hydraulický gradient prúdiacej vody na druhej strane.

Prehľadný zoznam literatúry, autorský a vecný register doplňujú túto veľmi cennú publikáciu, po ktorej siahajú najmä tí odborníci, ktorí chcú hlbšie vniknúť do problémov mechaniky zemín, ako aj ti odborníci — najmä projektanti — ktorí majú záujem o spresňovanie výpočtov a zdokonaľovanie navrhovania stavieb. Kniha by sa mala dostať do rúk každému špecialistovi mechaniky zemín, vedeckému pracovníkovi, všetkým diplomantom a ašpirantom, ktorí v nej nájdu odraz najnovších poznatkov fyzikálnych vlastností zemín, podaných zrozumiteľnou formou.

Prof. P. Peter, DrSc.

Строителство
1/1980 (Bulgaria)

Árpád Kézdi. **Soil physics. Selected topics (Физика на почвите. Избрани проблеми).** Akadémiai Kiadó, Budapest, 1979, 160 с., 215 фиг. 5 табл., 88 лит.

Изучаването на физическите свойства на почвите стана необходимо пред вид все понарастващите проблеми, свързани с изграждането на инженерни съоржения. Действието на огромни динамични и статични товари върху земната основа на много големи и тежки съоръжения значително разшири областта на приложение на земната механика. Тези

проблеми изискват задълбочен анализ на физическите свойства на изходните материали. Това доведе до подем на изследователската работа в областта на физиката на почвите, т. е. физиката на зърнестите материали, в целия свят: открити бяха нови закони и зависимости, които улесняват количественото изразяване на поведението на почвите. Понастоящем теориите, прилагани при фундирането и земните работи, се базират главно върху опростени приемания и модели. Обикновено те не отчитат обстоятелството, че почвите се състоят от три материала с различни състояния на материята и разпръснати в една сложна система. Законите на Дарси, Хук, Моор, Кулон и др. отразяват с много грубо приближение поведението на зърнестите материали. До голяма степен съвременната изследователска работа в областта на физиката на почвите цели да се избягнат тези приближения. Обаче резултатите от тези изследвания са все още слабо отразени в теориите и в практиката. Целта на автора, известен учен в областта на земната пеханика и член на Унгарската академия на науките, е да допринесе за подобряване на това положение на нещата и за изясняване на свойствата на зърнестите материали.

Както показва заглавието, книгата включва някои избрани проблеми, като се базира на изследванията, проведени в лабораторията на катедрата по геотехника при университета в Будапеща.

Авторът класира почвите в три главни групи: пясък, преходни почви (ситен пясък, каменно брашно, прах) и глина. Освен това той класира почвите според основните им механични свойства: якост, деформация и движение на фазите.

Книгата се състои от три части. Първата част включва понятията зърна и зърнести агрегати и свързаните с тях явления. Разпределението на едрите зърна се разглежда според техния обем, тъй като според автора традиционната крива на зърнометричния състав не отразява достоверно размера на зърната поради различната им форма. По-нататък се разглеждат разпределението на размера на порите в ситнозърнестите материали, зависимостта между зърнометричен състав и уплътняемост, нови изследвания върху свойствата на смеси от различни зърнести материали, уплътняемост на „преходните почви.

Втората част включва проблеми на якостта на пясъците, които се разглеждат като несвързани или свързани почви в зависимост от плътността и степента на водонасищане и на „преходните почви, върху които играят голяма роля капилярните сили. Авторът отделя специално внимание на якостта на опън и на срязване на свързаните почви.

В третата част авторът разглежда с примери от лабораторни изследвания някои случан на движение на фазите с цел да се постигне по-добро разбиране на това явление. То може да се класира по различни признаци, а именно въз основа на силовите полета, предизвикващи движението или

средата, в която става движението. Най-важните силови полета са гравитация, капилярност, вакуум, термоосмоза и електроосмоза. Разглежда се движението на водата в наситен пясък, пропадане под действието на водата, суфозия, ерозия, движение на фазите в трифазна среда, движение на фазите в свързани почви. Книгата е съвместно издание на английски език на Издателството на Унгарската академия на науките Akadémiai Kiadó и на известното холандско издателство на научна литература Elsevier Scientific Publishing Company.

Инж. Г. Понов

Die Eisenbahntechnik 6/1979 (German Dem. Rep.)

Soil Physics Selected Topics (in engl. Sprache). Von *Árpád Kézdi,* Budapest: Akadémiai Kiadó 1979. 160 Seiten, 215 Bilder, Format 16,7 cm × 24,8 cm, Broschur. ISBN 963 05 1478 8

Als Ergänzung und Fortsetzung zum vierbändigen, in mehreren Sprachen erschienenen „Handbuch der Bodenmechanik" gibt der Autor in dem vorliegenden Werk eine Analyse zum wissenschaftlichen Erkenntnisstand in der Bodenphysik sowie Impulse für künftige Forschungsarbeiten auf diesem Gebiet. Das Anliegen wird besonders im 1. Kapitel (Einführung) umfassend dargelegt. Aufgezeigt wird hierbei, daß die Weiterentwicklung von Berechnungsmethoden der Bodenphysik nur dann möglich ist, wenn die bisher bekannten bodenphysikalischen Teilergebnisse durch ein einheitliches Bild gefaßt werden. Dies gilt für die drei fundamentalen Erscheinungen „Festigkeit, Deformation, Phasenbewegung" gleichermaßen.

Mit den Kapiteln 2 bis 4 (Körner und Körnerhaufen, Fragen der Festigkeit, Einige Fälle der Phasenbewegung) stellt der Autor ausgewählte Abschnitte der Bodenphysik unter diesem Aspekt vor. Neue Forschungsergebnisse aus dem Geotechnischen Laboratorium der Technischen Universität Budapest und zahlreiche Ergebnisse anderer Forschungseinrichtungen werden einbezogen.

Unter dem Titel „Fragen der Bodenphysik" erschien das vorliegende Buch 1976 in deutscher Sprache. Es wurde mit einer ausführlichen Rezension von *Bobe* im Heft 7/1977 der Zeitschrift „Bauplanung — Bautechnik" bereits vorgestellt und zur vielseitigen Anwendung empfohlen. Vorzüge der englischen Ausgabe gegenüber der deutschsprachigen bestehen darin, daß SI-Einheiten und Symbole entsprechend den internationalen Empfehlungen herangezogen werden. Mit dem Buch „Soil Physics" („Fragen der Bodenphysik") besitzt vor allem der in der Forschung tätige Ingenieur einen wertvollen Leitfaden, der Wege und Methoden zum Auffinden bodenphysikalischer Gesetzmäßigkeiten aufzeigt. Besonders zu

empfehlen ist diese Literatur als spezielles Lehrmaterial zur Aus- und Weiterbildung von Fachingenieuren für Baugrundfragen sowie für Diplomanden des Fachgebiets Geotechnik. Auf diese Weise läßt sich ein wesentlicher Grundgedanke von *Á. Kézdi* verwirklichen: die rasche Praxiswirksamkeit wissenschaftlicher Erkenntnisse.

Helga Hubáček

Geotechnique 12/1980 (England)

Soil physics (selected topics). Arpad Kezdi. Elsevier Scientific Publishing Company. 1979. US$47.80.

Like Prince Metternich and Dorothea Lieven, Soil Mechanics and Soil Science have carried on a mild flirtation for upwards of 30 years; meeting infrequently but each clearly intrigued by the other and wondering what exactly is on offer. The title of this book may give the impression that an eminent soil engineer is once again tying to bring the bashful couple together. This is not the case.

One senses that Professor Kezdi shares the feeling of many readers of *Géotechnique* that mathematics has been so closely harnessed to the necessarily experimental approaches used in soil mechanics that there is little room between for any serious consideration of the physical principles involved.

The book has been written primarily for students at the University of Budapest, and the soil physics is that of the Bernard Keen era rather than the purely thermodynamic approach pioneered by Kenworthy Schofield, although energy concepts are briefly discussed in the last chapter.

The first third of the slim volume deals with particle size distribution and particle shape and their influence on compactibility. The treatment is not new but it is assembled in an attractive and highly readable manner.

In considering the strength of soil, Professor Kezdi considers first granular and then cohesive soils. The soil physicist might prefer to reverse this treatment and first deal more fully with the relationships between strength, pore water pressure and suction for cohesive soils. However, with admirable economy of space the author summarizes current thinking on shear strength and critical state soil mechanics.

The last chapter, entitled 'Cases of phase movement', is concerned almost entirely with fluid flow in sands. This is perhaps the most useful chapter from the point of view of the practising engineer. One would have liked to see the treatment extended to clay soils and the subject of consolidation.

This well-produced and carefully translated book must provide a valuable supplement to any soil mechanics course, and engineers in general will find it both stimulating and easy to read.

D.C.

STABILIZED EARTH ROADS

Archiv für Acker und Pflanzenbau und Bodenkultur 10/1979 (German Dem. Rep.)

Kézdi Árpád: Stabilized earth roads (Stabilisierte Erdstraßen). Budapest, Akadémiai Kiadó 1979, 327 S., 296 Abb., 48 Tab., 256 Lit., Ft 520,—

Zu der ungarischen Originalausgabe „Stabilizált Földutak" und der 1973 publizierten deutschen Fassung ist 1979 nunmehr eine überarbeitete englische Ausgabe erschienen, die im Umfang und Aufbau der deutschen weitgehend gleicht. Durch die Auswertung der wesentlichen Literatur bis etwa 1975 und unter Berücksichtigung der über mehrere Jahrzehnte vorliegenden umfangreichen Erfahrungen des Autors wurden bei gleichzeitiger Straffung des Inhaltes — gegenüber der deutschen Ausgabe — auch einige neue Abschnitte in Kapitel 7 bis 9 aufgenommen.

Nach einleitenden Ausführungen über die Bedeutung von stabilisierten Erdstraßen und deren Lösungswege zum Bau werden ausführlich die physikalischen, chemischen und bodenmechanischen Grundlagen dargelegt. Mehr als die Hälfte des Buches wird den Ausführungen in Kapitel 3 bis 7 über die mechanische Stabilisierung sowie den Stabilisierungen mit Zement, Kalk, Bitumen und Teer bzw. Chemikalien gewidmet. Im 8. Kapitel „Entwurf von Stabilisierten Erdstraßen" wird im Detail auf die Anwendungsbereiche der Erdstabilisierung, auf die technischen Daten derartiger Erdstraßen und auf die Bemessung der Straßenbefestigung eingegangen. Im letzten Kapitel werden die Konstruktionsmethoden und einzelnen Bauphasen dargelegt sowie ein typisches Bauverfahren (in-situ mixing) näher ausgeführt.

Mit dem vorgelegten Buch werden keine fertigen, abgerundeten „Rezepte" oder ausführlichen Anweisungen zu Konstruktionen und Bauweisen stabilisierter Erdstraßen unterbreitet, sondern multilaterale Grundlagen geboten bzw. komplexe Zusammenhänge dargelegt, um daraus optimale Lösungen abzuleiten, die den jeweiligen Standortbedingungen und territorialen Ressourcen hinsichtlich der anzuwendenden Stabilisierungsmittel angepaßt sind. Auf diese Weise wird dieses Buch zu einem unentbehrlichen Rüstzeug für jeden auf diesem Gebiet tätigen Ingenieur werden, um eine fachgerechte und material-ökonomisch fundierte Stabilisierung der Erdstraßen vom Entwurf bis zur Bauausführung zu sichern. Das besondere Anliegen des Autors ist es, durch zahlreiche Hinweise und gebotene Zahlenbeispiele den Praktiker zu befähigen, selbst die Einflüsse und Auswirkungen zu beurteilen, die sich aus den wechselnden Eigenschaften des Untergrundes und der Erdstoffe ergeben.

Sowohl die Gestaltung und der Druck als auch die Art und Weise der Wiedergabe von umfangreichen Schwarz-weiß-Abbildungen, von teils recht komplizierten graphischen Darstellungen, von zahlreichen Tabellen und Formeln verdienen als besonders gelungen herausgestellt zu werden.

A. Kullmann

Acta Technica Academiae Scientiarum Hungaricae, 98 (1—2), pp. 59—75 (1985)

RELIABILITY OF THE PREDICTION OF THE LOAD-SETTLEMENT DIAGRAM OF A PILE

E. E. De Beer*

The reliability of the prediction of the behaviour of a ringle pile is investigated by using different methods of evaluation. Results gained by cone penetration test (CPT), λ-method, DPA test, DPB test and WST test are compared with large scale tests. It turned out that CPT tests provide the test result because there is much similarity between the stress field around the cone and the pile.

Introduction

Numerous are the contributions of Professor Kézdi to the problem of Pile Foundations (London 1957, Montreal 1965, Budapest 1976, Weimar 1976, Foundation Engineering Handbook 1975, Mexico 1976). In these contributions he stresses the various parameters influencing the problem, and making a correct prediction based on laboratory and field tests very difficult. In order to get an idea of our actual capabilities of such a prognosis it can be worthwhile to describe the predictions made for some actual pile loading tests, and to compare them with the measured values.

In this contribution the influence of the grouping of the piles will not be considered, which does not mean that this influence should be neglected, especially for the problem of the settlements.

General considerations

1. Ultimate bearing capacity — Conventional Rupture Load — Limit Load

A first attempt is to predict the ultimate bearing capacity of a pile. But at first should be clearly defined what is meant by "ultimate bearing capacity".

In principle the ultimate bearing capacity of a pile is the load for which the gradient $\Delta s : \Delta Q$ of the settlement increase Δs to the pile load increase ΔQ becomes infinite. However the load-settlement diagram of many test piles do not show such a

* Prof. em. Dr. ir. E. E. De Beer, Keise–lisk Plein 3, B 9300 Aalst, Belgium

gradient, or if so, this gradient is only reached after relative settlements $s:D$ (D = diameter of the pile) larger than 30%. Ultimate bearing capacities corresponding to such large deformations have mostly no longer any practical meaning.

Instead of the wording "ultimate bearing capacity" preference is given to two other wordings.

"*Conventional rupture load*" Q_r^{con}

The conventional rupture load of the soil surrounding a displacement pile is by definition the smallest of the two following loads:

1. the load for which the gradient $\Delta s:\Delta Q$ becomes infinite,
2. the load for which the relative settlement $s_b:D$ becomes equal to 10%, where s_b is the settlement of the base of the pile.

"*Limit load*"

Besides the notion "rupture load" is defined the notion "limit load" Q_l.

1. When in a double log representation of the load-settlement diagram of a loading test, the points corresponding to the large loads are located on a straight line, the limit load is by definition the smallest load whose representation is located on that line (see Fig. 1 point Q_l).
2. When in a double log diagram the representative points corresponding to the larger loads are not located on a straight line, the "limit load" is defined as the load causing an irreversible relative settlement $s:D$ equal to 2.5% (DIN 1975—1976).

Thus instead of the vague notion "ultimate bearing capacity" are introduced two clearly defined notions "conventional rupture load" Q_r^{con} and "limit load" Q_l. When making a prediction, one should clearly indicate to which notion the prediction belongs.

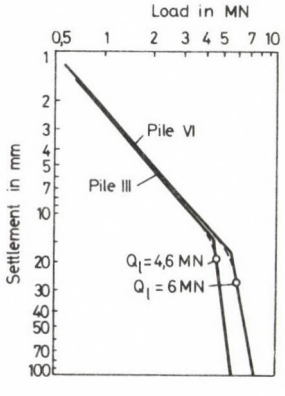


Fig. 1

Furthermore a prediction limited to one or both of these notions, is in many cases not sufficient for practical purposes. In fact one should be able to predict the full load-settlement diagram.

2. Prediction based on the results of laboratory tests

It is very difficult to make a close prediction of the load-settlement diagram of a pile based on laboratory tests on "undisturbed" samples. There are several reasons for this.

The bearing capacity of a pile is influenced by all the layers surrounding the pile. Mostly these layers are heterogeneous and consequently a large number of samples should be tested to obtain the characteristics of all these layers.

In general the properties of the soil layers show differences from one vertical to another, making the prediction for a pile not located at the same vertical as the boring more difficult.

The mantle friction on the pile depends on the stress field induced in the soil by the placement of the pile, and by the subsequent load on the pile. It is difficult to simulate these effects in the laboratory.

The rupture load at the base of a circular or square pile for the case of a cohesionless soil can most approximately be expressed by the equations of Vesic (1972, 1975)

$$q_r = N_q^* \sigma'_{m,0} \tag{4}$$

with $\sigma'_{m,0}$ = the mean effective stress in the soil mass surrounding the base before application of the load

$$\sigma'_{m,0} = \frac{\sigma'_{v,0} + 2\sigma'_{h,0}}{3}, \tag{5}$$

$\sigma'_{v,0}$ = effective vertical stress in the soil mass,
$\sigma'_{h,0}$ = effective horizontal stress in the soil mass,

$$\sigma'_{h,0} = K\sigma'_{v,0}. \tag{6}$$

The values of K and $\sigma'_{v,0}$ depend not only on the original state of stress in the considered soil mass (normally consolidated or overconsolidated), but also on the way the pile has been introduced into the soil (for instance driven piles versus bored piles, or buried piles).

The value of N_q^* depends not only on the angle of shearing strength φ', but also on the reduced rigidity index $I_{r,r}$, defined by

$$I_{r,r} = \frac{I_r}{1 + I_r \Delta} \tag{7}$$

in which I_r represents the rigidity index

$$I_r = \frac{E}{2(1+\nu)(c' + \sigma'_{m,0} \tan \varphi')} \tag{8}$$

where

Δ = the mean unit volumetric strain in the plastic zone surrounding the base,
E = the mean deformation modulus in the elastic zone surrounding the base
ν = Poisson ratio of the soil mass.

The value of φ' to be used is that corresponding to the mean stress in the soil mass surrounding the base.

The deformability modulus E and the mean volumetric strain in the plastic zone can, in principle, be obtained from appropriate oedometric and triaxial tests.

The values of the bearing capacity factor N_q^* are given versus the angle of shearing resistance φ', with the reduced rigidity index $I_{r,r}$ as a parameter in Figure 2. The figure also shows the appreciably large influence of the rigidity on the rupture load of a pile.

The great sensitivity of the results to the values of K, $\sigma'_{m,0}$, φ', E and Δ demonstrate how difficult it is to make a correct prediction of the rupture load of a pile, depending only on the results of usual laboratory tests.

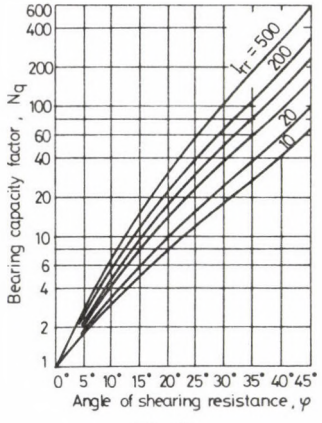


Fig. 2

3. Prediction based on the results of CPT tests

The rupture load of the soil surrounding a pile not only depends on the shearing strength of the soil, but also on its compressibility characteristics, the relative depth of embedment, the initial stress tensor in the soil and also on the stress tensor produced by installing the pile. In order to predict the rupture load correctly, it is necessary to know

all these factors, and it is no easy task to simulate them all in laboratory tests performed on more or less undisturbed samples.

On the other hand a cone penetration test (CPT test) provides an in situ value which is influenced by several of these factors.

The degree of similarity between the stress field around the cone and the pile depends largely on the way the pile has been placed in the soil (driven, jacked, bored, buried). There is, furthermore, a large scale effect. Approximate methods exist which take this scale effect into account (Begemann 1963, De Beer 1963, 1971).

An example of the results of a CPT test is given in Figure 3, showing the cone resistance q_c with depth, obtained by a cone with a diameter of 36 mm. Also shown are the calculated values of unit rupture load, $q_{r,\,calc}$, for driven cylindrical piles, with diameters of ∅ 188 mm, 270 mm and 619.5 mm derived by the De Beer Method. It is clearly seen that at most depths the unit rupture load of a driven pile, 619.5 mm diameter, is much smaller than the q_c values, and that it would be a big mistake to simply transpose the q_c values to the pile problem.

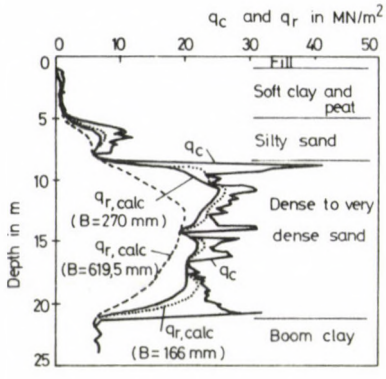


Fig. 3

Reliability of the prediction of the behaviour of a single pile by different methods

In order to check the reliability oóf the prediction, the best way is to compare the scattering of the results obtained by different prediction methods, and when possible to compare them with the results of pile loading tests (McLelland, 1977). Of course only predictions introduced before knowing the results of the pile loading tests are to be considered. Afterwards, it is mostly easy by adopting or adapting some parameters to get a "fairly good agreement" with the calculated method and the reality.

1. Driven piles in stiff clay layers

In Belgium the Boom clay is a stiff fissured clay whose shearing strength characteristics have been determined by several laboratory and field tests. In order to predict the length needed for a driven pile ∅ 1.37 m to reach a rupture load $Q_r = 34\,600$ kN use can be made from the CPT results found in the Boom clay (Fig. 4), and from

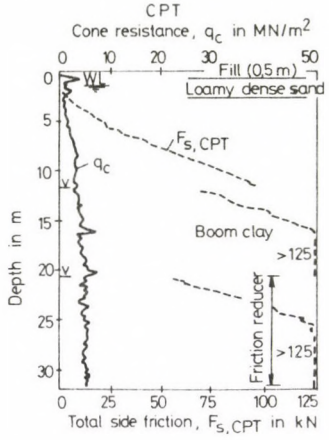


Fig. 4

some experimental factors deduced from the results of pile loading tests performed on real piles driven in this clay, in order to take account on the scale effect.

Another prediction method is the method of Focht (λ method, 1972). The coefficient λ is an empirical coefficient, deduced from a large number of field tests.

The values obtained with both methods are given in Table I. Both methods give the same depth of 62.2 m, thus backing each other. However it should be observed that the CPT method gives a larger base resistance $Q_{b,r}$ and a somewhat smaller shaft resistance $Q_{s,r}$ than the λ method.

Table I. Pile ∅ = 1.37 m in Boom clay

Prediction method	$Q_{b,r}$ kN	$Q_{s,r}$ kN	Q_r kN	Depth m
Belgian CPT (M4)	9 400	25 200	34 600	62.2
Focht (λ Method)	6 490	28 000	34 590	62.2

2. Large scale tests at Houston (Vesic)

At the Stockholm Conference Vesic (1981) showed the results of predictions made by 10 eminent specialists in piling problems, as compared to the measured "ultimate bearing capacity". The Figure 5 taken from Vesic gives the rather large scatter between the predicted values.

Vesic based his prediction on the results of the previously performed CPT tests. It appears that his prediction is one of the nearest to the experimental values.

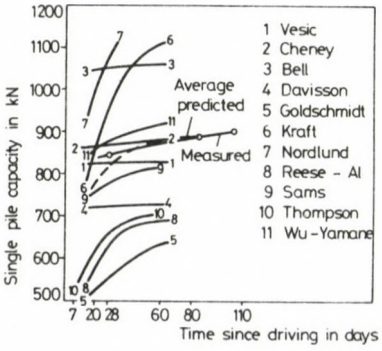


Fig. 5

3. Test piles at ESOPT II — Amsterdam

At the occasion of the Second European Symposium on Penetration Testing at Amsterdam, the Organizing Committee sent to the participants the data of a soil investigation at a test site, asking for a prediction of the bearing capacity of a test pile to be driven at that site.

The available soil data were a CPT test (Fig. 6), a DPA test, a DPB test and a WST test, all performed according to the European Standard (Tokyo, 1977) and at a distance of 2 m from the location of the pile. Also available were the results of a boring performed at a distance of 6 m from the location of the pile, and of the SPT tests performed according to the European Standard (Tokyo 1977).

The test pile to be tested was a prefabricated concrete pile, 250 mm × 250 mm, with a length of 15 m and a flat shaped toe, driven to the level − 13.00 m NAP.

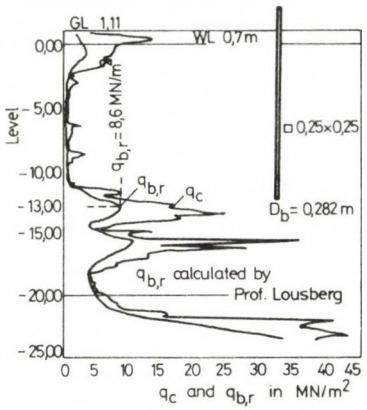


Fig. 6

Calculation of the conventional rupture load

The unit rupture load under the base $q_{b,r}$ is calculated from the cone resistance q_c with the method De Beer (1971) for an equivalent diameter of the pile base D_b, given by

$$D_b = \sqrt{\frac{4 \times 0.25 \times 0.25}{\pi}} = 0.282 \text{ m}. \tag{9}$$

The calculations are based on the following data and assumptions:

Phreatic level: 0.70 + NAP

Volume weight above phreatic level: $\gamma_d = 16$ kN/m³,

beneath phreatic level: $\gamma_n = 20$ kN/m³.

The values $q_{b,r}$, calculated with the computer program of the Université Catholique de Louvain, are given versus depth on Fig. 6. At the level 13.00 m – NAP of the pile base one obtains:

$$q_{b,r} = 8.6 \text{ MN/m}^2. \tag{10}$$

The base resistance $Q_{b,r}$ is given by

$$Q_{b,r} = A_b q_{b,r} = 0.0625 \times 8.6 = 0.538 \text{ MN}. \tag{11}$$

In Belgium, the shaft friction resistance $Q_{s,r}$ is usually calculated from the total lateral friction resistance Q_{st} measured in the CPT test. As the total penetration resistance was not measured in the CPT test in Amsterdam, the Belgian method could not be applied.

In place of the usual Belgian method, the shaft friction resistance was deduced from the values q_c of the cone resistance in the following way:

— In sands, the unit friction resistance $q_{s,r}$ on displacement piles, is related to the cone

resistance q_c by the following empirical formulae:

$$q_{s,r} = \frac{q_c}{200} \qquad \text{when} \quad q_c \geqq 20.0 \text{ MN/m}^2, \tag{12}$$

$$q_{s,r} = \frac{q_c}{150} \qquad \text{when} \quad q_c \leqq 10.0 \text{ MN/m}^2. \tag{13}$$

For intermediate values of q_c, the value of $q_{s,r}$ is calculated with a formula obtained by linear interpolation between $q_c/200$ and $q_c/150$.

In cohesive soils and for displacement piles one disposes on several experimental data (Carpentier, 1970; De Beer et al, 1977) from which a relationship between the unit shaft friction $q_{s,r}$ on driven prefabricated concrete piles and the cone resistance q_c (M4) is deduced, as shown by the curve IJ of Fig. 7.

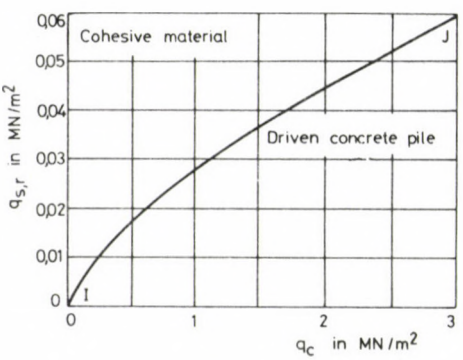


Fig. 7

Shaft friction resistance in the bearing stratum

The bearing stratum is a dense sand layer. For the layer between 11.39-NAP and 12.99 m-NAP one has (Fig. 6)

$$(q_c)_m = 10.01 \text{ MN/m}^2, \tag{14}$$

$$(q_{s,r})_m = \frac{10.01}{150} = 0.0667 \text{ MN/m}^2, \tag{15}$$

and

$$Q_{s,r,1} = \Delta h_1 \chi_s (q_{s,r,})_m, \tag{16}$$

$$\Delta h_1 = \text{thickness of the layer} = 1.60 \text{ m},$$

$$\chi_s = \text{perimeter of the shaft} = 4 \times 0.25 = 1.00 \text{ m}, \tag{17}$$

$$Q_{s,r,1} = 1.60 \times 1.00 \times 0.0667 = 0.107 \text{ MN}. \tag{18}$$

Shaft friction resistance in the soft layers

For each of the considered strata the mean value $(q_c)_m$ of the cone resistances and the corresponding values $q_{s,r,i}$, deduced from Fig. 7 are given in Table II.

Table II

Layer m — NAP	$(q_c)_m$, MN/m² (Fig. 6)	Δh_i m	$q_{s,r,i}$ MN/m²	$Q_{s,r,i}$ MN
2.19– 3.09	1.12	0.90	0.030	0.027
3.09– 5.19	0.49	2.50	0.018	0.038
5.19– 6.29	0.41	1.10	0.016	0.018
6.29– 9.19	1.23	2.90	0.032	0.093
9.19–10.99	0.67	1.87	0.035	0.038
10.99–11.39	1.65	0.40	0.039	0.016
				$Q_{s,r,2} = \sum Q_{s,r,i} = 0.230$ MN

Finally in the last column the estimated total friction resistance $Q_{s,r,2}$ is given:

$$Q_{s,r,2} = \sum Q_{s,r,i} = \sum q_{s,r,i} \chi_s h_i. \tag{19}$$

Shaft friction resistance in the upper sand layer

For the sand layer between 1.01 + NAP and 2.19 − NAP we have

$$(q_c)_m = 7.43 \text{ MN/m}^2, \tag{20}$$

$$(q_{s,r,})_m = \frac{7.43}{150} = 0.0495 \text{ MN/m}^2, \tag{21}$$

$$Q_{s,r,3} = \Delta h_3 \chi_s (q_{s,r})_m = 3.30 \times 1.0 \times 0.0495 = 0.158 \text{ MN}. \tag{22}$$

Total shaft resistance:

$$Q_{s,r} = 0.107 + 0.230 + 0.158 = 0.495 \text{ MN}. \tag{23}$$

Total bearing capacity:

$$Q_r = 0.538 + 0.495 = 1.033 \text{ MN}. \tag{24}$$

As the pile base is located beneath the cricital depth in the bearing stratum, according to the method De Beer, the calculated total bearing capacity Q_r corresponds to the conventional rupture load (this is the load which in a monotonously increasing loading corresponds to a settlement of 10% of the pile base diameter, except if before total rupture should occur).

Calculation of the pile cap load-settlement diagram

In order to obtain a safe estimation of the pile cap settlement, the residual forces in the pile due to driving are disregarded. It is assumed that for loads Q_b on the pile base which are much smaller than the rupture load $Q_{b,r}$, the settlements of the soil due to the lateral displacements can be neglected, and therefore with a sufficient accuracy, the oedometric law of compressibility of Terzaghi can be applied.

Let us consider a load

$$Q_b = \frac{Q_{b,r}}{3} = \frac{0.538}{3} = 0.179 \text{ MN}. \tag{25}$$

As the stress level at the pile base generated by the load Q_b is in any case smaller than the stress level during the driving of the pile, not the constant C of virgin compression, but the recompression constant A has to be introduced.

The settlement of the pile base s_b is then given by the expression

$$s_b = \sum \frac{\Delta h_i}{A} \ln \frac{p'_0 + i(q_b - p_0)}{p'_0} \tag{26}$$

with

p'_0 = the initial effective stress at a depth h underneath the pile base;
p_0 = the initial effective stress at the level of the pile base;
q_b = the considered unit pressure on the base;
i = coefficient giving the variation of the stress increase with depth underneath the singular point of the pile base (De Beer, 1949);
Δh_i = the thickness of the considered sub layer;
A = the recompression constant.

One has:

$$p_0 = 18 \times 0.40 + (20 - 10)(14.10 - 0.40) = 144.2 \text{ kN/m}^2, \tag{27}$$

$$q_b = \frac{Q_b}{A_b} = \frac{0.179}{0.0625} = 2.864 \text{ MN/m}^2. \tag{28}$$

In sand safe values of the compression constant C are given by (De Beer, 1949)

$$C \geqq \frac{3}{2} \frac{q_c}{p'_0}, \tag{29}$$

p'_0 = the initial effective stress at the level where q_c is measured. In pure sands the recompression constant A can at most be 10 times larger than the virgin compression constant C.

Thus

$$A \leqq 10C. \tag{30}$$

In order to obtain a safe value of the recompression constant A, instead of q_c the base resistance $q_{b,r}$ has been introduced:

$$A = 10 \cdot \frac{3}{2} \cdot \frac{q_{b,r}}{p'_0}. \tag{31}$$

At the level of the pile base one has

$$A = 10 \cdot \frac{3}{2} \frac{8600}{144.2} = 895. \tag{32}$$

This value has simply been considered constant over the whole thickness of the compressed layer.

The calculations of the pile base settlement for $Q_b = 0.179$ MN are given in Table III.

The calculated settlement amounts to

$$s_b = 1.508 \text{ mm}. \tag{33}$$

Table III

h in m	Δh in m	h/b	i	p'_0 kN/m²	$i(q_b - p_0)$ kN/m²	$\log_{10}$	s in mm
0							
	0.0625	0.125	1	144.5	2720	1.297	0.209
0.063	0.0625	0.375	0.7	145.1	1904	1.150	0.185
0.125	0.0625	0.625	0.46	145.8	1251	0.981	0.158
0.188	0.0625	0.875	0.33	146.4	898	0.853	0.137
0.250	0.0625	1.125	0.24	147.0	653	0.736	0.118
0.313	0.0625	1.375	0.19	147.6	517	0.653	0.105
0.375	0.0625	1.625	0.15	148.3	408	0.574	0.092
0.437	0.0625	1.875	0.12	148.9	326	0.504	0.081
0.500	0.0625	2.125	0.10	149.5	272	0.450	0.072
0.563	0.0625	2.375	0.09	150.1	245	0.420	0.068
0.625	0.0625	2.625	0.08	150.8	218	0.388	0.062
0.688	0.0625	2.875	0.07	151.4	190	0.354	0.057
0.750	0.0625	3.125	0.06	152.0	163	0.317	0.051
0.813	0.0625	3.375	0.05	152.6	136	0.277	0.045
0.875	0.0625	3.625	0.04	153.3	109	0.233	0.038
0.937	0.0625	3.875	0.03	153.9	82	0.185	0.030
1.00							

$s_b = 1.508$ mm

It is assumed that for $Q_b \leqq Q_{b,r}/3$ the settlement of the base varies linearly with the load on the base Q_b.

For $Q_b > Q_{b,r}/3$ the settlements caused by the lateral movement of the soil cannot longer be neglected and with increasing Q_b they become predominant.

For the value $Q_{b,r}=0.538$ MN the conventional rupture load is reached, and consequently the settlement

$$s_{b,r}=\frac{D_b}{10}=\frac{282}{10}=28.2 \text{ mm}. \tag{34}$$

The assumed relationship $s_b=F(Q_b)$ is given by the curve OAB of Figure 8.

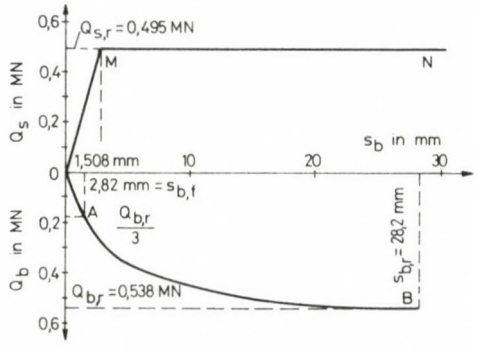


Fig. 8

Mobilization of the shaft friction resistance

It is assumed that the shaft friction resistance is completely mobilized for a relative displacement soil-pile shaft of 0.01 D_s (D_s = diameter of the pile shaft). For a settlement of the pile base $s_b=0.01D_s$, all points of the pile shaft will undergo a settlement at least equal to 0.01 D_s and therefore the shaft friction resistance will be completely mobilized.

According to this assumption, for a settlement of the base $s_{b,f}=282/100=$ 2.82 mm the side friction $Q_{s,r}=0.495$ MN is completely mobilized.

It is assumed that for $s_b \leqq s_{b,f}$ the mantle friction varies linearly with s_b. In absence of peak shear strength values, for $s_b > s_{b,f}$ the mantle friction is considered to remain constant. The assumed variation of the side friction Q_s versus the settlement of the base is given by the broken line OMN of Figure 8.

From the Figure 8 for a given settlement of the base s_b, one obtains the force at the base Q_b and the mantle friction Q_s, and consequently the total force on the pile

$$Q=Q_b+Q_s. \tag{35}$$

Pile cap settlement s_{cap}

The pile cap settlement is given by

$$s_{cap} = s_b + \frac{Q_b L}{EA_b} + \frac{1}{2}\frac{Q_s L}{EA_b}. \tag{36}$$

— For instance for:

$$Q_b = 0.179 \text{ MN}; \quad s_b = 1.508 \text{ mm}; \quad Q_s = \frac{1.508}{2.82} \times 0.495 = 0.265 \text{ MN},$$

$$E = 30.000 \text{ MN/m}^2; \quad A_b = 0.0625 \text{ m}^2; \quad EA_b = 1875 \text{ MN}, \quad L = 15 \text{ m}, \tag{37}$$

$$s_{cap} = 1.508 + \frac{0.0179 \times 15000}{1875} + \frac{1}{2}\frac{0.265 \times 15000}{1875} =$$

$$= 1.508 + 1.432 + 1.060 = 4.000 \text{ mm}, \tag{38}$$

$$Q = Q_b + Q_s = 0.179 + 0.265 = 0.444 \text{ MN}. \tag{39}$$

— Under the conventional rupture load, one obtains

$$Q_{b,r} = 0.538 \text{ MN}; \quad Q_{s,r} = 0.495 \text{ MN}; \quad s_{b,r} = \frac{D_s}{10} = 28.2 \text{ mm},$$

$$s_{cap,r} = 28.2 + \frac{0.538 \times 15000}{1875} + \frac{1}{2}\frac{0.495 \times 15000}{1875},$$

$$s_{cap,r} = 28.2 + 4.304 + 1.98 = 34.5 \text{ mm}, \tag{40}$$

$$Q_r = 0.538 + 0.495 = 1.034 \text{ MN}. \tag{41}$$

The predicted relationship between the pile load and the pile cap settlement is represented on the Figure 9 by the curve OCD.

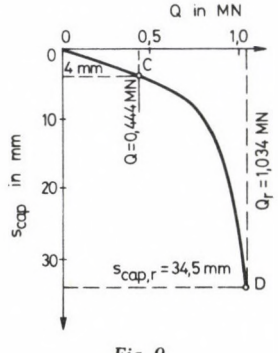


Fig. 9

Critical comparison with the observed values

The Organizing Committee received 36 predictions of the "ultimate bearing capacity". But only 13 predicted the full load-settlement diagram. The Belgian prediction, based on the CPT test in the described way, received the n° 27. On the Figure 10 the heavy dash dotted curve gives the observed settlement of the pile cap versus the pile load, as given by the Organizing Committee. The other lines give the 13 predicted load settlement diagrams. The very large scatter of the predicted curves can easily be observed.

The Belgian prediction n° 27 is represented by an heavier line. It can be seen that for a given load the predicted settlement is always larger than the real one, the maximum difference reaching about 100%.

The predicted conventional rupture load (load corresponding to $s_{r,b}:D=0.1$), $Q_{r,\text{predicted}}=1.035$ MN is somewhat smaller but does not differ very much with the measured conventional rupture load of 1.089 MN.

Although based uniquely on the results of a CPT test, and with rather simple and crude assumptions a load-settlement diagram is obtained, giving safe results, and certainly as good as predictions based on other, possibly much more sophisticated methods.

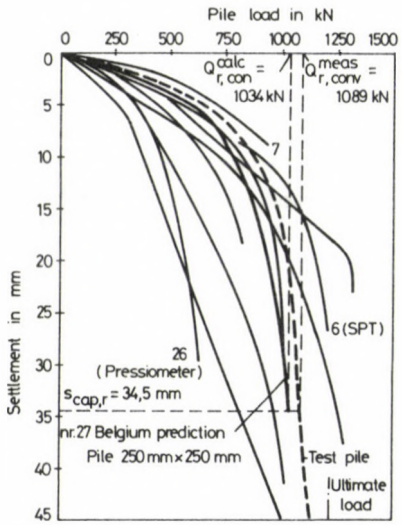


Fig. 10

2. Other examples

The method of prediction of the ESOPT 2 Test pile has been given in detail. This pile was a cylindrical prefabricated concrete pile.

Of course the way of introducing the pile, its geometry, and the nature of the skin of the shaft play an important role. In Belgium very extensive pile loading tests have been performed or are underway in order to define empirical factors, which take into account the influence of the mentioned factors, and are to be introduced in the prediction method.

Some of the results already obtained can be found in the literature, and are related to driven H steel piles (De Beer et al, 1981, 1982) to piles introduced by driving or vibration, and with different geometries (De Beer et al, 1977, 1979, 1981).

Conclusion

The correct prediction of the load-settlement diagram of a pile is a very intricate problem, as such a diagram is influenced by a large number of parameters which depend on the way of introduction of the pile, and on its loading history.

In comparison with other laboratory and field tests, the CPT tests present the advantage that their results are influenced by some of these parameters in an analogous way as displacement piles are. Therefore it is not astonishing that by using the results of CPT tests, duly taking into account the scale effect, one can mostly obtain predicted load-settlement curves which are among the best, obtainable with the methods yet at disposal.

Prediction exercices, checked against big scale tests, are the best means to improve the precision in the prediction of the behaviour of piles.

References

1. Begemann, H. K. S.: — The use of the static soil penetrometer in Holland, New Zealand Engineering Journal, (1963), February.
2. Carpentier, R.: – Vergelijking van de resultaten van enkele vinproeven en van de overeenkomstige weerstanden opgemeten in sonderingen, Tijdschrift der Openbare Werken van België, (1970) nr 3, Juni, 179—186.
3. McLelland: — Geotechnical Problems in Ocean Engineering, Proceedings IXth I.C.S.M.F.E., Tokyo, 3 (1977), 513—523.
4. De Beer, E.: Grondmechanica, Deel II, Funderingen. Standaard Boekhandel, Antwerpen 1949.
5. De Beer, E.: The scale effect in the transposition of the results of deep sounding tests on the ultimate bearing capacity of piles and caisson foundations, Géotechnique 13 (1963), 39—75
6. De Beer, E.: Méthodes de déduction de la capacité portante d'un pieu à partir des résultats des essais de pénétration, Annales des Travaux Publics de Belgique, (1971), 191—268; 321—332; 351—405
7. De Beer, E., Lousberg E., Wallays M., Carpentier R., De Jaeger J.: Bearing capacity of displacement piles in stiff fissured clays, Comptes Rendus des Recherches I.R.S.I.A., Brussel 1977 n° 39, March

8. De Beer E., Lousberg E., De Jonghe A., Carpentier R., Wallays M.: (1979) — Prediction of the bearing capacity of displacement piles penetrating into a very dense sand layer, from the results of CPT tests; Proceedings of the 7th European Conference on Soil Mechanics and Foundation Engineering, Brighton, 3, (1979), 51—59
9. De Beer E., Lousberg E., De Jonghe A., Wallays M., Carpentier R.: Partial safety factors in pile bearing capacity, Proceedings Xth I.C.S.M.F.E., Stockholm, 1, (1981), 105—110
10. De Beer E., De Jonghe A., Carpentier R., Hever M., Scholtes P.: (1981) — H steel piles in dense sand, Proceedings Xth I.C.S.M.F.E. Stockholm, 2 (1981), 693—698
11. De Beer E., Scholtes E., Carpentier R.: Draagvermogen van stalen liggerpalen, Tijdschrift der Openbare Werken van België, (1982) n° 3, 4, 5.
12. Deutsche Normen DIN 4026 — Raumpfähle-Herstellung-Bemessung und zulässige Belastung, Deutsches Institut für Normung, 1975
13. Deutsche Normen DIN 1054 — Zulässige Belastung des Baugrunds, Deutsches Institut für Normung, 1976
14. Kézdi A.: Bearing Capacity of Piles and Pile Groups, Proceedings IVth I.C.S.M.F.E., London, 2 (1957), 46—51
15. Kézdi A.: Deep Foundations — General Report, Proceedings VIth I.C.S.M.F.E., Montreal (1965)
16. Kézdi A.: Pile Foundations, Foundation Engineering Handbook. Van Nostrand Reinhold Company, 1975, 556—594
17. Kézdi A.: Philosophy of deep foundations, Third Nabor Carrillo Lecture, 8th National Meeting of the Mexican Society on Soil Mechanics, 1976 November
18. Kézdi A.: Bemerkungen zur Anwendung der Pfahlgründungen. II. Int. Symposium des D.D.R. Komitees für Bodenmechanik und Grundbau-Bauforschung-Baupraxis, 1976, Heft 9
19. Kézdi A., Marczal L., Farkas J.: Measurement of skin friction and point resistance of Benoto piles, Proceedings 5th Budapest Conference on Soil Mechanics and Foundation Engineering 1976 October, 313—325
20. Vesic, A. S.: Expansion of cavities in infinite soil mass, Soil Mechanics Series n° 25, School of Engineering, Duke University, Durham North Carolina 1972
21. Vesic, A. S.: Principles of pile foundation design, Soil Mechanics Series n° 38, School of Engineering, Duke University, Durham, North Carolina 1979
22. Vesic, A. S.: Behavior of pile Groups, Proceedings Xth I.C.S.M.F.E. Stockholm, 4, (1981), 808—810
23. Vijayvergiya, V. N., Focht S. A.: A new way to predict capacity of piles in clay, Offshore Technology Conference, Dallas, Texas 1972

Acta Technica Academiae Scientiarum Hungaricae, 98 (1—2), pp. 77—86 (1985)

BEHAVIOUR OF CLAYS AFTER LOADING

N. JANBU*

The short and longterm stress-strain behaviour of clays after static loading is analysed with particular emphasize on the effect of stress history. Emphasize is placed on analysing the lateral changes in stress and strain. Simplified models and numerical examples are used, but even so information of fundamental nature seem to have emerged from these analyses.

Simplified model

The idealized model is shown in principle in Fig. 1. The main assumptions are as follows:

— Isotropic initial stress in situ;
— Thin clay layer, i.e. H/B is small;
— Smooth top and bottom of layer;
— Smooth side boundaries, at distance L;
— Plane strain (or axi-symmetry);
— Intermediate stress $\sigma_2=(\sigma_1+\sigma_3)/2$.

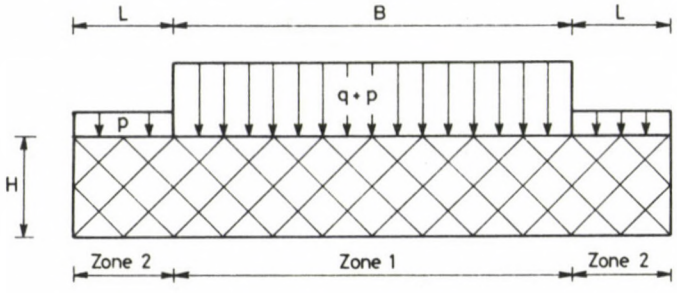


Fig. 1. Idealized model of loaded clay layer

The clay layer carries an additional load q at a level where the vertical surcharge is equal to p. Due to the assumptions made the maximum shear stress equals

$$\tau_{max}=\frac{1}{4}q\,. \tag{1}$$

* Prof. Nilmar Jambu, Norwegian Institute of Technology, Hoegskoleringen 7, 7034 Trondheim, Norway

The maximum shear stress trajectories are inclined at $\pm 45°$ with the horizontal.

The major principal stress σ_1 acts vertically under the load, but horizontally outside the load. Hence, the state of principal stresses are accurately defined over the entire layer.

Initial stress changes

Immediately after load application the total stress changes become:

$$\text{Zone 1: } \Delta\sigma_1 = q\,,\ \Delta\sigma_3 = q/2\,,\ \Delta\sigma_m = 3q/4 = \text{mean}\,,$$
$$\Delta\sigma_d = q/2 = \text{max. deviator}\,,$$
$$\text{Zone 2: } \Delta\sigma_1 = q/2\,,\ \Delta\sigma_3 = 0\,,\ \Delta\sigma_m = q/4\,,\ \Delta\sigma_d = q/2\,.$$

To express undrained, excess pore pressure changes, our institute prefers the following expression, based on total stress change

$$\Delta u = \Delta\sigma_m - D\Delta\sigma_d\,. \tag{2}$$

Laboratory experiences have given D-variations from about $+0.5$ to -0.5. For simple classification of the clays:

$D > 0$ for OC-clay, max $+0.5$
$D = 0$ for 'elastic' behaviour,
$D < 0$ for NC-clay, min. -0.5.

Using Eq. (2) on the model example one obtains the following initial pore pressure changes, u_i; see Fig. 2.

$$\text{Zone 1: } \frac{\Delta u_i}{q} = \frac{3-2D}{4}\,,$$

$$\text{Zone 2: } \frac{\Delta u_i}{q} = \frac{1-2D}{4}\,. \tag{2a}$$

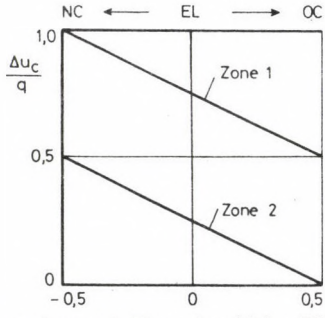


Fig. 2. Undrained excess pore pressure changes in Zones 1 and 2 for different clays. Initial values Δu_c at time $t = 0$, theoretically

A main conclusion to be drawn from Fig. 2 is the following; for zone 1:

> Only in very rare cases ($D = -1/2$) will excess pore pressure be equal to the excess load q. In most cases in practice $\Delta u_i < q$, and theoretically as low as $\Delta u_i = q/2$ for overconsolidated clays.

Of course, in Zone 2 the excess pore pressure is much lower, the difference between the Zones is $q/2 =$ constant. The abrupt change over the zone boundary is in more realistic models substituted by a smooth transition zone, see eg. Janbu (1979).

Knowing the total stress changes and the pore pressures, the effective stress changes become

$$\text{Zone 1:} \frac{\Delta\sigma'_v}{q} = \frac{1+2D}{4}; \quad \frac{\Delta\sigma'_h}{q} = \frac{2D-1}{4},$$

$$\text{Zone 2:} \frac{\Delta\sigma'_v}{q} = \frac{2D-1}{4}; \quad \frac{\Delta\sigma'_h}{q} = \frac{1+2D}{4}. \tag{3}$$

These initial, effective stress changes are illustrated in Fig. 3 as a function of D.

The most important information gathered from Fig. 3 is as follows:

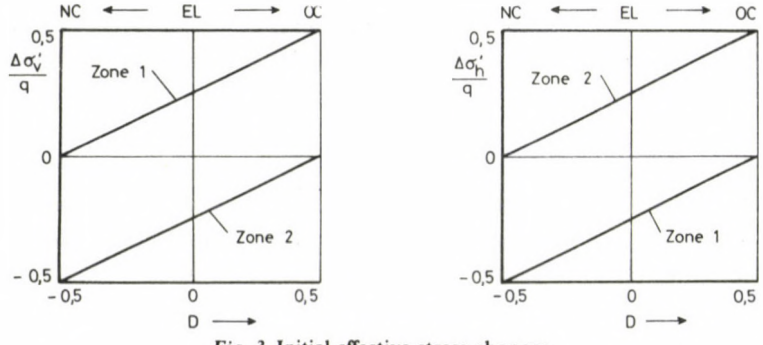


Fig. 3. Initial effective stress changes

> The immediate effective stress changes under undrained conditions are very appreciable for all types of clay. Under the load (Zone 1) the effective vertical stress increases while the horizontal effective stress decreases. Outside the load (Zone 2) it is opposite.

The magnitude of immediate effective stress changes are illustrated by the following two examples:

(1) NC-clay, $D = -0.3$:

$$\Delta\sigma'_v = 0.1q, \quad \Delta\sigma'_h = -0.4q \qquad \text{in Zone 1},$$

$$\Delta\sigma'_v = -0.4q \quad \Delta\sigma'_h = 0.1q \qquad \text{in Zone 2},$$

which will lead to horizontal swelling and vertical compression in Zone 1 under the load, and quite opposite in Zone 2.

(2) OC-clay, $D=0.3$:

$$\Delta\sigma'_v=0.4q\,, \quad \Delta\sigma'_h=-0.1q \qquad \text{in Zone 1}$$

$$\Delta\sigma'_v=-0.1q\,, \quad \Delta\sigma'_h=0.4q \qquad \text{in Zone 2}$$

which shows that immediate elastic vertical compression in Zone 1 dominates.

Initial deformations

The immediate, undrained deformations can now be analysed in terms of effective stress, since the effective stress changes are known. The one-dimensional tangent modulus concept, Janbu (1963 and 1967),

$$M=\frac{d\sigma'}{d\varepsilon}$$

will be used, because the ordinary ranges of M-values in onedimensional compression and swelling are well known from 20 years of experience for several types of clay.

With a lateral swelling modulus M_s the immediate, one dimensional lateral displacement becomes

$$\delta_{hs}=-\frac{1-2D}{8}\cdot\frac{qB}{M_s}. \tag{4}$$

For saturated soil, this undrained lateral displacement cause an immediate vertical settlement δ_{i0} which corresponds to *no volume change* equals

$$\delta_{i0}=\frac{1-2D}{4}\frac{qH}{M_s}; \qquad \text{for plane strain}\,,$$

$$\delta_{i0}=\frac{1-2D}{2}\frac{qH}{M_s}; \qquad \text{for axi-symmetry}\,. \tag{5}$$

Zone 2 must be laterally compressed an amount equals δ_{hs}. This compatibility requirements is satisfied theoretically, for

$$\frac{L}{B}=\frac{1-2D}{1+2D}\frac{M_c}{M_s}$$

when $M_c=$ horizontal compression modulus outside the loaded area. Usual ranges of D- and M-values lead to the conclusion that the expansion of Zone 1 is absorbed by Zone 2 for an extension $L=10\%$ to 30% of B.

Since an immediate change in effective stress $\Delta\sigma'_v$ takes place Eq. (3), it leads to an immediate, elastic vertical compression in Zone 1, equals

$$\delta_{ie} = \frac{1+2D}{4} \frac{qH}{M_c} \tag{6}$$

when M_c = vertical compression modulus.

The lateral deformations associated with δ_{ie} are usually small, and an insignificant lateral compression of Zone 2 is required even if Δvol. = 0.

Hence, the initial settlement δ_i in the conventional approach is equal to a sum of a no volume change part δ_{i0} and an elastic part δ_{ie}, as follows

$$\delta_i = \delta_{i0} + \delta_{ie} . \tag{7}$$

The two components of the initial deformation are shown as function of D in Fig. 4, from which one reads the following trend:

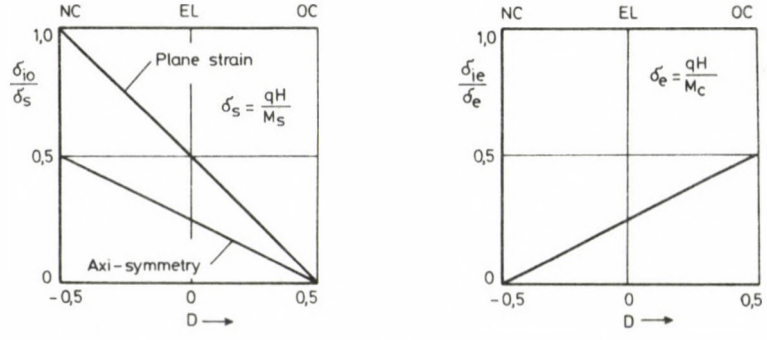


Fig. 4. Components of initial deformations

For a saturated normally consolidated clay with $D = -0.5$ the initial settlement consists only of the no volume change component δ_{i0}, while $\delta_{ie} = 0$. Since $\Delta u_i = q$ for $D = -0.5$ the subsequent consolidation corresponds to a 100% pore pressure dissipation. Therefore δ_{ie} is to be added to δ_c as in the conventional approach.

For a stiff overconsolidated clay with $D = +0.5$ the initial settlement consists only of the elastic part δ_{ie}, while $\delta_{i0} = 0$. Since $\Delta u_i = q/2$ and $\Delta\sigma'_{vi} = q/2$, the δ_i is caused directly by $\Delta\sigma'_{vi}$. The subsequent consolidation is hence caused by a pore pressure dissipation of only 50% of q, leading to a consolidation settlement of only 50% of the theoretical, classical value for M_c = constant. Hence the $\delta_i = \delta_{ie}$ is already included in the classical value of δ_c.

Therefore, it is directly wrong to add δ_i and δ_c for overconsolidated clay, when δ_c is calculated for the full load increase q.

For intermediate clay types, say for $D \cong 0$, the initial settlement contains both components, and now the consolidation due to pore pressure dissipation (δ_{cu}) is smaller than the classical value δ_c for total load q.

These considerations can be included formally in a generalized expression for total, vertical settlement.

$$\delta = \delta_i + \delta_{cu} + \delta_s \,, \tag{8}$$

where, theoretically

$\delta_i = \delta_{i0} + \delta_{ie}$ = initial settlement, at $t = 0$,
δ_{cu} = consolidation due to dissipation of Δu_i approximately over a time t_{cu}, since $\Delta u_i < q$ most often $\delta_{cu} \leq \delta_c$,
δ_s = creep (rheological, or so-called secondary compression) for $t > t_{cu}$.

A few comments regarding expression (8) is necessary:

In reality δ_i takes some time, but its timedependency has still not been explored sufficiently. The required t_{cu} is theoretically infinite, but in practice a fairly good approximate, finite estimate can be made. In the model δ_s is estimated for times $t \geqq t_{cu}$, but in reality creep also occurs during pore pressure dissipation. The adopted distinction between δ_{cu} and δ_s is therefore made for practical reasons only, due to a lack of a more realistic composite model. Research is going on to obtain a more satisfactory model for practical *engineering* purposes.

For *numerical examples* the following data are assumed (offshore–dimensions and loads):

$$q = 120\ \text{kPa}\,, \quad H = 20\ \text{m}\,, \quad B = 80\ \text{m} \quad \text{(circle)}\,.$$

Three types of model clays are selected, classified by their D-values:

NC: $D = -0.3$,	$M_c =$ 1.2 MPa,	$M_s = 3$ MPa,
EL: $D =$ 0,	$M_c =$ 4 MPa,	$M_s = 10$ MPa,
OC: $D =$ 0.3,	$M_c = 10$ MPa,	$M_s = 40$ MPa.

Using the derived formulas one obtains the numerical results shown in Table 1.

Table 1. Numerical examples

Clay	$\delta_{i0} + \delta_{ie} = \delta_i$ m m m	δ_{cu} m	meters δ_{tot}	$\frac{\Delta u_i}{q}$	$\frac{\delta_i}{\delta_{tot}}$
NC	0.32 + 0.20 = 0.52	1.80	2.32	0.90	0.22
EL	0.12 + 0.15 = 0.27	0.45	0.72	0.75	0.37
OC	0.01 + 0.10 = 0.11	0.14	0.25	0.60	0.44

The contribution δ_{cu}, due to pore pressure dissipation, is added for completeness, and also the magnitude of dissipated pore pressure Δu_i versus q. Table 1 shows that δ_i amounts to 22% to 44% of the total δ, with the largest value obtained for the OC-clays.

Pore pressure dissipation

The initial pore pressure Δu_i will dissipate with time. Theoretically, more than 90% of Δu_i is dissipated at a time t_{cu} equals

$$t_{cu} = \frac{d^2}{c_v}, \tag{9}$$

where d = drainage path for vertical one-dimensional drainage only, and c_v = coeff. of cons. (in m^2/year).

The corresponding consolidation settlement, δ_{cu} due to pore pressure dissipation, becomes

$$\delta_{cu} = \frac{3-2D}{4} \frac{qH}{M_c}. \tag{10}$$

This means that $\delta_{ie} + \delta_{cu} = qH/M_c = \delta_c$. The values of δ_{cu} for the three types of clay in the example are included in Table 1.

The time required for pore pressure dissipation will be shorter than expressed by Eq. (9) because horizontal drainage will also take place. For this example this horizontal effect is neglected herein.

Creep (or secondary consolidation), δ_s

When $\Delta u_i = 0$, i.e. $t > t_{cu}$, a constant state of total and effective stress exists. The deformation are now creep (rheologically), or so-called secondary consolidation in geotechnics. For the example shown one can express δ_s as follows (Janbu, 1969) in the simplest model:

$$\delta_s = \frac{H}{r_s} \ln \frac{t}{t_{cu}}, \tag{11}$$

where r_s = dimensionless time resistance, or simply the creep number.

If one prefers to estimate creep rate $\dot{\delta}_s$ at the time $t > t_{cu}$

$$\dot{\delta}_s = \frac{H}{r_s t}. \tag{12}$$

For the three model clays used herein, the proper time-related parameters are:

NC: $c_v = 3\ m^2/yr$, $r_s = 150$,
EL: $c_v = 10\ m^2/yr$, $r_s = 400$,
OC: $c_v = 25\ m^2/yr$, $r_s = 1000$.

When assuming double drainage $d = H/2 = 10$ m in the example, and with the data

above one obtains; Eqs (9) and (10):

$$\delta_{cu} = 1.8\,\text{m}, \quad 0.45\,\text{m} \quad \text{and} \quad 0.14\,\text{m},$$
$$t_{cu} = 33\,\text{yrs}, \quad 10\,\text{yrs} \quad \text{and} \quad 4\,\text{yrs}$$

for the three clays NC, EL and OC respectively.

From Eq. (11) one finds

$$\delta_s = 9.0\,\text{cm}, \quad 3.5\,\text{cm} \quad \text{and} \quad 1.4\,\text{cm},$$
$$\text{for} \quad t = 2t_{cu},$$

and from Eq. (12):

$$\dot{\delta}_s = 4\,\text{mm/yr}, \quad 5\,\text{mm/yr} \quad \text{and} \quad 5\,\text{mm/yr}$$
$$\text{at} \quad t = t_{cu}.$$

Settlement versus time

The informations obtained up to now are used to construct the curves for settlement versus time for the three types of clay, called NC, EL and OC. The result is shown in Fig. 5.

The indicated point of 50% consolidation δ_{cu} is plotted at time $t_{50} = 0.2t_{cu}$ according to classical theory for $M_c = \text{constant}$, and vertical drainage only.

In reality M increases with depth (particularly for NC-clay) and M increases with effective stress. A strain theory for consolidation (Janbu 1965) can take these effects into account and it leads to a faster consolidation process, particularly to begin with. This effect is in principle included in Fig. 5 for the NC-clay. Horizontal drainage also speeds up consolidation.

For the OC-clay the concept of $M = \text{constant}$ may be a fairly good approximation. To neglect lateral effects is also reasonable, so no correction is indicated on the OC–diagram in Fig. 5.

Concluding remarks

The idealized very simple analyses carried out herein has been triggered by acute problems arising in offshore engineering.

The basic geotechnical reasonings behind these analyses are the following:

— All deformations in granular soil has their main root in the grain skeleton response, and since the deformations of the grain skeleton is solely dictated by effective stress changes, irrespective of boundary conditions, it is necessary to explore the effective stress changes at any time after the load application has taken place, in order to get insight into the actual soil behaviour at various times after load application.

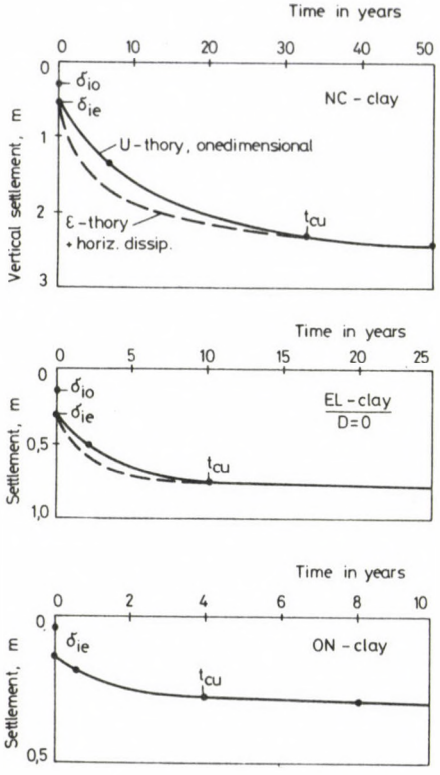


Fig. 5. Estimated settlement versus time for the three model clays

The analyses have lead to the following observations:

— The conventional approach of adding initial settlement and 100% consolidation settlement is justified only in very rare cases.

— The amount of consolidation settlement due to pore pressure dissipation is often overestimated by the conventional approach using $\Delta u_i = q$, since in reality $\Delta u_i < q$.

— The net result of these deviations is that the portion of the total settlement that occurs during the early stages after load application is often underestimated by the conventional approach while the total value itself is most likely overestimated, classically (creep neglected).

If one adds the influence of increasing M with depth and with effective stress, the rate of settlement is increased furthermore, particularly during the first part of consolidation.

The effect of load repetitions (wave action) superimposed on an average state of stress, is now subject to intense research to be able to predict the changes in rate of settlement and possible changes in total settlements due to cyclic loads.

In closing it is strongly emphasized that this oversimplified analyses are carried out solely for the purpose of trying to identify trends of behaviour which may not be properly understood as yet. The real behaviour near $t=0$ requires more intricate studies. Such studies are urgently needed.

References

The papers directly referred to for carrying out these analysis are; in order of reference:

1. Janbu, N.: Design analyses for gravity platform foundations. Proc. Boss'79, London, 1 (1979), 407—426
2. Janbu, N.: Soil compressibility as determined by oedometer and triaxial tests. Proc. ECSMFE, Wiesbaden, 1 (1963), 19—25
3. Janbu, N.: Settlement calculations based on the tangent modulus concept. (Moscow-lectures). Bulletin 2, SM, Norw. Inst. of Techn., Trondhein (1967), 1—57
4. Janbu, N.: The resistance concept applied to deformations of soils. Proc. 7, ICSMFE, Mexico, 1 (1969), 191—196
5. Janbu, M.: Consolidation of clay layers based on nonlinear stress-strain. Proc. 6. ICSMFE, Montreal, 1 (1965), 83—87

Acta Technica Academiae Scientiarum Hungaricae, 98 (1—2), pp. 87—95 (1985)

EVALUATION OF THE SMALL COHESION EXISTING IN NATURAL SANDS DEEMED TO BE COHESIONLESS

J. KERISEL*

Small cohesion of sandy media can play an important role in active and passive earth pressure problems or in arching effects. Because of the problems of taking undisturbed samples from sand the laboratory investigation of this problem is not possible. Site investigation should be used. Cohesion can be determined by pulling out a panel or a pile. Results of small and full scale tests and the theoretical investigation are presented.

Theoretical considerations

In the calculation of foundations and retaining walls relative to sandy media, one is inclined to refer only to the angle of friction whenever the cohesion is supposed to be no more than some $T.f/m^2$ i.e. a few tens of KPa. To proceed on such a line is correct and safe for dry clays retained by walls: they show large fissures through which the rain can percolate, ruining the cohesion. But it is different for non plastic media. In tropical countries, for example, the embankments almost vertical of deep cuts in sand often remain stable in the long term and to neglect the small cohesion of these sands would be a denial of the reality.

More generally, to neglect a cohesion of some tens of KPa associated with an angle φ of 30° leads to an important waste. For a retaining wall, the multiplicator of C is then:

$$\frac{1-K_a}{\tan \varphi} = 1.15$$

20 KPa of cohesion only corresponds to — 23 KPa for the thrust which compensates the thrust of 9 m of cohesionless sand. For a foundation resting on sand, the influence of the cohesion is still more important. The C multiplicator is:

$$\frac{K_p \exp \pi \tan \varphi - 1}{\tan \varphi} = 30.$$

20 KPa of cohesion correspond to an allowable additional ultimate pressure of 600 KPa.

* Jean Kerisel, Past President of the International Society of Soil Mechanics and Foundation Engineering, Simecsol, 115 rue Saint-Dom, 75007 Paris, France

Driving tunnels or underpinning foundations in sandy soils would be impossible without this small cohesion which explains all arching effects.

Many small scale models using sands artificially built layer by layer by compaction are not equivalent to large volume of natural sands not only due to scale effects but also to the small cohesion existing in the latter.

Many observations have shown that a small cohesion can take place in a cohesionless medium in-situ after a life of only a hundred of days. This cohesion cannot but increase with time as shown for clays by Bjerrum in his dissertations on ageing effect. The reasons of this built-in small cohesion in sands are multiple: often chemical but more generally under compressive stresses, they are due to an attrition followed by liaisons on small surfaces of grains. Autor had always in mind the construction of the Tower Maine-Montparnasse in Paris some ten years ago. The builder had to remove a 15 m high fill in sandy material which was worked in 1850 the top of it being the platform of the tracks in the Montparnasse Railway Station. So important was the cohesion 120 years after, that the contractor had to use pneumatic picks to remove the fióll.

Conversely, this small cohesion may be transitory: it has abused many young children playing in insupported trenches dug in humid sands.

Therefore, an important problem for engineers is to measure this small cohesion. Laboratory tests on undisturbed samples is almost impossible. The core samples are so fragile that trimming is a challenge.

How can a C some tens of KPa can be determined by an in-situ test, whatever φ? To do that, we propose pulling tests. Many tests of this category have been performed in the past and they have been reported, in particular by Ireland (1957), Sutherland (1965), Meyerhof (1973), Das (1972) but the difficult problem is the interpretation of the test. Our interpretation is based:

a) on the corresponding states theorem of Caquot (1934). This theorem shows that there is a correspondence between a medium C, φ and a medium O, φ by adding a spherical vector $C/\tan\varphi$ in every inner point and a vector $C/\mathrm{tg}\,\varphi$ normal to the outer surfaces;
b) on our tables, for the calculation of active and passive pressure (1948);
c) on pulling tests in cohesionless sand mixed with a small amount of cement giving birth to a cohesion which is thus known a priori.

In fact, one knows that beyond a certain depth, the equations ruling pulling tests are very different from those corresponding to a small depth. The analogy is obvious with the driving of a pile where down to a certain depth (one meter for small diameters) there is a relationship between the overburden and both point resistance and lateral friction.

Here, of course, there is no point resistance but the question is to write correctly the relationship between lateral friction and γD, γ specific weight of the $C\varphi$ medium, D depth of the small embedded pile to be pulled.

Formula

First, let us suppose that we are concerned by an indefinite panel to be pulled; we call $2R$ the thickness and D the depth.

Using the corresponding states theorem, the problem may be split into two others.

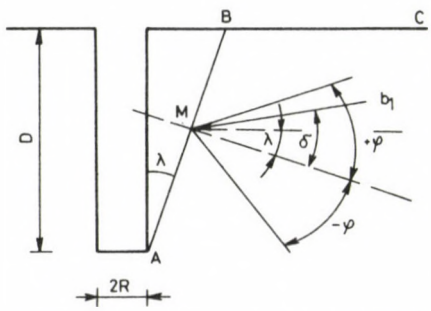


Fig. 1. Elements of the problem in a medium $C=0$, φ

a) Medium with weight and friction; no cohesion

When pulling up the panel, a passive pressure will act on a plane AB inclined at an angle λ to the vertical. λ is unknown but for every λ, we will determine the passive pressure and optimized λ. We call b_1 the passive pressure at point M at unit distance from B. Its obliquity is called δ with $-\varphi<\delta<\varphi$.

In order that b_1 gives a downwards component (opposed to the pulling force) it is necessary that: $\lambda<\delta<\varphi$.

For a given λ, we have first to optimize δ in order to get the maximum vertical component $b_1 \sin(\delta-\lambda)$.

For example, with $\varphi=30°$, $\lambda=10°$, the optimization corresponds to $\delta=2\varphi/3$ whatever φ.

We have calculated all the b_1 optimized in relation with δ and compared them, and have found that the max k of the vertical component is obtained always for $\lambda=0$ that is to say when plane AB is confounded with the lateral surface. Such a result is not obvious a priori and we have finally the Table 1.

Table 1. Maximum max. of the vertical passive pressure K

φ in degrees	20	25	30	35	40
k	0.33	0.40	0.49	0.60	0.72
δ for k	1	3/4	2/3	7/12	1/2

the pulling force in this first problem is therefore equal to:

$$Q = 2x \frac{1}{2} \gamma D^2 k$$

the factor 2 relates to the two lateral surfaces, or:

$$\frac{Q\varphi}{2D} = \frac{1}{2} \gamma D k,$$

k being given by Table 1.

b) Weightless medium with $C\varphi \neq 0$

We call here H the isotropical tensor $C/\tan \varphi$.

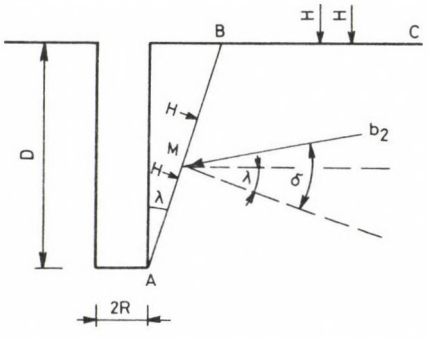


Fig. 2. Conesive weightless medium

If H is applied in every inner point, the equilibrium is not changed. Therefore, we have only to determine the passive pressure b_2 to be brought on plane AB by H applied on the outer surface BC, being well understood that from the normal component of that passive pressure we will have to deduce H acting on AB.

For every λ and a given δ, b_2 has been determined by the tables of Absi and Lherminier; afterwards, we have calculated the expression of the vertical component

$$b_2 \sin(\delta - \lambda) - H \sin \lambda$$

and found out the maximum of that expression varying δ and finally compared the results for several values of λ. As previously we found that the max. max. is obtained for $\lambda = 0$. Finally the max. max. of Q_c are shown in Table 2.

Table 2. Max. max. values of Q_c

φ°	20	25	30	35	40
$\frac{Q_c}{2DH}$	0.32	0.40	0.49	0.58	0.68
$\frac{Q_c}{2DC}$	0.88	0.86	0.85	0.83	0.81

$Q_c/(2DC)$ is decreasing very slowly with φ and in the fork 20° to 40° which interests us, with a max. error of 4%, we can write:

$$Q_c = 0.85 \times 2DC \text{ whatever is } \varphi.$$

To summarize the total result of the two subproblems

$$\frac{Q}{2D} = \frac{1}{2}\gamma Dk + 0.85C, \tag{1}$$

k given by Table 1.

For $\varphi = 30°$ and $C = 20$ KPa for instance

for D	1 m	2 m
$Q/2D$ in KPa	$4.2 + 17 = 21.1$	$8.4 + 17 = 25.4$

The second term, even with this small cohesion, is preponderant. Now, is this formula (1), is the pulling force divided by the screen lateral surface still correct for a pile of diameter $2R$? In other terms is $Q/(2\pi RD)$ given by the same formula?

Experimental results reported hereafter seem to show that the answer is yes and finally if A is the lateral surface of the pile or panel (1) may be rewritten as follows:

$$\frac{Q}{A} = \frac{1}{2}\gamma Dk + 0.85\ \mathrm{C}. \tag{2}$$

Experimental results

In all the following experimental tests we speak of net pulling force, i.e. we have deduced from the total pulling force the weight of the piles or panels.

a) Small scale pulling tests in sands where it was believed that C is zero

We have checked the experiments of DAS (1977) with the values given by our formula (2).

The correlation is good if a cohesion of 4 KPa existed in the sand in which DAS piles were pulled.

b) Full scale tests

A number of full scale tests were gathered by the Comité d'Etudes de la Conférence Internationale des Grands Réseaux Electriques (Barraud et al, 1965); the tests were performed in various countries to determine the stability conditions of pylons, in Australia by the State Electricity of Victoria (shown by *A* in Fig. 3), in Poland

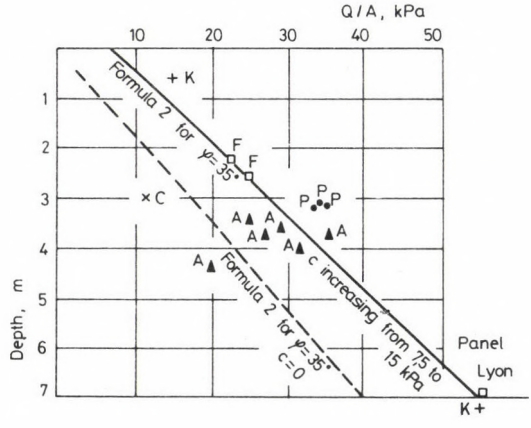


Fig. 3

by Energo projekt (shown by P), in France by Electricité de France (shown by F). The letters *C* and *K* correspond to experiments performed respectively by Cambefort and Kerisel.

Figure 3 concerns sands with $\varphi = 35°$ and *C* deemed to be zero.

On Figure 3 are drawn two straight lines:

— a dashed one corresponding to formula (2) with $\varphi = 35$ and $C = 0$,

— another full line corresponding to formula (2) with $\varphi = 35°$ and *C* increasing with depth from 7.5 to 15 KPa.

The agreement is reasonably good for the full straight line and the corresponding assumptions.

c) Pulling tests in a medium $C\varphi \neq 0$

1. Panels

Full scale tests are very rare for panels.

Exceptionally, we did such a test for the Lyon metro (Kerisel and al, 1972). The panel 0.60 m thick was 4.90 m in length and 7 m in depth. It was made of concrete cast in the Rhone alluviums composed of sand and gravels. The peak value of Q/A measured was 55 KPa.

$\varphi = 35°$ and $C = 12$ KPa were measured in a big shear box test. With these values formula (2) gives 42 KPa: the difference with 55 KPa can be explained by the fact that 35° and 12 KPa were intermediate between peak and residual value.

2. Small scale tests with piles

As explained here above, we built an artificial material which was a Seine sand mixed with cement (4% in weight). The pulled piles had a diameter of 63 mm.

To vary the experiments, the tests were performed after a setting time of the cement of 2 days and 8 days.

Fig. 4 gives the shearing strength τ of the sand, without cement and with 4% of cement at 2 and 8 days.

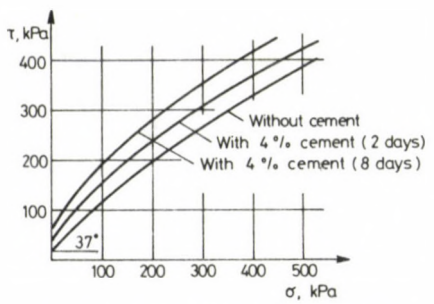


Fig. 4. Shearing strength with the box

Now, the pile was coated with a resin (araldite) and placed into a cylindrical hole worked out with an auger, the hole having a diameter slightly greater than the pile diameter.

After the time interval necessary for the setting of the resin, the pile is pulled and Q/A measured (Fig. 5)

In all tests, the pile came up not bare but enveloped with some centimeters of sand. In formula (2) A has been taken as the lateral cylindrical surface of that sand.

The depths D of the embedded pile were successively 0.25, 0.35 and 0.55 m.

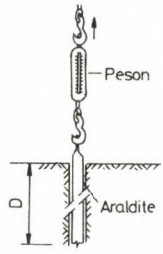


Fig. 5

With such depths it can be seen that in formula (2) even φ is greater than 45° the first term is neglectible in front of the second with the cohesion measured in the sand cement.

Practically with these data

$$\frac{Q}{A} = 0.85\ C, \tag{4}$$

or

$$C = 1.18 \frac{Q}{A}. \tag{3}$$

Finally the results of the tests are shown in Table 3.

Table 3

	D	*C* coming from experiments in						*C* and φ labo with the box	
		0.25 m		0.35 m		0.55 m			
Sand cement		peak	residual	peak	residual	peak	residual		
2 days		84	84	64	64	69	65	60	45°
8 days		98	80	95	82	90	75	80	45°

There is a pretty fair agreement between *C* and measured and deduced from the pulling tests.

Summary

Small cohesion in sandy media plays an important role in active and passive pressure as well in arching effects. One must try to measure it carefully. As trimming is a challenge for such sands, we propose to use in-situ tests and more precisely pulling tests, using small piles (depth *D* smaller than 2 m) as shown in Fig. 5. *C* and φ may be calculated by formula (2) or by the simplified formulas (3) and (4) when $D \leqq 1$ m and *C* greater than 50 KPa.

References

Absi E., Lherminier R.: Tables numériques de butée en milieu pulvérulent non pesant — Cahier de la Recherche 28 — Eyrolles — Paris 1965

Barraud Y., Martin D., Montel B.: Fondations profondes sollicitées à l'arrachement. Revue Construction 20 (1965), N° 4 Avril

Cambefort: Expérience d'arrachement d'un pieu de 3 m enterré dans du mâchefer. Revue Travaux (1947), Juillet — Août.

Caquot A., Kerisel J.: Table de poussée et butée. Gauthier-Villars, Paris 1948

Caquot A., Kerisel J., Absi E.: Tables de butée et de poussée, Gauthier-Villars, Paris 1973

Das M.: Pull out Résistance of Rough Rigid Piles on Granular Soil — Soil and Foundation — Journ. of the Jap. Soc. of S.M.F.E. 17 (1977), N° 3 Sept.

Davis A., Auger D.: La butée des sables: essais en vraie grandeur. Annales I.T.B.T.P. (1979) n° 375, Sept.

Ireland H. O.: Pulling test on piles in sand. Proc. 4th Int. Conf. on S.M.F.E. 2 (1957), 43—46

Kerisel J., Lherminier R., Tcheng I.: Résistance de pointe en milieux pulvérulents de serrages divers — Proc. Conf ISSFME Paris (1961), 265

Kerisel J., Adam M.: Fondations profondes — A.I.B.T.P. (1962) Nov. N° 179, 1073

Kerisel J., Ferrand J., Lareal P., Clement P.: Mesures de poussée et butée faites avec 42 paires de butons asservis. C.R. 5 th Conf.Eur.M.S. 1 (1972), 261—279

Kerisel J., Robert J., Schlosser F, Juran L. et al.: Expérimentation sur un mur à ancrages multiples — 10e Conf.Int.M.S.F.E. (1981)

Meyerhof G.G.: Uplift resistance of inclines anchors and piles: Proc. 8th Int.Conf.S.M.F.E. 2 (1973), 167—172

Sutherland H. B.: Model Studies for shaft raising through cohesionless soils. 6th Int.Conf.S.M. Montréal (1965)410—413

Acta Technica Academiae Scientiarum Hungaricae, 98 (1—2), pp. 97—103 (1985)

SOME EXPERIMENTAL STRESS–STRAIN RELATIONSHIPS FOR LOESS COLLAPSING SOILS

M. V. MALYSHEV—V. A. PUSTOGACHEV*

The problems of deformability of loess soils was investigated on undisturbed samples taken from a borehole of 30 m depth. Tests were carried out on samples with natural moisture content and on saturated soil. Collapsibility of loess was tested in both oedometer and triaxial tests. Special emphasis was placed on analysing the space stress condition by using triaxial testing.

The problems of deformability of loess soils of natural structure and moisture content taken from the holes 30 m deep through the whole depth of collapsing mass have been studied in the Laboratory of the Chair of Soil Mechanics, Foundation Beds and Foundations, MISI after Kuybishev V. V., USSR. These soils according to their collapsing characteristics refer to Type II. The deformability of the soils was tested using oedometer and triaxial compression apparatus. Oedometers were applied to study the deformability of saturated soils and those of natural moisture content. Special researches in collapsing deformation were conducted by the method of single and two curves, the depth at which specimens were taken into consideration. The single curve method was used in carrying on experiments with wetting soil under a pressure equal to γH. For the soil specimens taken at different depths it was found that the collapsing deformation defined under single curve method is always greater (by 20–50%) than that defined under two-curve method. The non-linear dependence of relative collapsibility on pressure was revealed: the higher the pressure, the smaller is the increment of relative collapsibility.

The deformability of loess soils under conditions of space stressed state was studied on triaxial compression apparatus. For this purpose several series of experiments were made. One series was devoted to testing the deformability of soils of natural moisture content and structure, another series treated soils fully saturated. A special series of experiments dealt with studying collapsing deformation under triaxial compression conditions to be discussed below.

The results of studying loess soil deformability at preloading showed that the dependence of the volumetric deformation on the pressure is linear. Such results were obtained both for the soils of natural moisture content and for saturated soils. The Table below gives the moduli of volumetric deformation through the depth of collapsing soil mass.

* Prof. Dr. Sc. M. V. Malyshev—ENG. V. A. Pustogachev, Moscow Civil Engineering Institute, Sluzovaya, Nabereznaja 8, Moscow-M114, USSR

Table

Depth, m	3	7	14	21
Module of volumetric deformation of soils of natural moisture content, MPa	5.3	9.5	18.3	23.0
Module of volumetric deformation of saturated soils, MPa	1.2	2.1	3.3	5.5
Moduli correlation	4.4	4.5	5.5	4.2

In case of clay soils the increment of volumetric deformation reduces with the pressure growth. The linear relationship of the volumetric deformation of loess soils and the hydrostatic pressure may be accounted for by their high porosity; for instance: the porosity coefficient for the soils studied varies from 0.9 to 0.6 with depth. For non–collapsing layers of soil taken at a depth of 27 m and more non–linear relationship was obtained.

The deformability of loess soils essentially increases when they are wetted. As is seen in the Table the moduli of volumetric deformation decrease 4–5.5 times. This change is most considerable at a depth of 14 m and these very layers are more susceptible to collapse than others. It follows that the greater collapsing properties of the soil, the more considerable is the change of the module of volumetric deformation at wetting.

The deformability of loess soil both of natural moisture content and saturated when deviatorily loaded was studied along the line of crushing, the deviatory loading being applied at hydrostatic pressures of various values. Consideration was taken of the depth at which the specimen was taken, as the character of the deformation considerably depends on the prehistory of loading. Special attention was paid to the studies of soil deformability at the hydrostatic pressure equal to γH.

At the start of the deviatory loading in all the experiments non–linear dependence was found (ascertained) and linear dependence at the subsequent loading, vertical and horizontal deformations being essentially non–linear which depends on the density and moisture content of the soil.

The size of the non–linear section of the chart of the volumetric deformation dependence on the deviator depends on the rate of preloading and on the moisture content of the soil. If the hydrostatic pressure is close to γH and the soil is saturated the relationship is slightly curvilinear. Depending on the density of the soil the non–linear section of the chart extends up to the deviator

$$\sigma_1 - \sigma_3 = 0.02 \div 0.05 \text{ MPa}.$$

In case the hydrostatic pressure is lower than γH or the soil is of natural moisture content the non–linear relationship is preserved up to

$$\sigma_1 - \sigma_3 = 0.10 \div 0.15 \text{ MPa}.$$

In both the cases the volumetric deformation makes 1–1.5%. With further loading linear deformation occurs. Such a behaviour of loess soils can be accounted for by their structural strength and the availability of water-colloidal bonds between the soil particles. Until these forces are overcome there exists non–linear relationship, but, further on, linear relationship becomes pronounced, the prevailing role being played by the friction forces between the particles (Fig. 1).

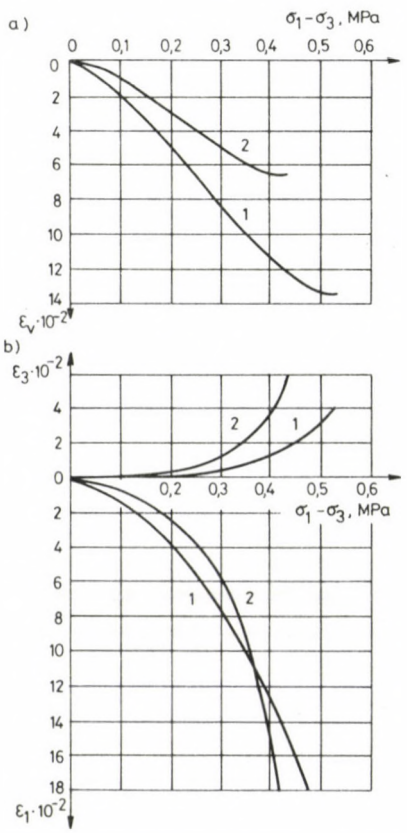


Fig. 1. Graph of loess soil deformation due to deviator; a) volumetric deformation, b) vertical and horizontal deformations; 1-soil of natural moisture content, 2-saturated soil

The deformation of soil at the horizontal pressure lower than γH differs considerably from that described above. Here, the non–linear section is preserved within 1–1.5% of the volumetric deformation, then comes abrupt destruction of the soil, at which vertical and horizontal deformations develop intensively, the volumetric deformation of compaction being observed at it. The value of the deviator when the destruction of the soil occurs depends on the depth at which the specimen is taken and on the value of the horizontal pressure.

The deformation of soil at the horizontal pressure higher than γH is close by its nature to that at the pressure equal to γH, but the deformation module increases.

Figure 2 shows the graphs of variations of Poisson coefficient and the linear deformation module at deviatory loading (both being defined by incrementing deformations for every step of the deviator). For saturated soils the Poisson coefficient at the starting moment of loading decreased from 0.3 to 0.1, then increased, and at a certain value of the deviator the rectilinear dependence was maintained up to the values close to 0.5 I_f the graph of the coefficient variations is compared with the graph of changes of volumetric deformation due to the deviator it is seen that the decrease of the Poisson coefficient corresponds to the non–linear section of the volumetric deformation. In the same way, but at higher values of the deviator, the Poisson coefficient decreases for the soil of natural moisture content. It follows that the Poisson coefficient for loess soil depends essentially on moisture and density.

The linear deformation module for initial steps of the deviator is considerable (ranging from 25 to 6 MPa). It corresponds to non–linear dependence of volumetric deformation on the deviator. Further on, it decreases gradually from 4 to 1 MPa.

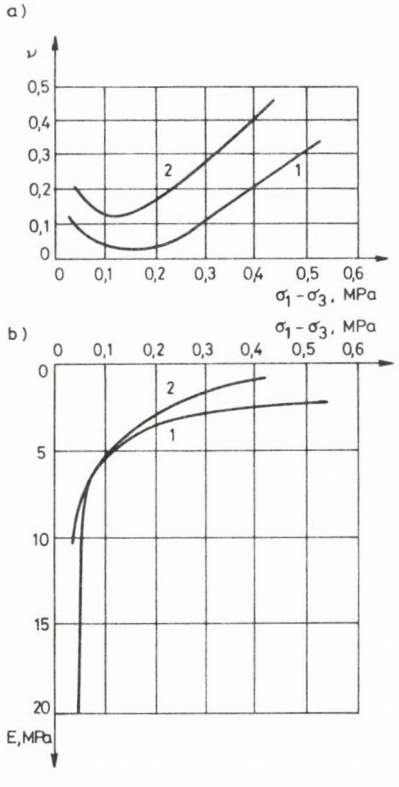


Fig. 2. Variation of deformation characteristics of loess due to deviator; a) Poisson coefficient, b) linear deformation module; 1-soil of natural moisture content, 2-saturated soil

If linear deformation moduli of saturated soil and of that of natural moisture content are compared at the same horizontal pressure equal to γH, they don't actually differ, except the initial and final steps of loading. The reason is the following: being preloaded saturated soil experiences considerable deformation and, consequently, its density increases as compared to the soil of natural moisture content. Though these soils differ in their physical state, the effect of the density and moisture of the soils on their deformability results in the fact that their moduli of linear deformation are actually the same. It follows that when loess soil is wetted its physical state changes, collapse occurs but linear deformation module computed from increments remains unchanged at the linear section of the volumetric deformation variations due to the deviator.

The collapsing deformation was separately studied. The experiments conducted on the triaxial compression apparatus were similar to those made using oedometer when soils are preloaded under the method of single and two–curves. Like in case of using oedometer the volumetric collapsing deformation received under the single curve

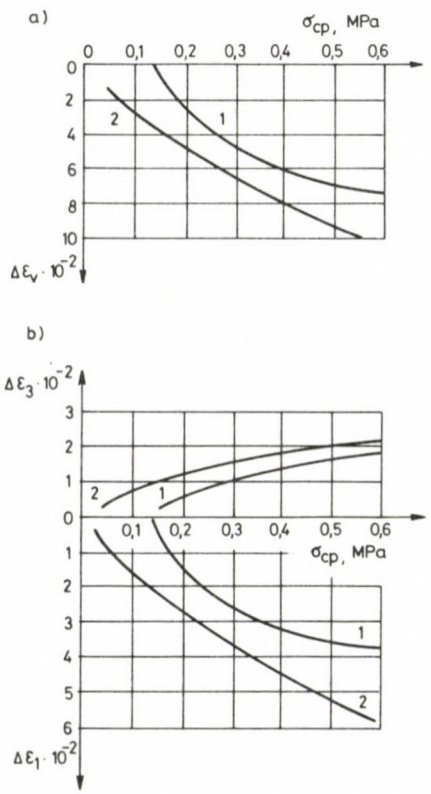


Fig. 3. Dependence of collapsing deformation on hydrostatic pressure; a) volumetric collapsing deformation, b) vertical and horizontal collapsing deformation; 1-under the single curve method, 2-under the two curve method

method is found to be smaller than that obtained under the two curve method. As to vertical and horizontal collapsing deformations they are found using the two curve method in different tests, but under the single curve method the vertical collapsing deformation may be greater and vice versa, which depends on the rate of wetting the specimen. The difference of collapsing deformation obtained under these two methods is accounted for by different effect of water on the soil subjected to no pressure and on that being under pressure.

If comparison is made of the collapsing deformation of loess soil under some pressure in the oedometer with the volumetric collapsing deformation in the triaxial compression apparatus at the same hydrostatic pressure, it is found that in the latter case the volumetric collapsing deformation will be greater. It is explained by the fact that the stressed state in the oedometer is different due to side pressure than in the triaxial compression apparatus at preloading; this difference increases with the growth of pressure.

The collapsing deformation as dependent of the deviator was studied in the triaxial compression apparatus. For this purpose three series of tests were carried out. The soil to be tested was taken at a depth of 14 m where the natural pressure

$$\gamma H = 0.24 \text{ MPa} .$$

In the first series of tests the soil was wetted at different values of the deviator, the horizontal pressure remaining constant and equal to the natural pressure γH. In the second series the soil was wetted at different values of the deviator but here, the vertical pressure remained constant and equal to γH. In the third series of tests the soil was also wetted at different values of the deviator but this time, it was the average pressure that remained constant, equal to γH. The results of the tests are cited in Fig. 4.

The tests confirmed that the collapsing deformation is considerably dependent on the deviator. The volumetric collapsing deformation depends on the deviator and is considerably dependent of the average pressure. If in the first and the third series of tests the volumetric collapsing deformation always increases with the increase of the deviator, in the second series it reduces starting from a certain value of the deviator which occurs due to the decrease of the average pressure. The vertical collapsing deformation depended to a considerable extent on the deviator: the greater the deviator, the greater the vertical collapsing deformation, the latter being dependent on the value of horizontal pressure—the lower the horizontal pressure, the greater the vertical collapsing deformation. The horizontal collapsing deformation in this case may be the deformation of compression as well as the deformation of expansion of the specimen. If the value of the deviator is insignificant, the soil is always compressed irrespective of the value of the horizontal pressure. At larger values of the deviator there occurs horizontal expansion of the specimen.

Thus, the influence of the stressed state of the soil on its collapsing deformation is evident.

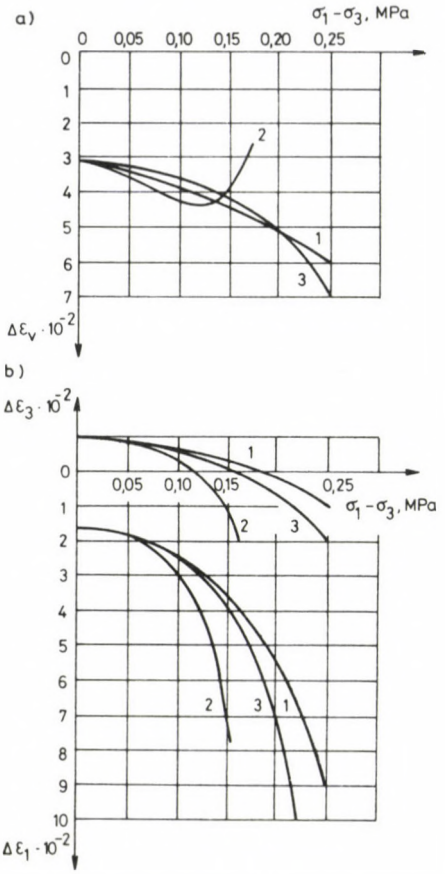


Fig.4. Dependence of collapsing deformation on the deviator; a) volumetric collapsing deformation, b) vertical and horizontal collapsing deformation; 1-at $\sigma_3 = \text{Const}$, 2-at $\sigma_1 = \text{Const}$, 3-at $\sigma_{aver} = \text{Const}$

Acta Technica Academiae Scientiarum Hungaricae, 98 (1—2), pp. 105—113 (1985)

BEHAVIOUR OF PILE GROUPS UNDER LOAD IN GRANULAR SOILS

G. PETRASOVITS*

Presented in this paper are the results of conventional triaxial tests and of a simulated pile model test in triaxial cell, proving the effect of test conditions on the value of shear strength parameters. Furthermore results of laboratory model and in situ tests conducted on pile groups with different length, number, and spacing of piles are presented.

1. Introduction

Considering interaction between pile and surrounding soil, and the effect of this interaction on pile load bearing capacity, investigations of changes caused by piles when driven into the soil are of great importance. Both the load bearing capacity of the piles and its mechanism shall be taken into consideration in accordance with the actual conditions. As is well known, increased emphasis has recently been laid on investigation of point resistance and skin friction, their ratio, and the factors affecting this ratio. The majority of theories consider the soil to be a continuous elastic material, although the deformation characteristics are determined empirically. These theories suppose the angle of shear resistance and cohesion to be constant, independently of the stresses and stress conditions prevailing. These theories agree in that the point resistance increases proportionally to the depth of penetration. However, Kérisel's tests carried out in the sixties prove that, depending on the pile diameter, a constant point resistance not increasing above a certain penetration depth occurs. The value of point resistance and of the depth where it is reached depends significantly on initial soil compactness.

In these theories, the fact that when the pile is driven into the soil, the soil gets deformed and the density increases considerably in the vicinity of the pile is not taken into consideration. At the beginning, the soil gets compacted only under the point where stress concentration takes place while it yields at the mantle with the initial density decreasing. With increasing penetration depth, the yield reduces and after formation of a dense soil core, the soil under the point applies a pressure to the subsequent strata in horizontal direction. The soil is most compact next to the pile, then the compactness decreases symmetrically and after a certain distance from the pile,

* Prof. Géza Petrasovits, Technical University Budapest, Geotechnical Department, Műegyetem rkp. 3, H-1111 Budapest, Hungary

neither deformation nor increase in density can be found. With increasing compactness, not only the stress conditions change but also the modulus of compressibility increases, this increase being 10—15 fold under the point while 5—10 fold next to the mantle. The existing theories leave these variation of soil characteristics out of consideration.

A theoretical method to determine the distribution of stresses around the pile is, therefore, necessary, making use of the change of compressibility modulus of soil. This is particularly important in case of pile groups, especially if the distance between the piles is less than 3—4 d (diameter).

As is well known from the literature in agreement with many authors' opinion, soil deformations caused by pile driving can be detected at distances of up to 3 d.

The first works concerning load bearing capacity of pile groups are connected with Kézdi's name. In his theoretical and practical studies between 1957 and 1959, he studied the distribution of horizontal stresses around piles using the passive Rankine state, determined the variation of skin friction–point resistance ratio versus load, and set up an equation for the value of friction resistance along the mantle of pile.

Kézdi carried out in-situ large-scale model tests to study the behaviour of pile groups of different size under load The results showed that in granular soil and with minimum distance between the piles within a group, the soil fenced by the piles got compacted to such a degree that it settled together with the piles when load was applied to the pile group. Therefore, the load bearing area in the depth of pile points increased significantly and the so called 'pillar effect' occurred.

Laboratory tests were carried out by the staff of Geotechnical Department, Technical University Budapest, with a view to investigate the variation of density, and modulus of compressibility, of the soil between the piles versus distance between piles, and to determine numerical values for the variation of compactness and compressibility modulus. Using the finite element method and experimental data, efforts were made to develop a generalizable theoretical method permitting the variation of compressibility modulus due to pile driving to be taken into consideration.

2. Effect of the test method on the value of measured parameters

As is well known, the technology applied affects the load bearing capacity and the interaction between soil and piles considerably. Depending on the construction technology used, the bearing capacity of engineering structures for foundation might vary considerably, even in case of uniform soil. The effect of technology on both load bearing capacity and settlement is extremely important in case of foundations in considerable depths.

The interaction between soil, structure, and technology for shallow and deep foundation is schematically illustrated in Fig. 1.

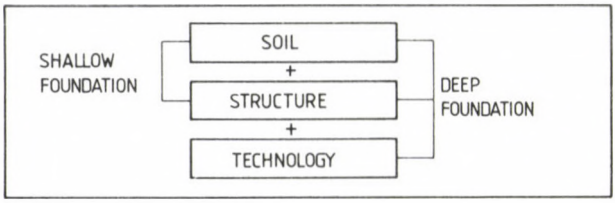


Fig. 1

The values of load bearing capacity calculated on the basis of laboratory test results are often considerably lower than those obtained in in-situ tests. This difference can be attributed to changes in stress conditions due to pile driving, and to mobilized shear strength along the mantle. It is well known that a displacement of some millimeters is enough to mobilize the shear strength in dense soil while a multiple of this displacement is required in soft soil.

Tests were carried out using a model pile built in a large-diameter sample, and tested in a triaxial test fixture, to simulate shear resistance along the pile mantle in different soils. The same soil was tested triaxially also under normal conditions.

In non-cohesive soil, the coefficients of shear resistance obtained in conventional triaxial or direct shear tests are considerably lower than those calculated on the basis of pile-force. Because of the limited displacement, there may be an increase of 40% in the angle of internal friction for non-cohesive soils, depending on the initial density.

Figure 2 shows the results of a series of experiments carried out in sand silt and clay to determine the angle of internal friction. Triaxial tests were made and the angle of internal friction was found to vary between 30° and 34°, depending on initial compactness. A model pile of a diameter of $d = 28$ mm was driven in a soil sample of the same compactness and of a diameter of $D = 100$ mm and a height of $H = 200$ mm and a constant volume test was carried out. The surface of the model pile was coated with synthetic resin and a layer of the sand tested. Thus shear takes place between sand and sand. For constant volume, considerable higher values were obtained for internal friction, amounting to $\Phi = 34°$ to 48°, depending on initial compactness. An explanation to this change is the phenomenon of dilatation.

Namely, after shear, a certain kind of dilatation occurs. As a result, the horizontal stresses in sands increase considerable. In silt, this phenomenon is less appreciable while it is not observed in clay at all, that is the same values were obtained for the angle of internal friction in all methods. The angle of internal friction has been determined in relation to the initial diameter of the model.

Figure 3 illustrates the compacted zones in a group of 9 piles. As shown in the Figure, in case of a centre-to-centre distance of 5 D the extent of intersections is reduced, and there is only a single overlapping of the zones at a relatively large distance from the different pile axes. If, however, there is a spacing of 2.5 D between the piles driven in the soil, overlapping of the zones is multiple. It can be clearly seen in the pattern so obtained that the different piles within the group are surrounded by zones

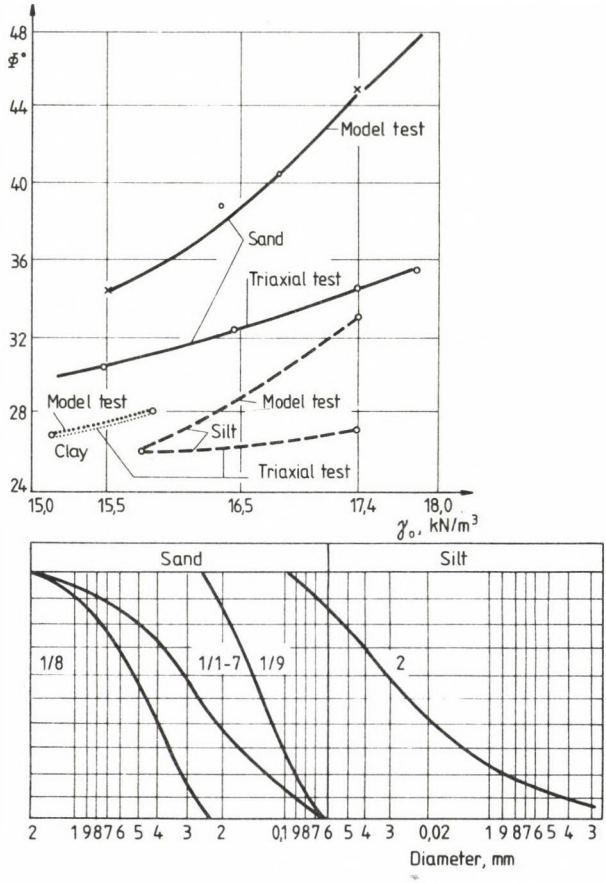


Fig. 2

with significantly different densities. The number of overlaps is largest at the central pile, reducing at the so called outer piles,the smallest number of overlaps being found at the corner piles. With the investigation extended to pile groups containing 16 and 25 piles, it can be seen that there is no increase in the number of overlaps nor a change in the conditions for the 4 piles in the centre of the group of 16 piles as compared with those in the centre of 9-pile group. In a group of 9 or more piles, three different pile types can be distinguished such as internal piles designated I, outer piles designated 0, and corner piles designated C. On the basis of the number of overlaps, no numerical values can be determined for the changes in density of the surrounding soil but it indicates the role of the different piles according to their position within the group.

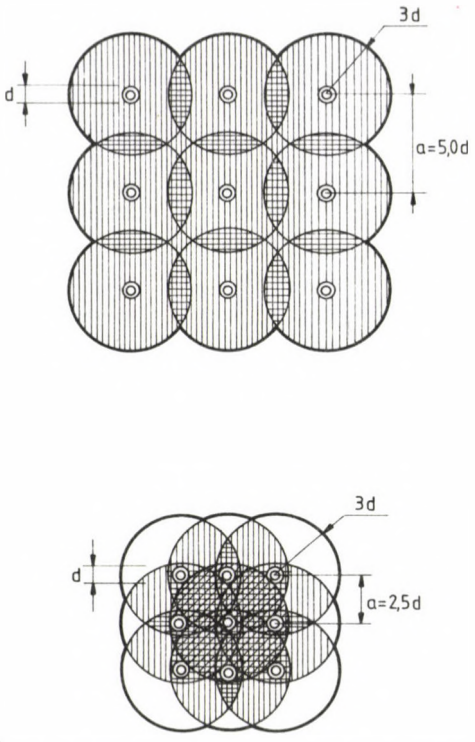


Fig. 3 Overlapping of compacted zones for different pile distances

3. Effect of pile length and pile spacing on the load bearing capacity of pile groups

In spite of the fact that piles are used usually in groups, the behaviour of pile groups under load has not been comprehensively investigated so far. Laboratory tests, but first of all field tests included the investigation of a few factors only, and a number of questions concerning behaviour of pile groups remained unanswered. Because of the scientific and practical importance of the problem, the authour tested pile groups with a view to study the behaviour of the individual piles in the pile group under load. Studied were

— the variation of magnitude and ratio of point resistance and skin friction versus group size (number of piles within the group),
— effect of distance between piles,
— effect of pile length,
— effect of location of pile within the group,
— components and magnitude of efficiency coefficient for different pile groups.

In laboratory experiments, model pile groups of $n = 4, 9, 16, 25$ piles with a length of $H = 10, 20, 30, 40$ d and a spacing of $a = 2.5, 3.33, 5.0\ d$ were tested in homogeneous granular soil. The experiments included 140 tests with model piles to investigate the behaviour of pile groups. Each pile was equipped with a load cell at either end and driven into the soil.

The load resulting in 10 mm pile penetration was considered to be the limit load bearing capacity of the pile group. The load applied to the piles was increased gradually, and settlement, load taken by the piles and that taken by the point of the piles were measured continuously for each step or load increase. Skin resistance of the pile was given by the difference between total load applied to the pile, and the load measured at the point of the pile. This method permitted the magnitude of point resistance and skin friction and their ratio for different levels of utilization factor of load bearing capacity to be measured.

Load was increased in steps of about 1/10 of the estimated limit load. Increase of the load was continued after no settlement had been detectable over 10 minutes under actual load.

Before analyzing the effect of factors mentioned in the previous chapter, the author not only had to work up and arrange the abundant experimental data but also to isolate these effects bearing closely upon each other while at the same time including them in the same complex of questions and, finally, to represent them in a demonstrative form.

The problem, difficult enough, was still more complicated by the fact that each factor proved to be significant in respect of behaviour or both pile groups and individual piles under load. Any change in one factor modified the effect of other factors.

The results of experiments concerning the effect of different factors on the load bearing capacity of pile groups are given below.

As seen in Figs 4a and b showing load-settlement curves for pile groups with piles of different length ($H = 20d, 40d$), pile groups with a larger number of piles in them and a smaller spacing ($a = 2.5d$) have not reached their load bearing capacity at a settlement of 16 mm. This can be attributed to the large number of internal piles (I) as both the corner piles (C) and outer piles (O) have reached their load bearing capacity. The marginal piles in the group fenced the compacted soil around the internal piles and loaded the soil on the pile point level, thus contributing to the load bearing capacity of the group due to the significant increase of their point resistance.

For groups with a larger number of piles, the settlement is 4 to 6 times as much as for smaller groups, the degree of utilization being the same. This results first of all from the increased layer thickness of soil involved in load bearing because the soil under the pile point gets compacted significantly through a layer of a thickness of two pile diameters. The extent of settlement depends largely on initial soil compactness. Considering the limited permissible settlement for buildings, the load bearing capacity of groups with a larger number of piles in them can not be utilized.

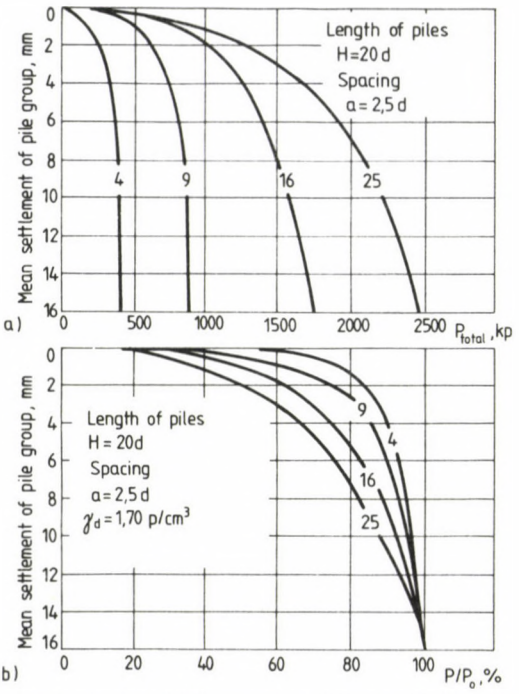


Fig. 4. Load bearing capacity for groups of different size and degree of utilization of load bearing capacity, $H = 20\ d$, $a = 2.5\ d$

a) Load-settlement curves for groups of $n = 4$, 9, 16, 25 piles

b) Settlement of pile group versus degree of utilization of load bearing capacity

A comparison of the results of experiments with pile groups of different size and pile length permits the following conclusions to be drawn:

— Increasing pile length results in a higher rate of increase in load bearing capacity of the group with, however, the settlement not increasing.

— In case of two pile groups of identical layout and loaded in the same way, increased settlement occurs for the group where more load is transferred to the soil by the point of the pile.

This suggests obviously that longer piles are preferable in homogeneous soil.

Figure 3 shows the compacting effect of piles and pile groups driven in the soil. It can be clearly seen that the compression around piles I, O, C is different, depending on pile spacing.

Soil density within the area surrounded by piles driven at a spacing of $a = 2.5\ d$ increases by about 20 to 30%. As a result of this compression, the load bearing capacity of internal piles increases significantly, the outer piles having a higher load bearing capacity than corner piles.

This effect increases with reducing spacing. Figure 5 shows the load bearing capacity and point resistance ratio of characteristic piles of a group of $n = 9$ piles driven

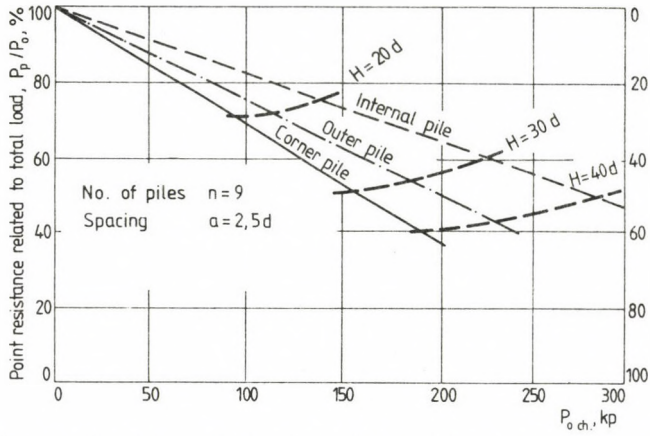


Fig. 5 Load bearing capacity of characteristic piles, and point resistance/total load ratio for different pile lengths

at a spacing of $a = 2.5\ d$. On the basis of the Figure, the following conclusions can be drawn:

— Within a group of piles, the load bearing capacity is lowest for the corner piles.
— The point resistance ratio is lowest, and the effect of skin friction is highest for corner piles within a group.
— The load bearing capacity of internal piles (I) is by 40 to 50% higher than that of corner piles.
— With increasing load bearing capacity of internal piles, both the point resistance and the skin friction increase, the increase in skin friction amounting to 15 to 20% while in point resistance to 40 to 80% as compared with corner piles. Responsibility for increased load bearing capacity lies in majority on increasing point resistance.
— The behaviour of outer piles (O) represents a transition between internal piles and corner piles, the mantle friction of outer piles being by 5 to 10% while their point resistance by 20 to 30% higher than the same values for corner piles.

The considerably higher skin friction of internal piles and outer piles can obviously be attributed to compacted zones around the piles while the significant increase in point resistance results from the compacted soil fenced by outer piles and corner piles as well as from the load applied to the soil by the adjacent piles.

In a 9-pile group with a pile spacing of $a = 5.0\ d$ (Fig. 6), the behaviour of the different piles is rather indistinct although the load bearing capacity of internal piles is higher also in this group. If the spacing is increased, the ratio of point resistance will reduce. Here the nonspecific behaviour of the different piles is due to the fact that within a group with a pile spacing of $a = 5.0\ d$, each pile behaves like an independent pile because of the poor overlap of the compacted zones. Increase in the bearing capacity of internal piles results from the increased point resistance.

The results of tests concerning effect of pile spacing suggest that the load bearing capacity is favourably affected by the group effect in case of a pile spacing of $3.0\ d$ or less.

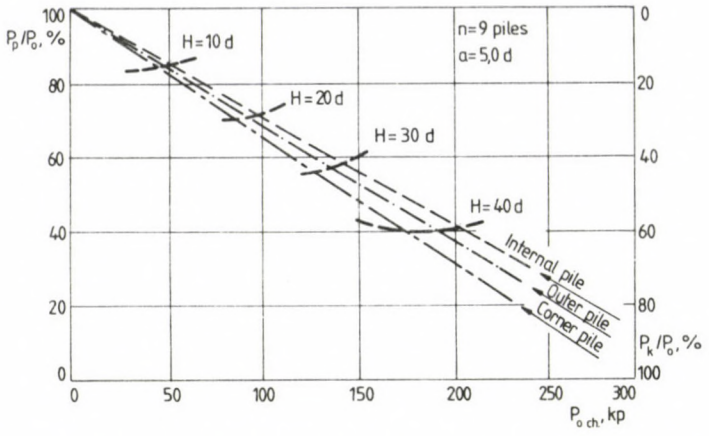


Fig. 6. Load bearing capacity for characteristic piles, and point resistance ratio for different pile lengths

4. Analysis of the results of in-situ pile group tests

In-situ test supervised by Prof. Széchy were carried out, driving piles of a cross section of 12 × 12 cm and a length of 2.5 m and 3.5 m in silty sand soil. The piles were driven first to a depth of 20 d (spacing: $a = 12$ cm) and test loaded, then driven to a depth of 30 d and loaded again. In this way, not only the group effect but also the effect of length could be studied. Included in the tests were groups of 4, 5, 7, 9 piles and with two different spacings. The spacings were so selected that the area fenced by the piles would remain the same within series of tests, using $a = 5$ to 9 d in one series while $a = 2.5$ to 3 d in the other series (see Fig. 7).

The load-settlement curves obtained show that the increase in pile length affects the load bearing capacity favourably. Within the same test series, the load bearing capacity increases proportionally with the number of piles in a group and with pile length. Maximum point resistance occurs for a pile length. of 20 d, and no increase of point resistance is brought about by an increase of the depth of driving. The increase in load bearing capacity is higher for a small-spacing group of type S than for a group of type G of larger pile spacing due to the increased compactness of fenced soil. The tests proved that the load bearing capacity is higher for a group of piles of number n than for independent piles of number n.

It can be seen in Fig. 7 that the load bearing capacity per pile is higher in group S than in group G. The specific load bearing capacity per pile in group S is by 25% while in group G by 10% higher than that of independent piles, the driving depth being $H = 30$ d in each case.

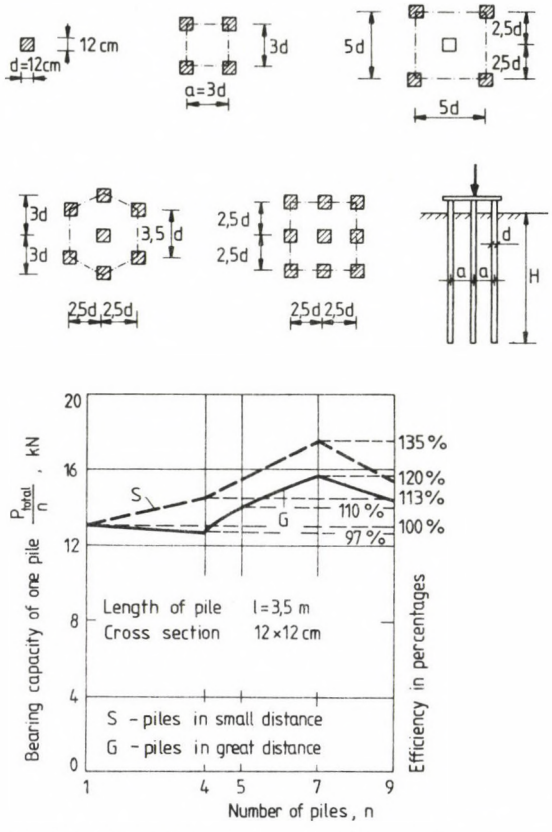


Fig. 7. Load bearing capacity and efficiency of a pile group

Acta Technica Academiae Scientiarum Hungaricae, 98 (1—2), pp. 115—136 (1985)

FLOW PRESSURES ON PILES AND PILE GROUPS*

K. STEINFELD**

Behaviour of piles and pile groups subjected to lateral earth pressures due to soil flow and creep is analysed in this comprehensive study. Experiences gained at construction of embankments, bridge abutments, slope restoration are collected and presented. Results of earlier and the newest theoretical investigations are also shown. Author points out that the Gudehus — Leinenkugel's method seems to be the best to determine the flow pressures on piles.

1. Introduction

Unfortunately, a soil seldom is seen to flow and creep, namely its displacement proceeds too slow to be observed.

Even after protracted, important deformation paths, flow phenomena are mostly not recognized on the terrain by its rough, embossed surface, often permanently altered by vegetation. Only cracks, displacements at crack surfaces are more perspicuous, that may, however, not arise even for displacements of the ten cm order, and are soon concealed by secondary deformations, traffic, rain, plant cultivation, etc.

Flow and creep are normally absent or negligible in brittle, high-strength soils such as rock or hard clay. Even in granular soils they hardly develop because of the high shear strength or but slightly near the ultimate strength, but here they are instantaneous, because of the absence of temporary strength changes due to swelling or shrinkage.

On the other hand, all cohesive soil types are prone to flow as a function of softness, the more protracted, the finer grained they are. These processes last for decades in clays, even, introducing the concept of "delayed compression" suggested by Bjerrum [1], for centuries or millennia.

Creep and flow processes in these soils, independent of the stress state, are often described in terms of the concept "rheology" i.e. mechanics of viscous fluids. For soils it is meant as relative flow displacements between soil particles.

Of course, millennia are no time for geohistory; it's millions of years that count. That is why there are only a few soil types of this cohesive, soft class among the small group only a few millennia old.

* Lecture at the annual meeting of the Federation of Structural Testing Engineers, September 19th, 1983, in Kassel, FRG.

** Prof. Dr. Ing. Karl Steinfeld, Beratender Ingenieur, VBI, VDI, ASCE; Alte Königstrasse 3, 2000 Hamburg 50, BRD

These are normally termed alluvia, infrequent already in medium highlands of Germany, as e.g. flood plain soils in river valleys, moors, sometimes as slope flow soils in form of secondary sediments, erosive soils and the like.

But in all alluvial zones of the world, in marshlands just as in North European coastal areas, in extended river estuaries often accommodating important harbors with their living areas, there are complete areas recently arisen and in continuous development.

Since times immemorial, stable constructions on soft, plastic soil, and even in shallow water have been built on pile foundations, as so-called pileworks, from Borneo to Lake Constance.

Piles are applied as foundation structures for an onshore construction if loads of a solid building mass have to be transferred through a soft, plastic subsoil to a deeper, solid ground.

Eventual embankments or spread foundations on the soft layers adjacent to the pile foundation not only compress but laterally displace them, arising flow phenomena up to soil failure penetrating the pile foundation; they are the most conspicuous in form of landslides in slopes at pilings under bridge abutment piers.

2. Large-scale experiments in the Netherlands

Netherland's lowlands have always been areas predestined to these experiments. In towns such as Rotterdam and Amsterdam, attention was paid to inherent risks of hazard already before World War II. This is why building authorities required — in addition to possible upfills or complementary surface loads (independent of the reinforcement for vertical loads and driving stresses) — a reinforcement for the maximum possible bending moment, or for 5 Mpm in common r.c. piles, usual at that time.

Large-scale experiments made in the '50s by the municipality of Amsterdam showed bending stresses in piles hence flow processes in cohesive and organic soils to have an importance much exceeding that assumed earlier.

These tests were made and evaluated by Heyman and Boersma [2]. Figure 1 shows a planned road embankment of 60 m base width and 7 m high, with three steel piles drive by a pile ram at 30 m in front of the fill, each joined by a mobile tube accommodating deflectometers for measuring soil displacements due to further fillings. Bending moments were measured in the relatively rigidly clamped piles.

Soil stratification is seen in the left part of Fig. 1. At a depth of about 12 m, dense, older diluvial glacial sand starts, superposed by sandy, here and there very sandy clays, underlying, in turn, besides of two marked peat layers, a natural, loose top sand cover 2.6 m thick. This is not so bad a soil condition as usual in alluvial marshlands in German coastal areas.

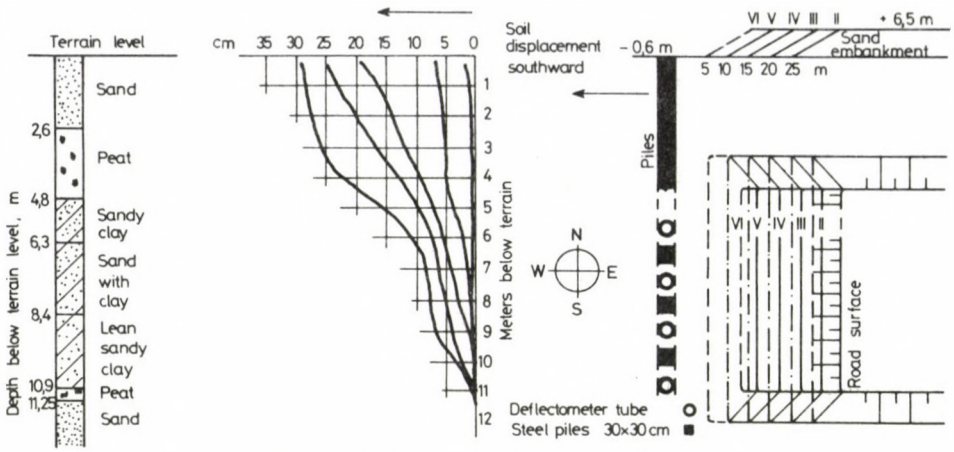


Fig. 1. Horizontal soil movement according to Heyman and Boersma [2]

After having incorporated the testing equipment, the sand embankment was further filled up in 5 m sections each two weeks, up to a distance of 5 m from the piles and tubes.

Representation of the soil displacement vs. depth measured by deflectometers in the tubes shows a 2 cm displacement at the surface of the tested cross section initiated by the 5 m wide base of the first fill spaced at 25 m, increasing to 30 cm for the last fill section spaced at 5 m from the test piles. The top sand is fully floated on the top peat. Just this last phase shows the full displacement to actively penetrate the top peat.

Let us notice first in general that, in turn, in sand zones before a rigidly built–in mass, much higher resistances develop than in cohesive, soft soils, and second, that these displacements have no soil failure or the like as concomitant. Solely slight heaping phenomena were felt on the surface in front of the practically rigidly clamped piles 30 cm wide, developing bending moments seen in Fig. 2.

During filling, bending moment maxima ranged from 20 kNm (or 2 Mpm) at a distance of 25 m to 140 kNm (or 14 Mpm) at 5 m, and appeared at a depth of about 2.5 m, that is, at the interface of top sand and top peat. The mentioned 5 Mpm moment specified in Amsterdam was exceeded in the pile already at a distance of 23 m from the embankment base.

The presented figure has been taken over from a 1967 lecture [3] in this scope, thus applying the previous units.

The dotted line presents the estimated line load Mp/m per running meter of pile length, originally rather inaccurately determined from the supporting force maxima of 30 to 130 kN [2], because of the moment distribution unknown to the Author.

The line load on a 30 cm pile exceeds 2 Mp/m already at a distance of 25 m, and 5 Mp/m at about 7.5 m from the embankment; values to be reminded of later.

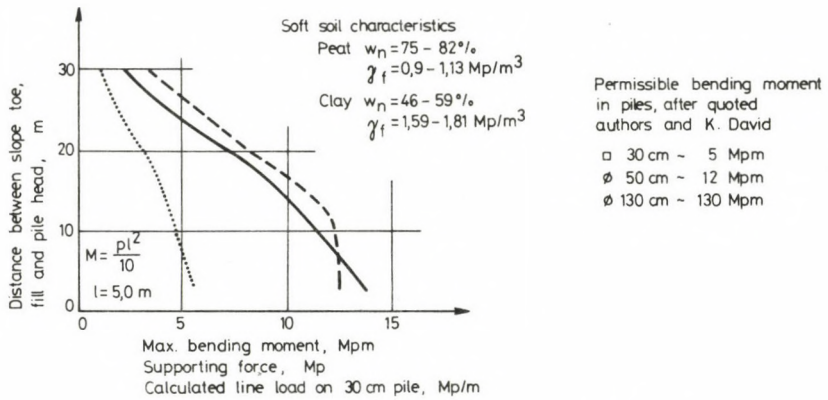


Fig. 2. Large-scale test results by Heyman and Boersma [2]

Soil characteristics are seen in the top part of Fig. 2, unfortunately only moistures and densities are given in [2]; shear values are missing.

The top right part of Fig. 2 shows bending moments supported by r. c. piles. Accordingly, a 30 cm pile can absorb 50 kNm = 5 Mpm measured at 23 m from the embankment, a 50 cm in–situ concrete pile about 120 kNm, and a ∅ 130 cm large bored pile some 1300 kNm.

3. Further measurement and test results

Leussink and Wenz [4] refer to further examples of similar displacements happened at Klöckner-Werke, Bremen (FRG). Figure 3 taken over from [4] serves only to exemplify 30 cm deflections along the flow line measured on 114 cm high "Doppelpeiners". Here also consternating failures causing collapse in the pile foundation arose.

Franke and Schuppener [5] present measurements in the area of highway BAB 7 near Hamburg similar to those in Amsterdam (Fig. 4). In a rest time of 2.5 years, displacements were measured almost continuously from the surface to a depth of 5 m in 8 m clay and peat in front of a 7.5 m high stepped spill dike mass, sloping across the steps as little as 1 : 5 (Fig. 5) such as:

at 8 m from the embankment base	12 to 19 cm,
at 15 m from the embankment base	4 to 6 cm,
at 22 m from the embankment base	3 to 5 cm.

Also the outermost row of large, ∅ 1.5 m bored piles for the highway was at 22 m from the spill dike base. The moderate displacement values measured there, and the great pile dimensions do not admit critical bending stresses, as demonstrated in [5].

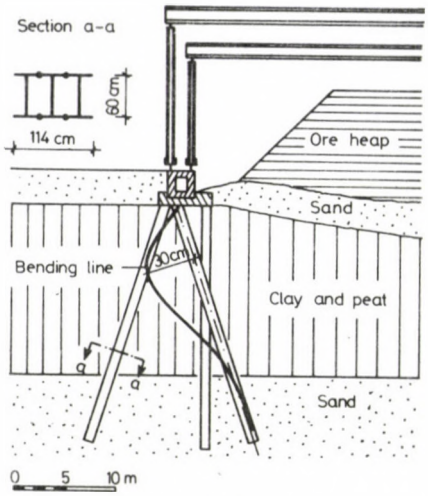


Fig. 3. Craneway foundation on skew piles with bending line after Leussink and Wenz [4]

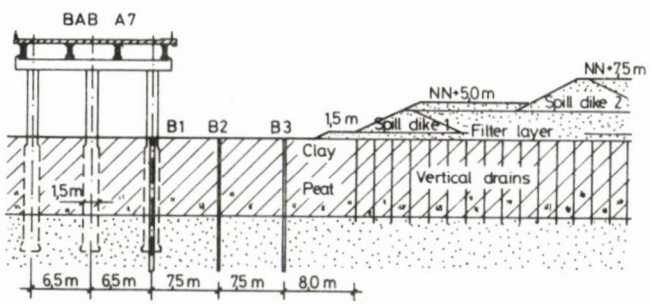


Fig. 4. Section of Highway BAB 7 after Franke and Schuppener [5]

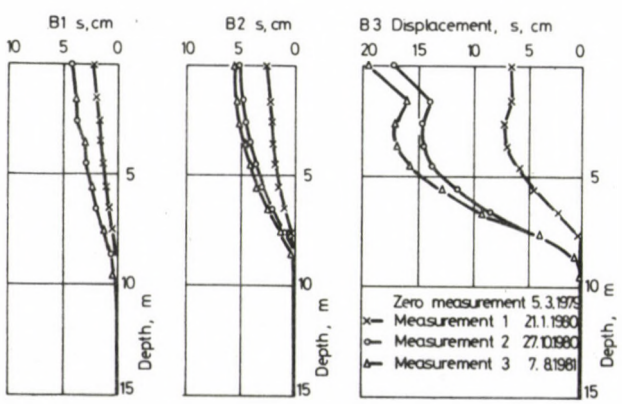


Fig. 5. Displacement measurements in section 1 after Franke and Schuppener [5]

By the early '70s, Nahrgang [6] made model tests to see how far horizontal flow displacements propagate from a surface load. Applying the boundary value problem on an incremental load of width B on a soft subsoil of finite depth H and a size across limited to 5 B, horizontal displacements vs. subsidence of the given incremental load shown in Fig. 6 arose.

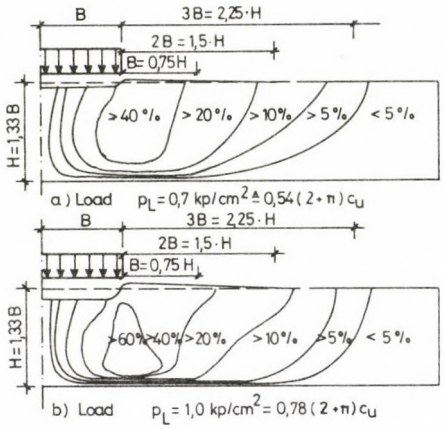


Fig. 6. Horizontal displacements referred to subsidence after Nahrgang [6]

The load has been referred to the well–known Prandtl ultimate semi–space load [7]:

$$p = (2+\pi)c_u \quad [\mathrm{kN/m^2}]$$

where c_u is the undrained soil shear strength determined in vane tests. Wenz [8] applies corresponding Prandtl's formulae for the ultimate load capacity of a soft full-space subsoil.

Deformation diagrams in Fig. 6 have been obtained by applying a special stress/strain relationship between elastic strain and purely plastic flow developed by Nahrgang as a special material law based on peculiar soil mechanic tests.

As to be seen in Fig. 6a, for 54% of the "Prandtl-load", hence, for a subsidence safety $\eta \approx 2.0$, the horizontal displacement is at the load edge about 40% and at a distance of 3 B 5% of the subsidence under vertical load. In Fig. 6b, the load amounts to 78% of the subsidence load, that is, $\eta \sim 1.3$ after Prandtl, the horizontal displacement is 60% at the load edge and at a distance of 3 B, only 5% of the subsidence under vertical loads; anyhow, the vertical subsidence is of course significantly higher under 78% of the "Prandtl load".

Numerical FE-calculations on an example yield for a soft layer 10 cm thick uónder half the Prandtl load, i.e. $\eta = 2$, a vertical subsidence of 20 cm (2% of the thickness), under the load edge a horizontal displacement of nearly 10 cm, and at 23 m still a displacement of 1 cm (5% of the 2% of thickness).

The subsidence is already 60 cm under $\sim 3/4$ of ultimate load hence $\eta = 1.3$, horizontal displacement is 45 cm under load edge, and 4 cm at a distance of 23 m.

The example shows critical horizontal deformation values near the loaded area edge (embankment base) to occur much before reaching the usual soil failure safety factors, demonstrating, in addition, the displacement fields calculated here according to the special material law and the FE-program to correspond to the velocity field developed by Prandtl as early as in 1920, provided the soft layer depth is 1–2 times the represented half-width of the loaded area.

4. Earlier, quasi-classic concept of the effect of horizontal earth forces on piles

Some examples of bridge abutments selected by the Author earlier, at random [3] are seen in Fig. 7.

The first example is a bascule bridge at Stutthof near Gdansk, much worrying young engineers Dr. Erlenbach and Wodtke (becoming later leading road authority officers), and making them engaged with this problematic for a life.

Already at the construction and backfilling of this engineering structure, the bascules got constantly clamped, requiring 10 cm to be cut off. The right-side abutment with pier footing and pile heads was displaced by 10 cm to midstream, for not at least unusual slopes below the bridge, 1 : 1.5 above, and 1 : 3 below water level. The access road upfill amounted to some 7 m. Represented bridges include the bridge across the river Eider at Breiholz in Schleswig-Holstein, so to say helped to publicity by Leussink and Wenz [4], where the abutment turned back obviously upon soil failure–like flow strains, and subsided. Here also, slopes ranged from 1 : 1.5 to 1 : 3. The road fill was here max. 6 m.

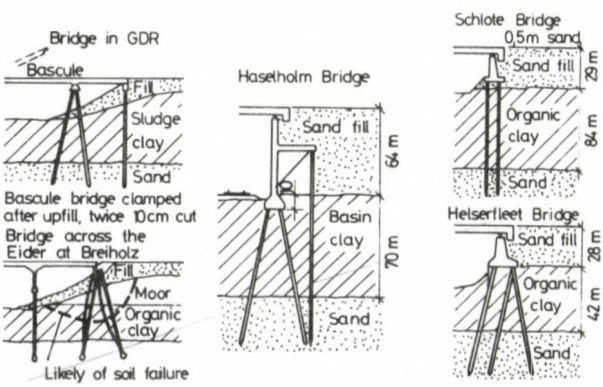


Fig. 7. Examples for the bending of bridge piles [3]

Up to the '60s, foundation engineering maintained the view that earth pressure bilaterally on piles was about equal also in slopes, stable in themselves, exerting no force on the piles. Sometimes it was conceded that in slightly flowing slopes little resistance could develop before slender piles. On the other hand, behind a pile pressed downslope, soon a much higher soil resistance would develop. All these in spite of clause 5.1.2 "Permanent Soil Loads" in DIN 1072 inserted in the development period from tentative in 1963 to finalized in 1967:

> "Earth pressure loads on thin structural members (e.g. piles, piers, slices of sectioned abutments) *standing in slopes* — provided no closer confirmed assumptions have been made — have to be determined as:
> on structures max. 1 m wide for three time the width;
> on structures 1 to 3 m wide for a width of 3 m;
> on structures wider than 3 m for the real width.

For piles driven or bored in grown, stable or previously filled and carefully compacted stable soils, direct earth pressure load on the piles may be ignored, *provided the soil is able to absorb earth thrust in itself*, without taking the bending stiffness of piles into consideration."

This part of DIN 1072 has been applied for little else but lost abutments "in slopes" — usually passing over the last sentence "provided the soil is able to absorb earth thrust in itself".

Thus, while earth forces on waterfront structures have always been considered throughout the height of all the construction, and the safety against landslides demonstrated by a gliding plane directly below the bulkhead base, free–standing pile grid and bridge abutment constructions were examined for the absorption of earth thrust usually only down to the lower grid edge, and the safety to soil failure was checked as above, by a gliding line below the pile tips.

Thus, checking the earth force course i.e., the safety to soil failure between bottom edge of abutment wall and piling down to pile tips had been missing.

In the domain of piles and the necessarily (because of interstices) free slopes before them, stability was tacitly presumed, just as that nothing could happen to the piles. It was even pointed out — partly with right — that piles prevent slides, without taking stress and strain constraints in them into consideration.

The subsequent three cross sections are those of bridges where the leading bridge construction officers of Schleswig-Holstein, Horch and Wodtke had the entire construction designed for earth thrust.

Without further details, it is obvious that a significant earth pressure acts at least in level with the pile head, counteracted by zero soil resistance because of the soil missing in the subway zone.

5. Recent approach to horizontal forces acting on piles

Continuous design for soil pressure differences was the first step to make up the "structural gap" in the design of abutments on piles.

Earlier experiments by Jäger [9] and Förster [10] made in the '20s and '30s on relieving bulkheads from soil pressure by means of upper pile grids already showed them to absorb 90% of the earth thrust if built in sandy soils at a ratio of 0.3 i.e. spaced at thrice the pile width, and to fully absorb it if spaced at twice the pile width.

In this case the embankment wall or apron wall acts only as sealing against soil outflow or dribble and not as a supporting structure if piles are of adequate design.

Similar expectations may be made in soft, cohesive, mainly clayey or organic soils. Unfortunately, it is not so easy to adequately design piles in soft soils for lateral pressure. Such soils may simply flow away, either piles have or have not been exposed to the assigned earth thrust difference, in particular, for larger pile spacings. Irrespective of the earth thrust value dependent on the imposed load and on the shear strength, the flow around piles depends of course also on the soil stability itself, thus, on the safety of shear failure, that is, on the shear strength of cohesive layers and on the slope inclination before pile foundations, namely there is often no bulkhead in front of e.g. bridge abutments.

For soft, cohesive soils, at least if exposed to high loads, Peck and Raamot [10] suggest design assumption of a perfectly plastic material — rather for safety, namely the assumption of an ideal viscous fluid is quite safe and leads to no utilizable result, in particular for high loads. Leussink and Wenz [4] state strains to prevail over stresses under (usually heavily loaded) storage areas on very soft clays and peats, suggesting strains to be taken as design criteria.

Safety of a piling should primarily refer to members most affected by high soil strains and the most likely to fail.

Examples for bridges in Fig. 7 clearly show the piles to be the weakest (most slender) members in the construction, and the most likely to fail under the effect of lateral soil strains, they being originally conceived as compressive or tensile bars.

In final account, pile deformations depend on the force applied by the surrounding soil, and on its power for a path long enough; thus, essentially, on soil shear stresses and shear strains, hence on the stability of the slope. It should be pointed out that even research on non-linear material laws greatly intensified by computerization in the last decade could not produce stress/strain hypotheses generally valid for all soil materials. Nahrgang in his work referred to [6] suggested a special material law for a soft soil, useful for certain boundary values of the model. In spite of the fair approximation of the test soil between conditions of elasto-plastic rigid failure and of a viscous fluid, no image is obtained of its behaviour in shear vs. e.g. consistency.

Wenz [8] has been concerned with loads and consistencies where critical lateral forces arise in piles, in particular, with the determination of forces arising upon the flow of soil past the piles.

Flow rate and path essentially depend on the safety to soil failure, they are the higher, the closer to the ultimate safety, hence to instability in the classical meaning of the word.

The safety to terrain or soil failure seems to be the first criterion to define the risk degree and range.

6. Considerations on soil landslides

Stripping e.g. bridges — represented in Fig. 5 with longitudinal sections — of their structures leads to soil sections in Fig. 8.

Engineers practiced in dam and dike constructions in marshlands would obtain surprising aspects. A dam of inflected slope 1 : 1.5 to 1 : 3 on organic marshy soil is not stable at all or only if filled in steps with protracted consolidation intervals. Constructions of dikes even with continuous slopes 1 : 3 with 3 to 7 m fills as in the examples above often cause shear failures becoming less frequent only for 1 : 4.

Obviously, embankment slopes "concealed" here by the structure have to be considered as critical. All critical gliding planes in Fig. 6 reach much below the grid structure, causing bending stresses in the piles due to displacement of the gliding soil mass. By the way, of course, such shear failure analyses on the "undressed" soil section require soil pressures (as active forces) absorbed before by the superposed closed abutment wall to be restituted as reactions in the considered vertical soil section parts, to be reckoned with in design (usually as moment of resistance) (see 4.2 in [13]).

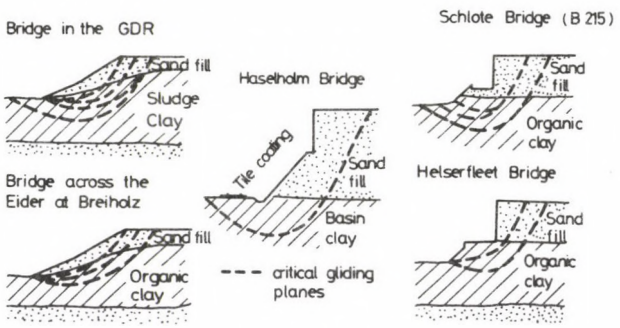


Fig. 8. Stripped soil sections for bridge examples [3]

Occurrence and knowledge of important soil failures in major ore mines by the turn of this century is apparent from Fig. 9 due to Peck and Raamot [11]; other interesting examples of pile failures due to soil flow have been described by Leussink and Wenz [4], and by Neumayer [12].

It is interesting to see the risk not to be restricted to values below safety limits admitted in DIN 4084. While in Kiel, 1967, the Author considered safety factors $\eta = 1.3$,

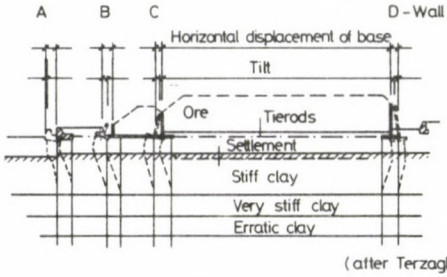


Fig. 9. Soil failures near ore deposits after Peck and Raamot [11]

and $\eta = 1.5$ for cohesive soils of consistencies $k > 0.5$ and $k < 0.5$, resp., to be satisfactory against soil failure (except for cohesive, organogenic soils with extremely high water contents, reacting a priori as viscous fluids) as criteria in pile design for yield pressure — matching the views of all engineers practiced in this field, and against economical considerations — and put it on issue [15], [3], in 1972 Wenz was right in pointing out [14] that, according to Nahrgang [6], even for higher safety factors, inadmissible high strains take place near piles, applying already earlier full flow pressures on them.

This is why the Committee specified under 4.2 of its recommendations [13] even a safety factor as high as $\eta \geqq 1.8$ for heavily organic soil with an ignition loss $v_{gl} > 15\%$ and moisture content $w \geqq 75\%$ as criteria to test whether considerable lateral pressures on piles may be expected or not.

For consistencies $I_c \leqq 0.25$ — in agreement with Peck and Raamot [11], Leussink and Wenz [4] and the Author [3] — the expected important soil lateral displacements justify to examine piles for the absorption of flow forces and of the entire soil pressure excess in every case.

At a difference from DIN 4084, for any other cohesive soil type, a safety factor $\eta \geqq 1.5$ against earth sliding is required if specifically checking the piles in bending shall be avoided.

For safety factors below the above ones for each soil type, piles must be designed for lateral pressure.

7. Determination of flow pressure

Various formulae for the line force on a pile of known diameter d vs. ultimate strength c_u of an undrained cohesive soil flowing past the piles have been compiled in Fig. 10.

The first one is due to Brinch Hansen and Lundgren [16] developed 1958 from the soil failure formula known to all check engineers. The shear angle–dependent term for depth and width is omitted because of the analysis $\rho = 0$, and the term for cohesion contains the depth factor $d_t = 1.5$ developed by Skempton [17] as a maximum for large

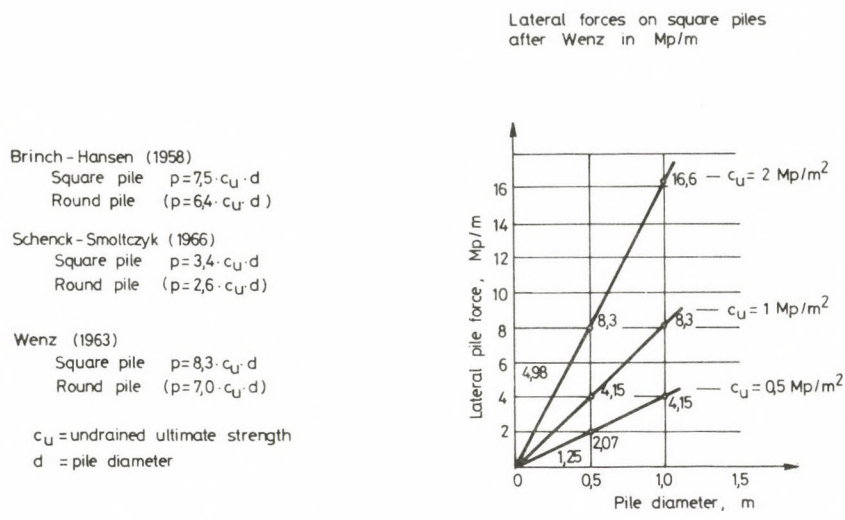


Fig. 10. Recent design formulae for lateral pressure on piles [3]

building–in depths. The formula is valid as a scientifically excellent approximate formula but it is not perfectly true for the flow past limited members.

Formulae developed by Wenz [8] in 1963 are attractive by relying on theoretically, physically and mathematically unobjectionable flow formulae by Prandtl [7], besides of being confirmed by detailed model tests.

Formulae by Schenk and Smoltczyk (1966) [18] resulted from the vectorial addition of pile cross sections flown past at boundary lines incident and leaving at 45° and deliver much lower values than both formulae above. They represent the possible lowest lateral friction resistances, neglecting the effect of constriction.

Practical values obtained from the quoted formula by Wenz for square piles have been plotted in Fig. 10. Accordingly, line loads on 1 m of pile are, for a relatively low, undrained shear strength of only 5 kN/m^2 and for 30 cm width, 12.5 Kn/m, for 50 cm width 21 kN/m, and for 1 m width 42 kN/m; for $c_u = 20$ kN/m^2, values grow already to 50 to 166 kN/m.

For shear values this high, the critical gliding planes mostly exhibit already sufficient soil failure safety factors $\eta \geqq 1.5$ — apart from extra high loads in bulk storages and in storage areas of steelworks. Such high-shear-strength soil types also exhibit much lower deformabilities in shear, making the design of piles for lateral pressure a priori needless.

Schemes of calculation underlying those by Wenz [8] are seen in Fig. 11, with stress fields in the bottom.

The mentioned classic Prandtl formula for the semi–space

$$p = (2 + \pi) c_u d = 5.14 c_u d$$

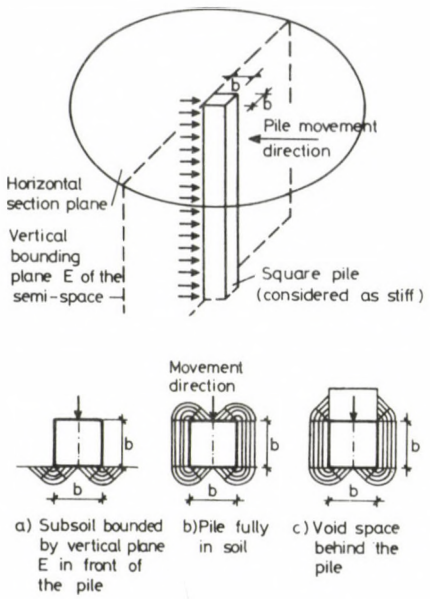


Fig. 11. Cross section of pile and soil in semi-space and full space after Wenz [8]

refers to the diagram in bottom left. The left–side drawing refers to the full space with a void behind the pile, expressed by:

$$p=(2+2\pi)c_u d=8.28c_u d$$

corresponding to the value in Fig. 10. The closed full space is presented in the middle and formulated:

$$p=(2+3\pi)c_u d=11.42c_u d.$$

Since values measured by Wenz in his tests fairly agreed with this latter formula, and the closed full space seemed to be most plausible for constant–volume flow, Wenz argued for its application in design problems.

Members — among them the Author — of Working Committee 5 for elaborating the quoted recommendation [13] decided — somewhat Salomonic, somewhat in the manner of the Papal Court, and also from the aspects of simplification and economy — to base the design of piles for flow pressure on:

$$p=10c_u d \quad [\mathrm{kN/m}]$$

becoming a routine in the years after 1978 in the FRG.

8. Sophisticated flow pressure formulae reckoning with flow rate

Wenz's study [8] was published in the Leussink era at the Karlsruhe University, and since then, his successor Gudehus dynamically furthered research in this scope.

Leinenkugel's thesis [19] published in 1976 aimed at examining the velocity dependence of the resistance to deformation. A result of that will be presented in Fig. 12.

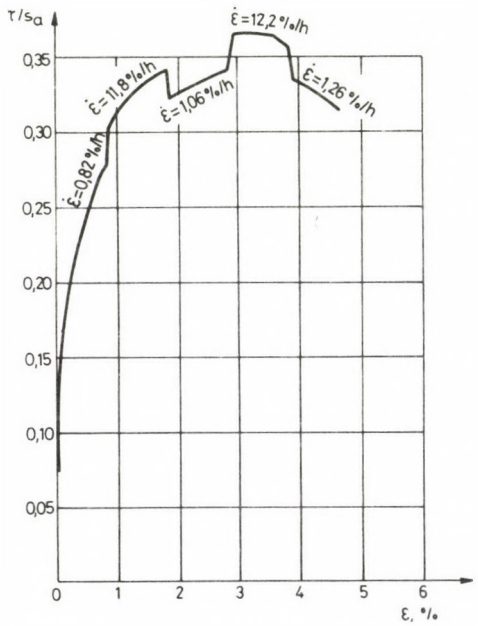


Fig. 12. Path-controlled test results with different, section-wise constant deformation rates after Leinenkugel [19]

Deformation, strain has been plotted in abscissa, and shear stress, resistance to deformation vs. equivalent stress in ordinate. The test started with a strain or deformation rate of 1% per hour, got abruptly increased to nearly 12% per hour, to drop again to 1%/h, etc.

Almost directly with the acceleration, resistance to deformation hence shear resistance also abruptly increases.

To conceive it as a new law in soil mechanics, the share of viscosity in the resistance may be distinguished by the regular variation of shear resistance, with shear resistance with feed velocity. The shear rate seems linearly to increase with the logarithm of shear strain rate $\dot{\varepsilon}$, nearly independent of the strain increase.

Also variations of pressure resp. stress produce little variation in the shear resistance at constant shear velocity, and Gudehus [20] deduces from the creep law by

Leinenkugel [19] that for a 1/10 deceleration of creep, stability increases by 10%; decrease of $\dot{\varepsilon}$ to 1/100 increases the stability by 10%.

Again, this model test result corresponds to the logarithmic law of viscosity developed by Prandtl as early as in 1928, and, according to Gudehus and Leinenkugel [22], it can be written in simplified form as:

$$\Delta c_u = I_v c_u \ln(\dot{\varepsilon}/\dot{\varepsilon}_0) \quad [\mathrm{kN/m^2}]$$

where:

I_v — non–dimensional proportionality factor (as a function of the moisture content at the liquid limit of the tested soil);
Δc_u — (positive or — mostly — negative) increase of the undrained shear strength;
c_u — undrained shear strength from laboratory test;
$\dot{\varepsilon}_0$ — reference velocity in laboratory test (mostly 1%/h in the reference test)
values to be obtained in simple, possibly either path-controlled or load–controlled triaxial tests.

[22] presents practical values for three soil types, viz. a lacustrine clay of Lake Constance, a kaolin, and a North–German clay.

I_v values, practically linearly dependent on moisture content, are 2.6 to 4.4% (clay) at the liquid limit w_L.

Accordingly, the c_u value determined in the laboratory reference test at a rate $\dot{\varepsilon}_0 = 1\%/\mathrm{min}$ decreases for 1%/day by 8 to 14%, and for 1%/year by 26 to 36%.

A further thesis submitted in Karlsruhe by Winter [23], examining a flow law taking variation of undrained shear strength due to flow rate differences into consideration, presented newly developed FE-methods for numerical solutions for steady flow motions imposed by significant non-linearities, and demonstrated the involved solutions to be mathematically correct approximations.

Gudehus and Leinenkugel [22] applied this recent velocity–dependent undrained shear strength to calculate the flow pressure in a rather common example, of a natural and a built slope, flowing — as possible — 1 cm a month, to be:

$$p = 4.5 c_u d$$

obtaining less than half of the assumed value suggested by the Committee for designing the pile for the lateral pressure:

$$p = 10 c_u d$$

The author deems this new method for determining the flow pressure on piles to be unobjectionable, both theoretically and physically confirmed, a fair achievement of theoretical and practical research at Karlsruhe.

9. Laboratory and field flow velocities

Anyhow, estimation of flow rates is felt actually to be more difficult, and little safe knowledge is available in this scope.

Practical measurements have shown several abrupt rate changes in slopes; partly attributable to e.g. precipitations, fluctuations of water level, thixotropy phenomena and the like, and partly inexplicable.

Shear box instruments, generalized earlier, had been usually load controled. Krey's shear box of $6 \times 6 = 36\ cm^2$ surface was regularly load controlled at 1/36 kg/min, equivalent to a stress increase of 1 kg/cm^2 in half an hour.

It should be noticed that, in the FRG, change to path control in shear tests, often involved introduction of $1/36 = 0.03$ cm/min — so to say as a comparison or calibration velocity — corresponding to about:

hourly 1.8 cm,
daily 44 cm,
yearly 160 m,

which, referred to the 6 cm specimen edge, would yield an approximately 0.5%/min or 30%/h deformation or shear rate.

Recent path–controlled shear box instruments operate sometimes in velocity ranges of 1 : 1000, and normally, of 1 : 100, smoothly adjustable from 0.0001 cm/min to 0.01 cm/min (or 0.1 cm/min), corresponding to

hourly $0.006 \div 0.6$ (6) cm,
daily $0.14 \div 14$ (140) cm,
yearly $0.5 \div 50$ (500) m.

Again, referring the specimen edge length to rate, strain or deformation rates of 0.1 to 10%/h result, as a rough approximation.

Leinenkugel [19] a priori refers rate to the deformation or strain of the specimen, usually applying 1%/h as reference value. Referred to the specimen length, an equivalent control feed rate of approximately $v = 0.1$ cm/h is again obtained.

His Farnell-type biaxial instrument permits constant feed rates of 0.0001 cm/min to 0.4 cm/min, corresponding to:

hourly $0.006 \div 24$ cm,
daily $0.14 \div 576$ cm,
yearly $0.5 \div 2100$ m

practically similar to those of commercial triaxial instruments.

The bigger model tester of Wenz [8] worked of course much faster, in the range from 0.2 to 4.5 cm/min. Converted values are:

hourly $12 \div 270$ cm,
daily $2.9 \div 65$ m,
yearly $1 \div 23$ km.

On the other hand, naturally occurring, protracted flow and creep phenomena are known with rates, irregarded upper extremal values and values below the measurement range, ranging from about 0.5 to 30 cm/year, that is, 0.005 to 0.3 m/year. Comparison of laboratory and field rates shows the latter to be slower by 2 to 4 decimal powers, even 100 000 times lower than those in model tests by Wenz.

2 cm/day horizontal velocities in slopes under constant load are already catastrophal, entraining immediate building legislational evacuation measures; practically they are considered as shock–like slope and soil failures.

Wenz reckoned with max. soil flow rates of 1 cm/min $\sim$ 15 m/day, knowing that practically occurring flow displacements were surely slower than that (see p. 56 in [8]).

In conformity with the engineering knowledge of his time, he considered the c_u value to be independent of rate, and only "dynamic viscosity" — he himself demonstrated to be negligible — to be rate-dependent.

On the other hand, it is an important and convincing argument in favour of works by Leinenkugel [19] and Winter [23] that their method, applying model rates and soil characteristics according to Wenz [8], yielded flow pressures indicated by the latter, according to Prandtl's formula

$$p=(2+2 \text{ or } 3\pi)c_u d.$$

Accordingly, in the version [24] revised by the Road Research Institute, Köln, of the chapter "Lateral Pressure on Piles" by the Working Committee "Effect of Backfill on Constructions" in the Directives first issued in 1977, to be soon published in "Geotechnik", the flow pressure formula for single piles will be reduced to:

$$p_f=7c_u d \quad [\text{kN/m}].$$

No doubt, this is a conservative (the Author being member of the Committee) but economical treatment of the problem.

This cautious application of recent theories developed under the guidance of Gudehus is due, as mentioned, to the insufficient knowledge of naturally occurring flow rates and of their often unpredictable changes.

Irrespective of that, such important extrapolations of model tests have always to be considered as critical, until these low flow pressure values get confirmed in large-scale or in–situ tests.

Besides, in earthworks relatively high rates may practically occur or can be produced. In tests by Heyman and Boersma [2], the five fillings followed each other two–weekly, rather slowly for an actual construction, nevertheless exhibiting rates of 0.4 cm/day resp. or 1.57 m/year, much exceeding e.g. the assumed 12 cm/year flow in a slope described in Chapter 9 of [22] by Gudehus and Leinenkugel.

10. Pile design for resultant soil pressure

Lateral forces in piles cannot exceed the resultant soil pressure, i.e. difference between the soil pressure on the load side, and an equivalent soil resistance on the off-load side, of piles or piling (see 7.8 and 7.9 in [3], and 2.2 and 4.4.2 in [13]).

Soil pressure Δp due to soft layers just loaded, applying the undrained shear strength (i.e., $\rho=0$, hence $\lambda_a=1$) becomes:

$$e_a=\gamma z_a+\Delta p-2c_u \quad [\mathrm{kN/m^2}]$$

and needs but slight deformation paths to be active. Equivalent soil resistance is that activated as partial resistance for the same deformation path, taken equal to natural soil pressure (again $\lambda_p=1$):

$$\mathrm{cal}\, e_p \sim \gamma z_p \quad [\mathrm{kN/m^2}]$$

and the resultant soil pressure:

$$\Delta e=e_a-\mathrm{cal}\, e_p \quad [\mathrm{kN/m^2}].$$

In evenly spaced a pile rows normal to the force direction, a single pile is loaded by:

$$\bar{E}_h=a\Delta e \quad [\mathrm{kN/m}]$$

that is, the difference of soil forces acting on the total width of the pile row has to be divided by the number of piles.

Design value for single piles exposed to lateral pressure is the lowest value resulting from the soil pressure difference or the flow formula. In assuming the flow pressure on a single pile, of course, its increase for pile rows as a function of built-in depth according to Wenz [8] has to be considered. Anyhow, this is only significant if piles are closer than $3a$ as seen in Fig. 13.

For further details see Recommendations [13] or the recent version of Directives [24].

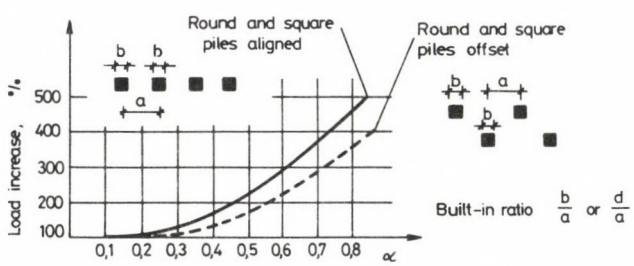


Fig. 13. Increase of flow pressure on single piles due to group effect as a function of building-in ratio, after Wenz [8]

Let us still mention that, according to outworn views of some pile experts, no significant lateral pressure acts on the piles if during the consolidation process of each filling, only deformations $\leqq 3$ cm arise. But then, the soil pressure difference will be absorbed by single piles if they are dense enough. For spacings over $3d$ (d = pile thickness), Recommendations [13], [24] limit the resultant soil pressure maximum to $3d\Delta e$.

Winter deems full development of the flow pressure to be only possible — also demonstrated by model tests — for a relative displacement of at least $0.1d$ between pile and soil, hence e.g. 3 cm for a pile width of 30 cm, the view conventional for this pile type. For greater pile diameters, no such a linear relationship has been proved to now.

The effect of loads not adjacent to, but at some distance from, the piles depends — as mentioned — on the deformation arising there in the soil. A design suggestion made by Horch [26] got adopted in the recent version of Directives [24].

Accordingly, for piles spaced at a distance equal to the soft layer thickness, about 20% of the design lateral pressure, and at twice this spacing about 10% has to be taken into account.

11. Effect on pile groups

Research on the distribution of such lateral forces in pile groups is going on, without — as far as I know — satisfying results to now. Neither did works by researchers from India (Prakash [27], [28]), Australia (Poulos [29] to [34]), and the USA (Vesić [35], [36]) or general reports by van de Beer [37] and Broms [38] produce agreeing conceptions.

For pile grids such as those under bridge piers, Horch [26] conceived the distribution of the resultant soil force over the overall grid width according to Fig. 14 a suggestion adopted by the Road Research Institute for the recent version of Directives [24], to be generally introduced by the Road Authority.

Essentially it assigns the main part of the lateral force always to the first pile row, rather than to evenly distribute between piles the lateral force resulting from the soil pressure difference. If the design is controlled by flow pressure, its full value has to be assumed for each pile, except if piles are so dense in force direction that — as stated before — a flow pressure increased as a function of arrangement and spacing of the piles according to Wenz [8] becomes prevalent.

12. Conclusion

These considerations have been intended to simply and perspicuously present the latest development, and to point out that lateral forces due to unilateral loads on piles in soft soil types — where these are normally used — should not be ignored any more.

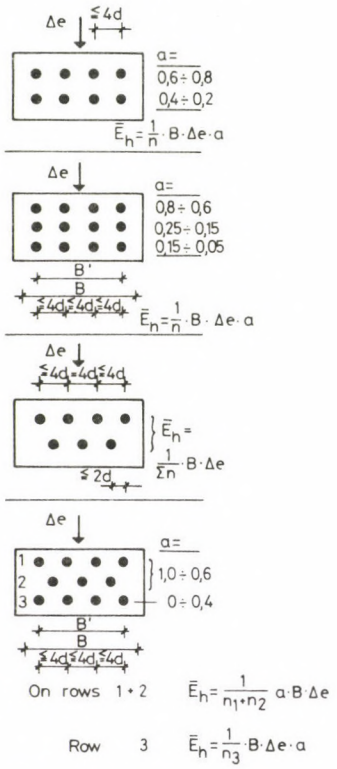


Fig. 14. Distribution of the resultant soil force before pile grids between single piles after Horch [26]

Undeniably, the problematic of piling behaviour has been treated rather briefly. It seems to be rather complex in research, imposing to be individually treated from economy aspects.

If flow velocity and drained shear strength values are available, the Author would suggest to consequently apply the method by Gudehus, Leinenkugel and Winter [22] from economy and precision aspects.

References

1. Bjerrum, L.: Engineering geology of Norwegian normally-consolidated marine clays as related to settlements of buildings. Geotechnique, 17 (1967), 83—118
2. Heyman, Boersma: Bending moments in piles due to lateral earth pressure. Proc. Int. Conf. Soil Mech. Found. Eng. 2 (1961), 425
3. Steinfeld, K.: Biegebeanspruchung von Pfählen unter Brücken durch seitliche Erdkräfte. Schriftenreihe Forschungsarbeiten aus dem Straßenwesen, Heft 74, "Straße und Untergrund IV", Kirschbaum-Verlag, Bonn-Bad Godesberg 1969
4. Leussink, Wenz: Über das Scherfestigkeitsverhalten von bindigen Erdstoffen im Bereich der deutschen Nordseeküste bei Belastung mit hohen Flächenlasten. Vorträge der Baugrundtagung 1966 in München, Deutsche Gesellschaft für Erd- und Grundbau e. V., Eigenverlag, Essen 1966

5. Franke, Schuppener: Horizontalbelastung von Pfählen infolge seitlicher Erdauflasten. Geotechnik, 4 (1982), 189—197
6. Nahrgang, E.: Verformungsverhalten eines weichen bindigen Untergrundes. Veröffentlichungen des Inst. für Bodenmechanik u. Felsmechanik der Universität Fridericiana in Karlsruhe, Eigenverlag, Heft 60, Karlsruhe 1965
7. Prandtl, L.: Über die Härte plastischer Körper. Nachrichten der Kgl. Gesellschaft der Wissenschaften, Göttingen, Eigenverlag, Februar 1920
8. Wenz, K. P.: Über die Größe des Seitendruckes auf Pfähle in bindigen Erdstoffen. Veröffentlichungen des Inst. für Bodenmechanik und Grundbau der Technischen Hochschule Fridericiana in Karlsruhe, Heft 12, Eigenverlag, Karlsruhe 1963
9. Jäger, E.: Versuch über die Verminderung des aktiven Erddruckes durch Pfahlreihen. Mitteilungen der Hannover Hochschulgemeinschaft, Heft 11, Eigenverlag 1929
10. Förster, K.: Abschirmung des Erddruckes vor Spundwänden durch Pfahlreihen. Mitteilungen der Hannover Hochschulgemeinschaft, Heft 17/18, Eigenverlag 1937
11. Peck, Raamot: Foundation behaviour of iron storage yards. Journal of the Soil Mechanics and Foundation Division ASCE, 90 (May 1964), No. SM 3, Part 1, 85
12. Neumeier, H.: Irrtümer beim Entwurf von Pfahlgründungen. Bauingenieur 38 (1963), 6, 241 (nach H. Lossier "Quelques cas d'erreurs de conception de fondation sur pieux")
13. Fedders, H.: Seitendruck auf Pfähle durch Bewegungen von weichen bindigen Böden — Empfehlung für Entwurf und Bemessung. Geotechnik 1 (1978), S. 100—104, Organ der Deutschen Gesellschaft für Erd- und Grundbau e. V.
14. Wenz, K. P.: Seitendruck auf Pfähle in weichen bindigen Erdstoffen. Vorträge der Baugrundtagung 1972 in Stuttgart, Deutsche Gesellschaft für Erd- und Grundbau e. V., Eigenverlag, Essen 1972
15. Steinfeld, K.: Diskussionsbeitrag zum Vortrag Leussink Wenz (4). Vorträge der Baugrundtagung 1966 in München, Deutsche Gesellschaft für Erd- und Grundbau e. V., Eigenverlag, Essen 1966
16. Brinch Hansen, Lundgren: Hauptproblem der Bodenmechanik. Springer-Verlag, Berlin 1960, 266
17. Skempton, A. W.: The bearing capacity of clays, Proc. Building Research Congress, London 1951
18. Schenck, Smoltczyk: Pfahlroste, Berechnung und Ausbildung. Grundbau-Taschenbuch, 1, 2. Aufl., 685, Verlag Wilhelm Ernst u. Sohn, Berlin 1966
19. Leinenkugel, H. J.: Deformations- und Festigkeitsverhalten bindiger Erdstoffe. Experimentelle Ergebnisse und ihre physikalische Deutung. Veröffentlichungen des Instituts für Bodenmechanik und Felsmechanik der Universität Fridericiana in Karlsruhe, Heft 66, Eigenverlag, Karlsruhe 1966
20. Gudehus, G.: Bericht über die Spezialsitzung der Baugrundtagung 1974. Monotone zeitabhängige Vorgänge im Baugrund. Vorträge der Baugrundtagung 1974 in Frankfurt/Main—Hoechst, Deutsche Gesellschaft für Erd- und Grundbau e. V., Eigenverlag, Essen 1974, 249—268
21. Prandtl, L.: Ein Gedankenmodell zur kinematischen Theorie der festen Körper. Zeitschrift für Angewandte Mathematik und Mechanik, 8 (1928), 85—105
22. Gudehus, Leinenkugel: Fließdruck und Fließbewegung in bindigen Böden: Neue Methoden. Vorträge der Baugrundtagung 1978 in Berlin, Deutsche Gesellschaft für Erd- und Grundbau e. V., Eigenverlag, Essen 1978, 411—429
23. Winter, H.: Fließen von Tonböden: Eine mathematische Theorie und ihre Anwendung auf den Fließwiderstand von Pfählen. Veröffentlichungen des Instituts für Bodenmechanik und Felsmechanik der Universität Fridericiana in Karlsruhe, Heft 82, Eigenverlag, Karlsruhe 1979
24. Schmiedel, U.: Seitendruck auf Pfähle Abschnitt zum "Merkblatt für die Hinterfüllung von Bauwerken" der Forschungsgesellschaft für Straßenwesen, Köln. Vorveröffentlichung geplant für die Zeitschriften "Bauingenieur" Springer-Verlag, Berlin und "Geotechnik" der Deutschen Gesellschaft für Erd- und Grundbau e. V., Eigenverlag, Essen, 1983
25. Winter, H.: Bemessung von Pfahlgründungen und Hangverdübelungen auf Fließdruck. Vorträge der Baugrundtagung 1980 in Mainz, Deutsche Gesellschaft für Erd- und Grundbau e. V., Eigenverlag, Essen 1980, 539
26. Horch, M.: Zuschrift zu: Seitendruck auf Pfähle . . . (13), Geotechnik, Jahrgang 3 (1980), Heft 4, S. 207—208, Organ der Deutschen Gesellschaft für Erd- und Grundbau e. V., Essen
27. Prakash, S.: Behaviour of pile groups subjected to lateral loads. Ph. D. thesis, University of Illinois, 1962
28. Prakash, Balasubramaniam: Behaviour of battered piles under lateral load. Indian National Society Soil Mech. Found. Eng., 4 (1965), 177—196 (siehe auch (27))

29. Poulos, H. G.: Displacements in a soil mass due to piles and pile groups. Australian Geomechanics Journal, 1 (1971) 29—35
30. Poulos, H. G.: Behaviour of laterally loaded piles: I. single piles. J. Soil Mech. Found. Eng. Div., Proc. ASCE 97 (1971) Nr. SM 5, 711—731
31. Poulos, H. G.: Behaviour of laterally loaded piles: II. pile groups. J. Soil Mech. Found. Eng. Div., Proc. ASCE 97 (1971) No. SM 5, 733—751
32. Poulos, Madhav: Analysis of the movement of battered piles. Proc. 1. Australia-N. Z. Conf. Geomech., Melbourne, 1 (1971), 268—275
33. Poulos, H. G.: Analysis of piles in soils undergoing lateral movement. J. Soil Mech. Found. Div., Proc. ASCE 99 (1973) Nr. SM 5, 391—406
34. Poulos, H. G.: Some recent developments in the theoretical analysis of pile behaviour. Chapter 7 N, Soil Mechanics New Horizons, London—News, Butterworths, 1974, 237—279
35. Vesić, A. S.: Tests on instrumented piles, Ogeechee River site. Proc. ASCE, J. Soil Mech. Found. Div. 96 (1970) Nr. SM 2, 561—584
36. Vesić, A. S.: Principals of pile foundation design. School of Engineering, Duke University, Soil Mechanics Series No. 38 (Boston Society of Civ. Eng. ASCE, Lecture Series on Deep Foundations, March—April 19755)
37. De Beer, E.: The effects of horizontal Loads on piles due to surcharge or seismic effects. Proc. 9th Int. Conf. Soil Mech. Found. Eng., Tokio 3 (1977) Special Session 10, 547—558
38. Broms, B. B.: Pile foundations — General Report. Proc. 10th Conf. Soil Mech. Found. Eng., Stockholm 4 (1981) Session 8, 427—439, Verlag A. A. Balkema, Rotterdam 1982

NOTICE TO CONTRIBUTORS

Papers in English* are accepted to the condition that they have not been previously published or accepted for publication.

Manuscripts in two copies (the original type-written copy plus a clear duplicate one) complete with figures, tables, and references should be sent to the

Acta Technica
Münnich F. u. 7. I. 111A
Budapest, Hungary
H-1051

Although every effort will be made to guard against loss, it is advised that authors retain copies of all material which they submit. The editorial board reserves the right to make editorial changes.

Manuscripts should be typed double-spaced on one side of good quality paper with proper margins and bear the title of the paper and the name(s) of the author(s). The full postal address(es) of the author(s) should be given in a footnote on the first page. An abstract of 50 to 100 words should precede the text of the paper. The paper should not exceed 25 pages including tables and references. The approximate locations of the tables and figures should be indicated on the margin. An additional copy of the abstract is needed. Russian words and names should be transliterated into English.

References. Only papers closely related to the author's work should be referred to. The citations should include the name of the author and/or the reference number in brackets. A list of numbered references should follow the end of the manuscript.

References to periodicals should mention: (1) name(s) and initials of the author(s); (2) title of the paper; (3) name of the periodical; (4) volume; (5) year of publication in parentheses; (6) number of the first page. Thus: 5. Winokur, A., Gluck, J.: Ultimate strength analysis of coupled shear walls. American Concrete Institute Journal 65 (1968), 1029.

References to books should include: (1) author(s) name; (2) title; (3) publisher; (4) place and year of publication. Thus: Timoshenko, S., Gere, J.: Theory of Elastic Stability. McGraw-Hill Company, New York, London 1961.

Illustrations should be selected carefully and only up to the necessary quantity. Black-and-white photographs should be in the form of glossy prints. The author's name and the title of the paper together with the serial number of the figure should be written on the back of each print. Legends should be brief and attached on a separate sheet. Tables, each bearing a title, should be self-explanatory and numbered consecutively.

Authors will receive proofs must be sent back by return mail.

Authors are entitled to 50 reprints free of charge.

* Hungarian authors should submit their papers also in Hungarian.

Periodicals of the Hungarian Academy of Sciences are obtainable at the following addresses:

AUSTRALIA
C.B.D. LIBRARY AND SUBSCRIPTION SERVICE
Box 4886, G.P.O., *Sydney N.S.W. 2001*
COSMOS BOOKSHOP, 145 Ackland Street
St. Kilda (Melbourne), Victiria 3182

AUSTRIA
GLOBUS,Höchstädtplatz 3, *1206 Wien XX*

BELGIUM
OFFICE INTERNATIONAL DE LIBRAIRIE
30 A venue Marnix, *1050 Bruxelles*
LIBRAIRIE DU MONDE ENTIER
162 rue du Mindi, *1000 Bruxelles*

BULGARIA
HEMUS, Bulvar Ruszki 6, *Sofia*

CANADA
PANNONIA BOOKS, P.O. Box 1017
Postal Station "**B**", *Toronto, Ontario M5T 2T8*

CHINA
CNPICOR, Periodical Department, P.O. Box 50
Peking

CZECHOSLOVAKIA
MAD'ARSKÁ KULTURA, Národní třida 22
115. 66 Praha
PNS DOVOZ TISKU, Vinohradská 46, *Praha 2*
PNS DOVOZ TLAČE, *Bratislava 2*

DENMARK
EJNAR MUNKSGAARD, Norregade 6
1165 Copenhagen K

FEDERAL REPUBLIC OF GERMANY
KUNST UND WISSEN ERICH BIEBER
Postfach 46, *7000 Stuttgart 1*

FINLAND
AKATEEMINEN KIRJAKAUPPA, P.O. Box 128 *SF-00101 Helsinki 10*

FRANCE
DAWSON-FRANCE S. A., P. 40, *91121 Palaiseau*
EUROPÉRIODIQUES S. A., 31 Avenue de Versailles, *78170 La Celle St. Cloud*
OFFICE INTERNATIONAL DOCUMENTION ET
LIBRAIRIE,48 rue Gay-Lussac
75240 Paris Cedex 05

GERMAN DEMOCRATIC REPUBLIC
HAUS DER UNGARISCHEN KULTUR
Karl Liebknecht-Straße 9, *DDR-102 Berlin*
DEUTSCHE POST ZEITUNGSVERTRIEBSAMT Straße der Pariser Kommüne 3 4, *DDR-104 Berlin*

GREAT BRITAIN
BLACKWELL'S PERIODICALS DIVISION
Hythe Bridge Street, Oxford OX1 2ET
BUMPUS, HALDANE AND MAXWELL LTD.
Cowper Works, *olney, Bucks MK46 4BN*
COLLET'S HOLDINGS LTD., Denington Estate *Wellingborough, Northants NN8 2QT*
WM. DAWSON AND SONS LTD., Cannon House *Folkstote, Kent CT19 5EE*
H. K. LEWIS AND CO., 136 Gower Street
London WCIE 6BS

GREECE
KOSTARAKIS BROTHERS INTERNATIONAL
BOOKSELLERS, 2 Hippokratous Street, *Athens-143*

HOLLAND
MEULENHOFF-BRUNA B. V., Beulingstraat 2,
Amsterdam
MARTINUS NIJHOFF B.V.
Lange Voorhout 9 11, *Den Haag*
SWETS SUBSCRIPTION SERVICE
347b Heereweg, *Lisse*

INDIA
ALLIED PUBLISHING PRIVATE LTD., 13/14
Asaf Ali Road,*New Delhi 110001*
150 B-6 Monunt Road, *Madras 600002*
INTERNATIONAL BOOK HOUSE PVT. LTD.
Madame Cama Road, *Bombay 400039*
THE STATE TRADING CORPORATION OF INDIA LTD.,
Books Import Division, Chanralok 36 Janpath, *New Delhi 110001*

ITALY
INTERSCIENTIA, Via Mazzé 28, *10149 Torino*
LIBRERIA COMMISSIONARIA SANSONI, Via Lamarmora 45, *50121 Firenze*
SANTO VANASIA, Via M. Macchi 58
20124 Milano
D. E. A., Via Lima 28, *00198 Roma*

JAPAN
KINOKUNIYA BOOK-STORE CO. LTD.
17-7 Shinjuku 3 chome, Shinjuku-ku, *Tokyo 106-91*
MARUZEN COMPANY LTD., Book Department, P.O. Box 5050 Tokyo International, *Tokyo 100-31*
NAKUA LTD. IMPORT DEPARTMENT
2-30-19 Minami Ikebukuro, Toshima-ku, *Tokyo 171*

KOREA
CHULPANMUL, *Phenjan*

NORWAY
TANUM-TIDSKRIFT-SENTRALEN A.S., Karl Johansgatan *41 43, 1000 Oslo*

POLAND
WEGIERSKI INSTYTUT KULTURY, Marszalkowska 80, *00-517 warsawa*
CKP-I W, ul. Towarowa 28,*00-958 Warsawa*

ROUMANIA
D.E.P., *Bucuresti*
ILEXIM,Calea Grivitei 64-66, *Bucuresti*

SOVIET UNION
SOJUZPECHAT IMPORT, *Moscow*
and the post offices each town
MEZHDUNARODNAYA KNIGA, *Moscow G-200*

SPAIN
DIAZ DE SANTOS, Lagasca 95, *Madrid 6*

SWEDEN
GUMPERTS UNIVERSITETSBOKHANDEL AB
Box 346, *40125 Göteborg 1*

SWITTZERLAND
KARGER LIBRI AG, Petersgraben 31, *4011 Basel*

USA
EBSCO SUBSCRIPTION SERVICES
P.O. Box 1943, *Birmingham, Alabama 35201*
F.W. FAXON COMPANY, INC.
15 Soutwest Park, *Westwood Mass. 02090*
READ-MORE PUBLICATIONS, INC.
140 Cedar Street, *New York,N.Y. 10006*

YUGOSLAVIA
JUGOSLOVENSKA KNJIGA,Terazije 27, *Beograd*
FORUM,Vojvode Mišića 1, *21000 Novi Sad*

HU ISSN 0001-7035 Index: 26.025

ACTA TECHNICA

ACADEMIAE SCIENTIARUM HUNGARICAE

EDITOR: M. MAJOR

VOLUME 98

NUMBERS 3—4

AKADÉMIAI KIADÓ, BUDAPEST 1985

ACTA TECHNICA

A JOURNAL OF THE HUNGARIAN ACADEMY OF SCIENCES

Acta Technica publishes original papers, preliminary reports and reviews in English, which contribute to the advancement of engineering sciences.
Acta Technica is published by

AKADÉMIAI KIADÓ
Publishing House of the Hungarian Academy of Sciences
H-1450 Budapest, Alkotmány u. 21.

Subscription information

Orders should be addressed to

KULTURA Foreign Trading Company
H-1389 Budapest P.O. Box 149

or to its representatives abroad

Acta Technica is indexed in Current Contents

CONTENTS

BOOK REVIEWS

PRINTED IN HUNGARY
Akadémiai Kiadó és Nyomda, Budapest

Acta Technica Academiae Scientiarum Hungaricae, 98 (3—4), pp. 137—154 (1985)

SIMILARITIES AND DIFFERENCES OF WATER MANAGEMENT IN IRAQ AND IN HUNGARY*

J. BALOGH**

[Received: 14 February 1984]

Having compared the most important features of Iraqi and Hungarian water management, their past, present and future concepts, the author is deeply convinced that the drainage requirement of the Mesopotamian valley may not be expressed by hundreds of m^3/s, but by thousands of m^3/s. A rough experiment has been made to express by numbers the approach as to how the magnitude of the drainage requirement of the entire Mesopotamian valley could be estimated. Closer cooperation between Iraq and Hungary would be advisable in this field.

1. Introduction

Iraq and Hungary are two countries significantly different in their natural and social conditions. Still there are a number of features in their water household that are similar. The immense quantity of experiences and knowledge accumulated in Hungary as the result of the early development of its water management together with the recent achievements of the Hungarian agriculture could be utilized for Iraq's Water Management and Agriculture.

Attempts are made by the author to enlist the most important similar and different features of Iraq and Hungary.

Climate
Agriculture
Rivers
Drainage and Irrigation as well as the
Concepts of further development.

Finally a rough approach has been prepared estimating drainage requirements of the Mesopotamian valley.

* The quoted data included in this study are taken over from the referred professional literature.

** J. Balogh, H-1213 Budapest, Damjanich J. u. 159/b.

Acta Technica Academiae Scientiarum Hungaricae 98, 1985
Akadémiai Kiadó, Budapest

2. Iraq

Iraq, this young republic, was one of the centres where civilization and culture began. At the same time its future is also rich in good hopes although its climate is extremely arid in the overwhelming part of its territory and large areas are absolute deserts Iraq's water resources, however, are abundant.

A great part of Iraq's arid territories is on the flood plain of the two big rivers, the Tigris and Euphrates, crossing the country.

Considering that the main tasks of water management and land development affect only this—so called Mesopotamian Valley (Figure 1)—will be dealt with in the following.

Fig. 1

Climate

The climate of Iraq is subtropical, continental and arid. Long hot and dry summers and cooler winters are characteristic. Springs and autumns are short or do not occur at all. Satisfactory precipitation for rainfed agriculture is only in the northern part of the country. The humidity of the air is generally low, especially in the central and southern part of the country.

Several characteristic data are included in Table 1.

Table 1

Region of	Average annual precipitation mm	Average yearly mean temperature °C	Average summer air humidity %
Mossul	382	19.4	22—24
Baghdad	134	22.7	12—15
Rutba	110	18.8	10—13
Basra	117	23.8	25—30

The temperature [1, 2] has a high average, long and lasting maximums in the summer, meanwhile in winter minimums sometimes reaching even the domain under the freezing point.

Yearly evaporation is around 2000 mm and more than 60% of it takes place in the months of June—September. In summer the average daily evaporation is about 10 mm but even 20 mm can frequently be reached. Especially on windy days.

Corresponding to these values the evapotranspiration of crops, supplied well with water (irrigated) may reach even 15—16 mm/day frequently in summer.

Agriculture

The agricultural production of Iraq—husbandry included—is very low. Intensive large scale farming is very rare and virtually unknown. Productive lands are in decrease and so is the agricultural population, although the Government made and makes serious efforts in promoting the organization of farmer cooperatives and state farms as well as in helping the productive work of individual farmers.

In spite of these efforts the yields of crop production are very low. The average yields are e.g.:

— barley around 0.8 t/ha
— wheat around 0.6 t/ha
— rice around 1.2 t/ha.

Although practically all vegetables and fruits can be grown in Iraq only a few and of bad quality may be found on the markets. Perhaps date palms are the only exceptions that can survive and develop under the hard climate and among the soil conditions of middle and southern Mesopotamia. The yield of these palms is quite the only considerable agricultural product that can be exported in great quantities. Practically all other foodstuff is to be imported.

The livestock in the agriculture is few and of low productive capacity. There are about 700 thousand heads of cattle, but the weight of an average cow is not more than 150–200 kg.

The hard climate and extensive conditions can be fairly well tolerated by

— arab horses (cca 150 000 heads)
— donkeys and mules (cca 450 000 heads)
— goats (cca 1,5 million heads)
— water buffaloes (cca 50 000 heads) and
— camels as well as
— poultry.

Due to the poor possibilities of plant growing and animal breeding at present the young generation has abandoned agricultural zones and has been accumulating in smaller and bigger towns in the country or in the capital. The agrarian population has decreased to its fraction, and the immigrants of the towns could make a much better living in the investments of the infrastucture, the developing industry, traffic, transportation and within it the frame of internal and external commerce.

The rivers

The two big rivers—Tigris and Eufrates—have their sources in Turkey and reach Iraq having crossed Syria. Several data can be found in Table 2.

Table 2

	Length of the rivers (km)	
	Total	Iraqi reaches
Tigris	1718	1418
Eufrates	2333	1213
Shatt el-Arab	110	110

The entire catchment area of the two rivers covers 705 500 squ. km from which 359 000 squ. km is Iraqi territory. The middle and lower reaches of the two rivers belong entirely to Iraq, meanwhile the upper one only partly. Schematic longitudinal sections of the two rivers are presented in Figure 2.

As for the transported water quantities some characteristic data are included in Table 3.

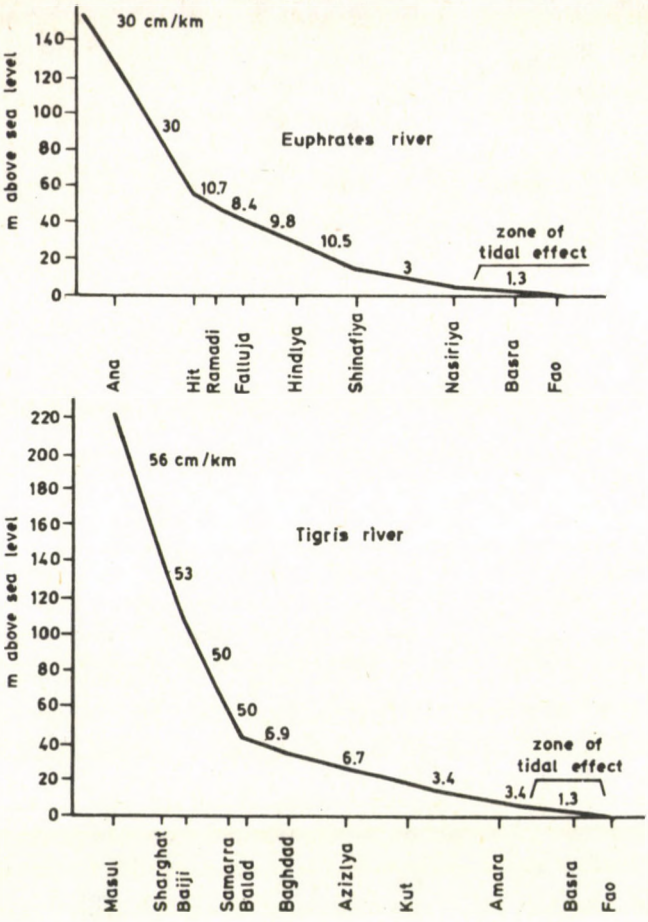


Fig. 2

Table 3

	Run offs (1902–1952)	
	Average	Low
	km^3/year	
Tigris (near Baghdad)	38.8	15.7
Euphrates (near Hit)	26.4	12.0
Total	65.2*	27.7

* According to contemporary estimation the average extent of the irrigated territory was the same as in recent years.

From the beginning of their middle reaches on, both rivers become meandering. The rivers are untrained. There is no flood protection. Therefore, floods were considerably frequent in the past. Due—however—to the big dams and reservoirs recently built in Turkey, in Syria and in Iraq itself, flood damages can almost be prevented.

Both rivers have pending riverbeds from the beginning of their middle reaches down to their confluence. A characteristic cross section of their valleys—after Buringh—can be studied on Figure 3.

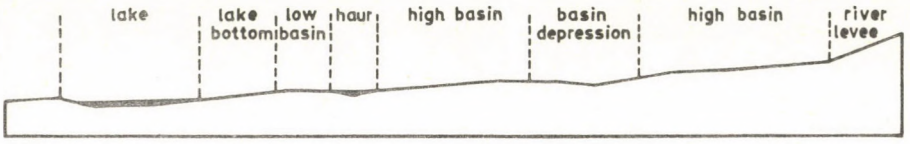


Fig. 3

The quality of their transported water is suitable for irrigation and also for drinking although it contains 200–400 g/m^3 solved salts. Harmful Na salts, however, have a share of 30–40 g/m^3 only. Silt content of the water in the two rivers is significant, i.e.: 1000–4000 g/m^3. Higher during floods and lower during low water periods.

The two rivers were connected with artificial canals already in the distant past due to the fact, that one of the rivers has higher bed than the other alternavely.

Irrigation

Irrigation—without which agricultural production here is impossible—has its roots in the prehistoric past. According to our recent knowledge, early primitive human societies began to build small earth dams in the (periodical) waterways of the northern hilly region. They observed, namely, that their animals found better grazing and higher yields of—in the beginning collected—palatable grasses (the ancestors of the barley and wheat) can be achieved if the ground had previously been soaked with water.

In this period—6–8 thousand years ago—the valleys of the twin rivers were more fertile than now. Natural floods of the rivers inundated considerable territories regularly. Abundant vegetation grew on these lands. Extent grasslands and savannah had to be found here, since in the Accadian and Babilonian epochs lions were sculptured on monuments. Lions, however, may live and breed on such territories where big flocks and herds of herbivorous animals find pastures, enough for them to live on.

The primitive societies of the hilly regions descended to the river valleys just because of their fertility. In the beginning settlements were limited to the riversides,

where natural floods fertilized the depressions. Later on the population occupied farther and farther lands in the basin and even in the region of the rivers' confluence.

A new irrigation method had been invented for those arid years, when the level of the two rivers was not enough to surpass the levees of the pending riverbeds. It consisted of simply cutting the river levees to let the water expand even at the period of low water. So fertilizing irrigation reached the lower plain lands of the basin even during droughts.

Primitive canalization systems were created, after cutting the river levees for irrigation water diversion and transport, to farther and farther deep lying fertile territories. The operation and maintenance of these constructions required discipline and division of work among specialized workers already in the early centuries. In this way entire states came into being with civilized welfare, flourishing agriculture, lively commerce, handicrafts and building industry.

Centuries and even thousands of years brought about the development of such urban states as Babylon, Eridu, Kish, Nimrud, Ninive, Ur, Uruk, etc.

The disadvantages of the primitive irrigation method already existed in these periods. One part of the irrigation water diverted to the productive lands percolated below the root zone. The level of the ground water began to rise higher and higher. The other part of the water quantities, flooding the cropped territories run off along the gullies and wadis to the deepest depressions, creating lakes, marshes and swamps. The third part of the irrigation water—that was retained in the root zone,—in the three phase layer of the soil—evaporated.

As the result of these phenomena, the good quality irrigation water, diverted from the rivers gradually deposited its

— sediments
  into the irrigation canals and
  on the surfaces of irrigated lands, while its
— salt contents remained
  in the originally sweet water lakes,
  in the fertile soils irrigated and
  in the ground water.

Due to the extreme evaporative demand of the air, salts have gradually accumulated in the groundwater, in the lakes and even in the upper layer of the soils alike. The fertility of the lands began deteriorating parallel with the accumulation of the salts. New and new distant and more distant territories were taken under irrigation and the same process has destroyed Mesopotamian agriculture almost entirely.

A strange irrigation practice has finally been developed in a kind of fallow system. Fifty percent of irrigable land is irrigated in each irrigation season. The abundant irrigation water percolates through the upper soil layer and takes soluble salts into the ground water, decreasing the harmful salt concentration in this way in the root zone. Crop production becomes possible but on a low yield level. Under the other fifty percent of the irrigable land ground water table rise and salt concentration

increases in it during the fallow period. But in the next irrigation season this part of the territory will be irrigated, and the process—described above—makes a low level crop production possible again.

From the 9 million ha of cultivable land 3–4 million ha is irrigated with the water resources of the twin rivers. Generally the irrigation method applied is flooding. Furrow irrigation is adopted only exceptionally on cotton and several vegetables. The significance of sprinkling and trickling irrigation is negligible.

Drainage

River regulating and flood control would also be included in this chapter. This activity, however, has far less traditions and roots reaching into the distant past than has irrigation in Iraq. Regulation of the meandering rivers had not begun yet. The few dams were constructed in this century in Turkey, Syria and even in Iraq solve the main problems of flood prevention.

On drainage works there are only a few references in ancient inscriptions and on monuments. It is quite understandable, because due to the relief and the pending riverbeds, there are only few or no possibilities to drain the excess irrigation or flood water back into the rivers. The natural runoff and the extreme evaporation partly solved the removal of the excess water quantities. In the deepest depressions salt water lakes were created by the accumulating surface runoff and at the outlet of the river basins along the Shatt al-Arab at the confluence of the rivers and above, marshes extension came into being due to the permanent increase of the ground water table under the irrigated lands and the progress of the river delta.

So drainage would have extraordinarily many and important tasks in Iraq. Apart from the necessity of mutual agreement between the countries of the catchment areas, on river regulating, flood protection and, last but not least, on water resources division, drainage would have the immense tasks of

i) leading excess irrigation water off the irrigated lands,
ii) decreasing too high ground water levels,
iii) draining off the leachets,
iv) conducting surface waters (excess precipitation and the water cover of marshes) away.

i) Other irrigation methods than flooding have no traditions here and have no economic bases either. The excess quantities of irrigation water distributed, run off the surface to the depressions and percolate into the groundwater instead of being drained away. After the development of power driven pumping stations lifting back excess waters was introduced but still only at a few places and with not enough discharge.

ii) The ground water level is high, in the entire Mesopotamian valley. 1—1,5 m can be considered as the average depth of the groundwater table. But this level is significantly higher than the critical one. For Central Iraq namely the critical groundwater level estimated according to Polinov (1957) is about *3,5 m under the surface* and above this salts are migrating upwards into the soil profile.

iii) Excess water quantities of flood irrigations have not only rised the groundwater table but even the salt concentration has been increasing permanently, since thousands of years. The lack of drainage, with the present irrigation practice the salts are leached down from one part of the territory by percolating irrigation water. They are accumulating under the other part and return into the surface layers after irrigation already during the next fallow season.

Although due to the fortunate soil texture salts can be removed by leaching, almost all lands of the entire Mesopotamia are saline because of lack of drainage. The leachates can not be transported away because there are no recipients and transporting systems either.

iv) Excess water quantities of the heavy showers precipitating locally during winter have not high influence on the dimensions of the prospective drainage network. The water cover on the surface of the extent marshes in the south, would have to be drained off, too.

The magnitude of the marshes is astonishing. Buringh (1960) estimates this territory at 35.000 km^2. This immense territory is covered by water permanently or periodically. The average depth of this water cover may be estimated to 1–2 m.

These are the main requirements that are to be met by drainage. Without the solution of these problems restoration of the agriculture can not be solved in the Mesopotamian valley.

Concepts

The leaders of the Republic of Iraq soon recognized that the modernization of agriculture has a decisive role in the development of the entire national economy. Serious steps have been taken to improve agrotechniques and to increase irrigation.

However, Buringh (1960) citing his words "It is astonishing that the problem of soil drainage has not been understood and the first development projects have been carried out without drainage; consequently these projects were doomed to fail". This can be also seen from the following data of Table 4 published by I. S. Zonn and P. D. Nosenko in the I.C.I.D. Bulletin in 1982.

Table 4

Characteristic Data on the Prospects of Iraqi Water Management

	(1000 ha)
Total area	43492
Cultivated territory	5290
irrigated from it	4300
drained from it	1550
Planned for future irrigation	8000
drainage	—

Considering the explanations above it is clear that now drainage development has become very urgent. It must be solved in the nearest future, because without proper drainage, development of irrigation and increase of agricultural production is quite impossible. But before treating this question a survey is presented on the situation in Hungary.

3. Hungary

Hungary is a country of temperate climate. Its territory is also crossed by two rivers, the Danube and the Tisza. Especially in the valley of the Tisza river similar features can be found as in Mesopotamia.

Climate

Several characteristic data of the climate of the Tisza valley on average precipitation and temperature are included into Table 5. Climatic belts of Hungary can be studied on Figure 4. The belt marked as I. is the most arid. It belongs to the Tisza valley to the greatest part.

Table 5

Climatic data of the Tisza valley

average precipitation (50 years)	540–610 mm
average precipitation in vegetation season	300–350 mm
minimum precipitation in vegetation season	125–150 mm
maximum precipitation in vegetation season	500–600 mm
average yearly temperature	10 °C
air humidity during summer	50–80%

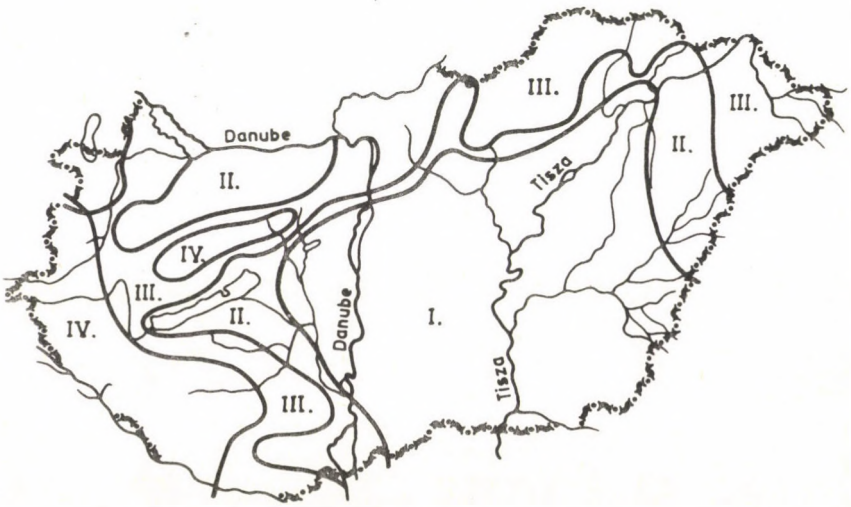


Fig. 4

Yearly evaporation is about 600 mm. The daily values are fluctuating between 3–8 mm/day from April to September.

Corresponding these values of the evapotranspiration of crops, well supplied with water may reach 6–8 mm/day as maximum.

Agriculture

Agricultural production (including husbandry) has reached a considerably high level. Intensive large scale farming has been developed during the last decades. Although cultivated lands and agricultural population has decreased due to industrialization, yields and husbandry products have been increasing continuously. Some average yields in 1982 were the following

barley	3.3 t/ha
wheat	4.4 t/ha
maize	6.8 t/ha

Orchards, vineyards supply the entire population with fruits, grapes and wines. Moreover, significant quantities of them are exported. Similar is the situation in vegetable production.

The livestock in the agriculture is numerous and of high productive capacity. The average weight of the cows is about 500 kg and their milk production exceeds 3500 kg/year. The structure of animal breeding can well be characterized by the following numbers for 1982

cattle	1 922 000
pigs (sow heads)	9 035 000
sheep	3 183 000

The number of horses is negligible and so are those of goats, donkeys, mules etc.

Pro capita meat production of 1982 (Table 6) may best show the productive capacity of Hungarian agriculture.

Table 6

Meat production of several countries in kg/inhabitant in 1982

Hungary	145
The Netherlands	137
Belgium, Luxembourg	116
German Democratic Republic	111
U.S.A.	109

Only 14% of the active working population is occupied in agriculture, that means about 1 million persons. It is less than 10% of the entire population.

The rivers

The Hungarian sections of the two rivers are also middle reaches. Several characteristic data of the rivers are presented in the following tabulation.

Table 7

Characteristic data of the Hungarian sections of the Danube and Tisza rivers

	Length (km)		Catchment area (km^2)	
	total	Hungarian reaches	total	in Hungary
Tisza, before regulation	1477	1477	157 186	157 186
after	977	977	157 186	
Danube	2860	691	817 000	93 036

The length of the levees built for protecting about 2,5 million ha-s against floods is more than 4000 km.

The water quantities transported by the two rivers across Hungary can be estimated

— for the average flow of the Danube at Budapest 71.5 km^3/year
— for the average flow of the Tisza at Szeged 21.3 km^3/year

The Tisza river was meandering (before its regulation) having reached the present Hungarian border. Since the riverbed of the Tisza is pending—like the twin rivers of Mesopotamia—flood waters have found here also an easy way to expand in the lower flood plain. Permanent and periodical lakes and ponds, marshes and moors came into being. Figure 5 clearly shows the meandering river and the lakes and marshes before drainage.

It was clear that water management works were to be begun with drainage (and not irrigation).

Drainage

Conscious and predetermined drainage activity had begun 150 years ago. (Two centuries after the Turkish occupation of Hungary.) Removal of excess waters from the Tisza valley had begun with

— the regulation of the river, then
— flood protection levees were next built and finally
— excess waters from the surface (lakes, marshes, moores) were drained.

The excess waters were collected and lifted back into the river solving the following tasks:

— liquidation of periodical and permanent lakes, marshes and moors,
— lowering the ground water table below the 2,0—2,5 m critical depth,
— draining off the surface runoff of the precipitations (irrigation etc.).

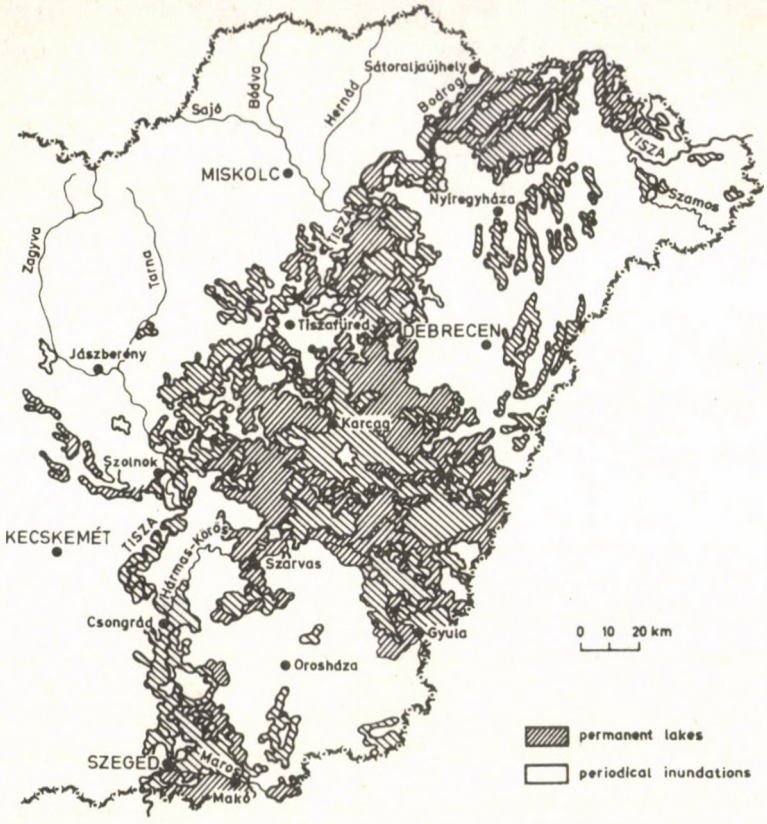


Fig. 5

Saline and alkaline soils have relatively smaller extent in the Tisza valley—about 0.5 million ha—and moreover they are alkaline and clayey, i.e.: they may not be reclaimed by leaching alone. But the removal of leachats has also been necessary.

Summing up: the total length of the drainage canals exceeds 30 000 km. The total discharge capacity of their outlets is about

$$800\ \mathrm{m^3/s}$$

that means a specific discharge capacity of about

$$20\text{–}25\ \mathrm{l/s \cdot km^2}$$

relating to the whole drained territory.

Irrigation

Irrigated farming has recently been developing in the Tisza valley, partly because due to its climate there is generally enough precipitation for the main crops, partly

because it was never populated too densely during the history, so that irrigation would have been necessary. Recent development of irrigation has begun together with large scale farming in the early sixties of this century.

Excess water quantities of irrigation—to be drained off—have never had any great importance. It is partly due to the relatively small extent of the irrigated lands (cca 400 thousand ha), on the other hand sprinkling irrigation method has been applied on the greater part of the irrigated territory.

Concepts

Having achieved the completion of the drainage works significant irrigation investments are planned for the future, due to the considerable resources available in Hungary. As for the prospects of drainage development—only relatively small tasks have been foreseen.

I. S. Zonn and P. P. Nosenko give the values of Table 8 for Hungary.

Table 8

Characteristic Data on the Prospects of Agricultural Water Management in Hungary

	(1000 ha)
Total area	9303
Cultivated territory	5471
irrigated from it	487
drained from it	4262
Planned for future irrigation	2013
drainage	238

It can be seen from the above survey that *Hungarian agriculture may be listed to the first ones of the World and its further development has no water management limits.* This position has been achieved by an activity where drainage was first solved and followed by the development of irrigation.

4. Drainage requirements of Mesopotamia

In areas where water management conditions are similar to Hungary, similar activity promises to have the best prospectives. It was soon recognized also by the Government of the young Republic of Iraq, that a well-planned harmony is necessary in the human activity carried out in the valleys of the twin rivers.

Existing plan

After many years of preparatory work a "General Scheme of Water Resources and Land Development in Iraq" was accomplished in 1975 by a team, involving Iraqi and Soviet experts. It is considered as a Master Plan for planning and designing irrigations and other works of water management. The General Scheme does not deal at all with marshes and marshy territories. Nevertheless these territories are still increasing. (Knappen et al. (1953)).

The backbone of the planned centralized drainage system of the General Scheme, the so called

Main Collector Canal

will have

$283\ m^3/s$

discharge capacity at its tail section to the Arab Gulf. Above the water quantities drained off by this network

$101\ m^3/s$

discharge will directly be drained into the different salty lakes, already existing in the depressions of Mesopotamia. And finally

$15\ ^3/s$

is planned to be lifted back into the Tigris. This way the total system for the entire Mesopotamian valley (that has an extension of cca 80 000 km^2) will represent a specific drainage capacity of roughly

$5.0\ l/s\ km^2$

It is thought that this quantity is highly underestimated and has to be considerably increased. The reexamination of the present values may follow the estimation as below.

Estimation

As has already been mentioned the water quantities to be drained off in the Mesopotamian valley include four main items. Therefore, the water quantities to be removed (drainage requirement) from the flood plain between the two rivers may (only) be estimated by the following formula.

$$D_r = I_e + G_{wl} + L + P_e + W_s$$

where D_r = Drainage requirement (water quantities to be removed)
I_e = Excess water quantities of irrigation
G_{wl} = Lowering the ground water table
L = Leachats (salty water quantities of leaching) to be removed
P_e = Excess water quantities of precipitations
W_s = Water cover on the surface of the marshes

I_e

Considering that the efficiency of surface flooding irrigations is quite low in the Mesopotamian valley and essential improvement of it may not be expected—not even in the future, it can be concluded that excess irrigation water will percolate under the root zone. This water quantity is to be removed, by all means, because it also contains harmful salts solved from the upper soil layers. If only 200 mm is estimated as water application rate for each of the 5 irrigations assumed in one year, then round 1000 mm is delivered to the irrigated territories yearly. Further—if on the base of F.A.O. publication ("Drainage Design Factors" 1979)— —30–40% is considered as a percolating water quantity, it can be estimated to be

350 mm/year.

G_{wl}

Lowering the ground water table is a necessity present almost over the entire flood plain of the twin rivers. If only lowering it by 1–2 m the existing high ground water table is taken into account, then another

100 mm/year

is to be drained off from the pore volume of the soil profile—setting it free from gravitational water.

$L + P_e$

Leaching is taken into consideration twice a year apart from the percolating excess irrigation water. If 50–100 mm may leach harmful salt concentration out of the 3 phase soil layer this item may be estimated to be 100–200 mm pro year. The local excess precipitations to be removed—on the other hand—can be neglected on estimating drainage requirement for the entire catchment area of the Mesopotamian valley.

In this way—for the time being not speaking of the marshes,

550–650 mm/year

— as a minimum is to be drained off. It means a specific drainage requirement (total discharge capacity of the drainage network) of

$$17 - 20\ l/s \cdot km^2 .$$

This specific drainage requirement would mean that the total discharge of the drainage outlets (main canal at the tail section outlets to the salty lakes lifting back to the Tigris) should be

$$1360 - 1600\ m^3/s$$

for the 8 million ha of irrigated land in the future, instead of the planned cca 400 m^3/s.

The above values may not be considered at all as an overestimation because it means a drainage factor

$$1.5 - 1.8 \text{ mm/day}$$

that represents quite a low value even according to the F.A.O. publication referred to above.

W_s

Reclamation of the enormous marshes is unavoidably necessary sooner or later. This gigantic task needs further investigations. Two main lines of these future but urgent examinations have to clarify

— the water quantities (depths and area) to be removed and

— where do these water quantities come from (floods of the rivers, underground seepage, etc.).

Anyhow, if the territory of these marshes can be isolated from restoration of its water cover, each 1 m high water column covering the surface to be drained in a 10 years' period needs an additional continuous discharge capacity of

$$100 - 150 \text{ m}^3/\text{s}$$

over the values mentioned above.

5. Conclusions

In spite of a number of differences between the conditions of water management in Iraq and in Hungary, there are several similarities showing that taking over Hungarian experiences for the water management of Iraq is possible and recommendable.

The following most important similar features are to be stressed.

— Riverbeds are pending in the case of Mesopotamia and in the Tisza valley too.

— Extent marshes and moors, permanent and periodical lakes and ponds were to be drained in the Tisza valley. At the same time such large territories still exist in Mesopotamia.

— Saline and alkaline soils were and are to be reclaimed on the drained flood plain of the Tisza valley. This task is still to be solved in Mesopotamia on even larger territories.

— Saline drain waters were and are to be removed as well as lowering the ground water table below the critical level was and is necessary in the Tisza valley. The same task is still to be solved in the Mesopotamian valley too. Etc.

Iraq is a developing country which is not able to feed its population without significant increase in its agricultural production. But this development may not be

carried out without land reclamation. It seems practical and promising to utilize experiences, knowledge and know-hows already accumulated and available in the Hungarian water management and agriculture.

It is therefore thought that the technical and scientific cooperation of the two countries should include the tasks connected to

Drainage and Agricultural Development of the entire Mesopotamian valley.

Considering that the reclamation of the whole Mesopotamia represents a gigantic task—that requires decades to realize it—the order of the actions required fully corresponds to the present situation of both countries. Namely

i) first the existing concepts of drainage are to be revised and detailed according to the requirements of agricultural development (time required: 1–2 years);

ii) the designs of a smaller pilot drainage basin of satisfactory size could be prepared, including not only water management works, but also those of agricultural development (establishing cooperatives and/or state farms, infrastructure, settlements, schools, roads, etc.) (time required: 2–3 years);

iii) realization of the designs under ii) i.e. construction of all works designed for the pilot drainage basin should be executed (time required: 3–4 years);

iv) operation of all works (water management and large scale farming) in the pilot drainage basin should be put into normal operation (time required: 5 years).

v) after several years of experiences with the operation of the pilot drainage basins the enormous task of reclaiming entire Mesopotamia can be planned, designed and executed (time required: several decades).

References

1. Balogh J.: *Irrigation Efficiency with Surface Irrigation.* I.C.I.D., Tokio 1963
2. Buringh, P.: *Soils and Soil Conditions in Iraq.* Min. of Agriculture, Baghdad 1960
4. *FAO, IRRIGATION AND DRAINAGE PAPER: Drainage Design Factors* F.A.O. Rome, 1979
5. *General Scheme of Water Resources and Land Development in Iraq,* "Selkhopromexport"—Min. of Irrigation, Baghdad–Moscow 1975
6. Petrasovits I.–Balogh J.: *Növénytermesztés és Vízgazdálkodás* (Crop production and Water management), Budapest, Mezőgazdasági Kiadó, Budapest 1975
7. Polinov, B. B.: *Izbraine Trudi,* Moscow, Selkhizdat 1975
8. Tippets–Abbets–Mmcarty–Stratton: *Report on the Zubair Irrigation Project,* Development Board, Baghdad 1956
9. Zonn, I. S.–Nosenko, P. P.: (1982) *Modern Level and Prospects for Improvement of Land Reclamation in the World,* ICID Bulletin, New Delhi 1982, July

Acta Technica Academiae Scientiarum Hungaricae, 98 (3—4), pp. 155—180 (1985)

INVESTIGATING THE OPTOELECTRONIC PARAMETERS OF SILICON PHOTOCELLS TAKING INTO CONSIDERATION OF PRODUCTION TECHNOLOGICAL FEATURES

J. BERKECZ*–K. SZENTIDAY**

[Received: 13 March 1984]

The publication investigates the optical and electrical properties of the silicon photocells with 7 mm^2 and 100 mm^2 light sensitive surfaces manufactured by the Enterprise for Microelectronics (Mikroelektronikai Vállalat). After summarizing the production technology of the photocells, it investigates the influence of the specific resistance of the basic crystal, of the orientation of the substrate-slice and of the penetration depth of the pn junction on the internal impedance of the photocells, on the temperature dependency of responsivity and on noise characteristics. It has been stated that, from the point of view of internal resistance and threshold sensitivity, photocells made of basic materials with a high specific resistance have proved to be better. The MEV products have been compared to those produced by Telefunken.

1. Introduction

Semiconducting photodetectors can be used in several fields of industrial electronics, measuring techniques and electronics for public consumption ranging from optical telecommunications systems and fibre-optical data transmitters to photometers. Earlier they used to make photodetectors with selenium and then germanium but they have gradually been replaced by silicon sensors with considerably more favourable parameters and these can be used within the range of visible light and the neighbouring infrared up to about the wavelength of 1050 nm.

The best known type of semiconducting photodetectors are the photodiode with one single pn junction and the PIN (*p*-intrinsic-*n*) diode. However, phototransistors with two pn junctions, photo-Darlingtons that can be made from two transistors and photothyristors with several pn junctions (TRIAC-s) are also wide spread.

The illuminated pn junction can operate in three different modes, as a photodetector, as a photodiode or avalanche-photodiode provided with reverse bias and as a photocell without bias (see e.g. Szentiday [10]).

Up-to-date silicon photocells may be applied in a wide field. They are relatively cheap photodetectors. Their photo-sensitive territory is large (as big as several sq.

* J. Berkecz, 1131 Budapest, Mosoly u. 42/b, Hungary.
** K. Szentiday, H-1013 Budapest, Attila út 25, Hungary.

mm . . . sq. cm), this makes their sensitivity great and the fact that they do not need reverse bias facilitates their use. In Hungary, it is the Enterprise for Microelectronics (Mikroelektronikai Vállalat) that makes photocells with a photo-sensitive surface of 7 sq. mm. and 100 sq. mm., in an unencased form. As, up till this moment, we have no other kind of home–made photodetector at our disposal,[1] the question arises whether the Si–photocells can be used in applications satisfying needs for greater precision, i.e. in territories where till now only imported photodiodes available at higher prices (such as the PIN–diodes) have been used. The primary goal of our investigation is the more thorough exploration of this problem.

2. Theoretical part

2.1. The current-voltage characteristic of the illuminated pn junction

When the photosensitive surface of the photodiode is illuminated, the electric field of the pn junction separates the electron–hole pairs brought about by the light, thus drifting the electrons into the *n* and the holes into the *p* type domain. Setting up a galvanic relation between the two poles of the photodiode, there is a current flowing in the circuit whose direction is identical with that of the diode's reverse current and is added to it. The photodiode with a reverse bias is a linear device. The photocurrent changes in direct proportion to the illumination. This is what the number of graphs in the first quarter expresses in Fig. 1.

When operated as a photocell, the pn junction works as a current source converting light energy into electrical energy. If the photocell's terminals are short–circuited, the short–circuit current may be measured, whereas, when unloaded, there is an open circuit voltage between the two points of the cell. The graphs of the photocell are in the fourth quarter of Fig. 1. In this domain, the device produces a reverse current and in the meanwhile on its junction an U_F forward bias voltage comes about. This is a state of non–balance and it can only be maintained at the expense of the radiation power used by the photocell.

When short–circuited, similarly to the reverse biased photodiode, the photocell operates linearly whereas its open circuit voltage changes approximately logarithmically with illumination. If the corners of the photocell are closed off with an R_L resistance, the work line suiting: the R_L value can be plotted. If the light current increases, the induced photoelectronic current increases as well and is distributed between the R_b internal resistance of the photocell and the R_L load resistance of the outer circuit. Both when increasing the light current and when increasing R_L, the internal resistance decreases since there is a larger and larger opening voltage coming

[1] In 1985 Enterprise for Microelectronics developed a phototransistor.

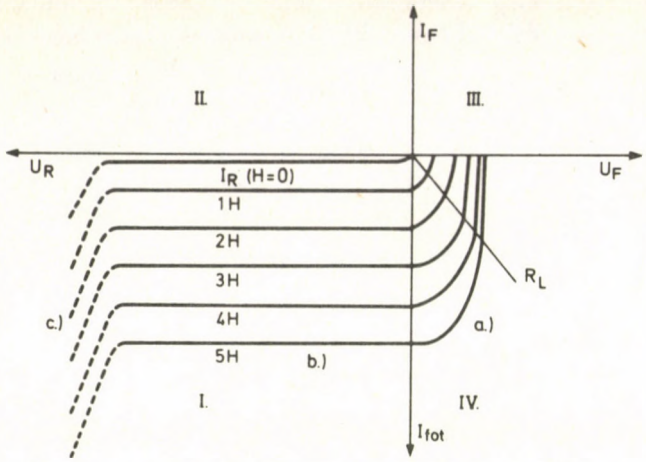


Fig. 1. Current-voltage characteristic graphs of an illuminated pn junction as a function of *H* illumination *a*) photon-voltage range, *b*) reverse bias range *c*) avalanche range

about in the junction. Therefore R_b shunts R_L and it is the consequence of this that the response signal changes non–linearly with the illumination.

The equivalent electrical circuit of the photocell operated in short circuit is identical with the short–circuited photodiode's switch (Fig. 2.). The photocurrent induced in the diode is represented by the current generator, whose R_b internal resistance is the dynamic resistance of the pn junction and the C_t junction capacity is connected to this in a parallel way. In *4.1.* we are going to detail the interpretation and role of the internal resistance of the photocell operated in a short–circuited mode in connection with the photocell joined to operational amplifiers. In an ideal case, with perfect short circuiting and very little illumination, the R_b internal resistance, just like the internal resistance of the short–circuited photodiode, has a very high value. When operating as a photocell, the junction-capacity is relatively large, since the depletion layer of the pn junction biassed in a forward direction is much narrower as compared to the depletion layer of the pn junction supplied with a reverse voltage. Apart from this, C_D diffusion capacity arising from charge accumulation is also added to the junction capacity of the illuminated photocell.

It follows from what has been mentioned above that the impedance of the short-circuited photocell—according to the substitution switch seen in Fig. 2.—is the parallel resultant of a resistance with high ohm value and of a relatively high value capacity. However, the internal impedance of the photocells is not constant even when short–circuited, but depends on illumination and temperature. On increasing the illumination and the temperature, the R_b internal resistance decreases. The reason for this may be found in the fact that the charge–carrying pairs induced by light and heat energy reduce the inner electrical field of the pn junction, the density of the charge-carriers accumulated in the transition region increases and thus the photocell with the pn junction behaves more and more like a photo-conductor instead of a rectifier.

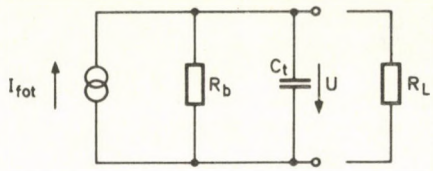


Fig. 2. Equivalent circuit diagram of a photocell

On increasing the illumination and the temperature, the capacity of the pn junction—most probably—is going to increase to a small extent, mainly as a consequence of the increase in the diffusion capacity.

2.2. The temperature dependency of the responsivity

The electrical response signal given as a result of a unitary incident light power output is called responsivity. At a first approach, this depends on the material of the photocell, the technology it was manufactured by and the size of its photo-sensitive surface. The spectral responsivity, i.e. the electrical response signal measured as a function of the wavelength, is also dependent on the material of the photocell and its constructional characteristics. If we wish to use the photocell for precision measuring purposes, the temperature dependency of the responsivity also has to be taken into account.

Most parameters of the semi-conducting devices are, to a smaller or greater extent, dependent on temperature as Fermi level is a function of temperature [9]. In the case of photo-sensitive semi-conducting devices, the temperature dependency of the photo–absorptional properties and the reflexion capacity of the semi–conducting mono–crystal are also added to this, though this latter is of a negligible extent.

In the case of an illumination perpendicular to the plane of the pn junction (Fig. 3.) the total current flowing through the plane of the junction may be expressed as

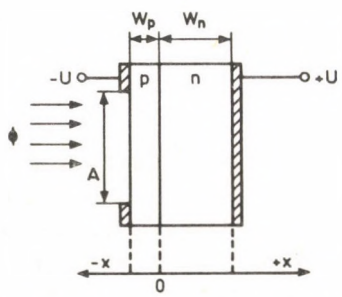


Fig. 3. Cross–section diagram of a pn junction for calculating the photon–current. The zero point of the coordinate system is in the plane of the junction from left of which is the *p* type domain and right of which is the *n* type domain

$$I = I_R - I_{fot} \tag{1}$$

where

$$I_R = I_s \left[\exp\left(\frac{qU}{kT}\right) - 1 \right], \tag{2}$$

is the diode's dark current and I_{fot} is the photo-current. The quantities in relation (2) are

I_s the saturation current member in the diode-equation (see formula [16]),
U the voltage applied to the pn junction,
k the Boltzmann constant,
T the temperature and
q the charge of the electron.

The photo-current is

$$I_{fot} = -qAg(0)[L_1 + L_2], \tag{3}$$

where

$$L_1 = \frac{L_p}{\alpha^2 L_p^2 - 1} \left[\alpha L_p - \alpha L_p \exp(-\alpha w_n) \operatorname{sech} \frac{w_n}{L_p} - \operatorname{th} \frac{w_n}{L_p} \right] \tag{4}$$

$$L_2 = \frac{L_n}{\alpha^2 L_n^2 - 1} \left[\alpha L_n \exp(\alpha w_p) \operatorname{sech} \frac{w_p}{L_n} - \alpha L_n - \operatorname{th} \frac{w_p}{L_n} \right] \tag{5}$$

$$g(0) = \alpha \eta_k (1 - R) \Phi \exp(-\alpha w_p). \tag{6}$$

In the above formulas,

A is the photo-sensitive territory of the photocell,
$g(0)$ the extent of generation in the plane of the junction at $x = 0$ point, see Fig. 3.
Φ the photon-current, i.e. the number of photons at a unit of place and time,
R the reflexion capacity,
η_k the amount of quanta,
w_p the width of the layer p,
w_n the width of the layer n,
L_p the hole diffusion length
L_n the electron diffusion length and
α the photo-absorption factor.

The given relations are valid for abrupt pn junctions [1].

Short-circuit current. If the equation system (1)–(2) is solved for the case of $U = 0$, the relation

$$I_k = I_{fot} - I_s \tag{7}$$

may be obtained for the I_k short-circuit current. In practice, the photocells are shaped in a way so that the illuminated range (in our case the layer with the p type) is chosen to be very narrow, whereas the opposite side of the junction (the n–type layer) is chosen to

be relatively wide. In this case the

$$w_p \ll L_n \qquad \text{and} \qquad w_n \gg L_p$$

inequalities are fulfilled and the

$$L_1 \cong \frac{L_p}{1+\alpha L_p}; \qquad L_2 \cong \frac{\alpha L_n^2}{\alpha^2 L_n^2 - 1}\left[\exp(\alpha w_p) - 1\right] \tag{8}$$

approximations may be applied. If the equations (6) and (8) are substituted in (3)

$$I_{\text{fot}} \cong C_\lambda \left\{ \left[\frac{\alpha L_p}{1+\alpha L_p} - \frac{\alpha^2 L_n^2}{\alpha^2 L_n^2 - 1} \right] \exp(-\alpha w_p) + \frac{\alpha^2 L_n^2}{\alpha^2 L_n^2 - 1} \right\} \tag{9}$$

may approximately be obtained for the photo–current, where

$$C_\lambda = Aq\eta_k(1-R)\Phi.$$

The responsivity is

$$S_p = \frac{I_{\text{fot}}}{P}, \tag{10}$$

where P is the incident light power

$$P = \frac{hc\Phi}{\lambda} A. \tag{11}$$

In the above formula, h means the Planck-constant, c stands for the velocity of light and λ for the wavelength of the light. Substituting (9) and (11) into (10) we get

$$S_p = \frac{\lambda q \eta_k (1-R)}{hc} \left[\left(\frac{\alpha L_p}{1+\alpha L_p} - \frac{\alpha^2 L_n^2}{\alpha^2 L_n^2 - 1} \right) \exp(-\alpha w_p) + \frac{\alpha^2 L_n^2}{\alpha^2 L_n^2 - 1} \right]. \tag{12}$$

In equation (12) the α absorption factor and the L_n and L_p lengths of diffusion are temperature dependent, provided we regard η_k and R temperature dependency as negligible. The temperature dependency of α can be estimated based on those said in the References [5], where the spectral values for $\alpha(\lambda)$ are given for the cases of 300 K and 70 K. Knowing the raw material and the technology used to produce the photocell, and being familiar with References [8], the value of the diffusion length and its temperature dependency may approximately be determined.

As the $\alpha(\lambda, T)$ absorption factor is a function of both the wavelength and the temperature, differing temperature factors are to be expected as regarding to the responsivity in the case of the various wavelengths and pn junctions with varying penetration depths.

Open circuit voltage. U_0 open circuit voltage may also be determined from equations (1)–(2) by substitutions $I = 0$ and $U = U_0$. In this case

$$U_0 = \frac{kT}{q} \ln\left(1 + \frac{I_{\text{fot}}}{I_s}\right). \tag{13}$$

If the photocell is illuminated and the $I_{fot} \gg I_s$ equation is taken into account,

$$U_0 \cong \frac{kT}{q} \ln \frac{I_{fot}}{I_s}. \tag{14}$$

Forming the differential quotient of (14) according to temperature

$$\frac{dU_0}{dT} = \frac{k}{q} \ln \left(\frac{I_{fot}}{I_s} \right) + \frac{kT}{q} \frac{d}{dT} \left[\ln \left(\frac{I_{fot}}{I_s} \right) \right]. \tag{15}$$

In this equation, we have the temperature dependent expression of the I_{fot} photocurrent and the I_s saturation current member. For I_s, we arrive at

$$I_s \cong K T^{1.4} \exp \left(- \frac{W_G}{kT} \right) \tag{16}$$

in the case of silicon, based on References [7]; here W_G is the forbidden bandwidth of silicon and K is a constant non–dependent on temperature. I_{fot} means the short–circuited current defined previously and its temperature dependency is small as compared to I_s where T is present in the exponent. By presuming that $I_{fot} \cong$ constant, we get

$$\frac{dU_0}{dT} = \frac{k}{q} \ln \left(\frac{I_{fot}}{I_s} \right) - \frac{kT}{qI_s} \left(\frac{dI_s}{dT} \right). \tag{17}$$

Performing the derivation according to T in (17),

$$\frac{dU_0}{dT} = \frac{k}{q} \left[\ln \left(\frac{I_{fot}}{I_s} \right) - 1.4 \right] - \frac{U_G}{T} \tag{18}$$

may be obtained as a result, where $U_G = W_G/q \cong 1.12$ V in the case of silicon.

If the photo–current is not too large as compared to the saturation the member in brackets in relation (18) is negligible and, for the thermical coefficient of the open circuit voltage (TK), we may get

$$TK \equiv \frac{dU_0}{dT} \cong - \frac{U_G}{T}. \tag{19}$$

As seen, TK is negative, therefore the open circuit voltage decreases when the temperature increases, contrary to the short-circuited current which generally increases when the temperature increases.

2.3. Noise, detectibility threshold

The greatest part of the noise of photocells without illumination is caused by the shot noise that can be calculated with the help of the saturation current member (I_s) and by the thermal noise of the series resistance of the silicon crystal forming part of the

photocell. Since the photocell can be represented by a current generator, as shown in Fig. 2., its noise is also generally characterized by noise currents. The effective value of the shot noise current is

$$I_{ns}^2 = 2qI_s \Delta f, \tag{20}$$

where Δf is the band width; the effective value of the thermal noise current is

$$I_{nt}^2 = \frac{4kT\Delta f}{r_{sc}}, \tag{21}$$

where r_{sc} is the series resistance of the crystal. In addition, depending on the technological procedures, $1/f$ noise may also arise and this generally becomes significant under 10 Hz. Noise currents are mostly given in reference to a unitary band width in units of $A \times Hz^{-1/2}$.

Similarly to other photodetectors, a noise equivalent power (*NEP*) is characteristic of the threshold sensitivity of the photocells and this is the value of the input light that can bring about an electrical output signal level equivalent to the output noise power arising without the input signal.

According to this, therefore

$$NEP = \frac{\text{noise current, } A \times Hz^{-1/2}}{\text{current sensitivity, } A \times W^{-1}}. \tag{22}$$

Therefore, in order to determine the *NEP* value of detectors, the effective value of the noise current and the responsivity in $A \times W^{-1}$ has to be measured. The dimension of the NEP is $W \times Hz^{-1/2}$ according to expression (22) [6].

3. Production technology of silicon photocells

In order to manufacture photocells, silicon monocrystal wafers with diverse specific resistance and diverse crystal orientation were used. So as to ensure a high value of responsivity, the p^+nn^+ structure seen in Fig. 4. was set up. The backside n^+ layer keeps the charge–carriers induced by the absorbed photons at a distance from the backside that has a great recombinational velocity. As a result the lifetime of minority charge-carriers should be approximately identical with the bulk lifetime free of surface recombination.

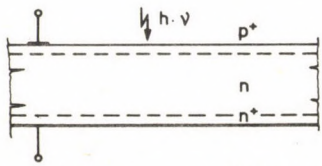


Fig. 4. Cross-section structure of a MEV made photocell

As it has already been pointed out, the α light absorptional factor is wavelength dependent, and, in the case of Si, the light-waves corresponding to the visible range are in practice absorbed within 5μm. That is why, so as to increase the responsivity, it is necessary to produce a so called shallow *pn* junction with a penetration depth as small as possible.

The technological process was modelled with the help of the HIPREM – 1 one-dimensional process simulation program [12]. The distribution profiles of the additives estimated for the various cases of raw material and technology variants are displayed in the a), b), c) and d) varieties of Fig. 5. In the figures, *C* concentration was shown as a function of d distance taken from the surface of the crystal substrate.

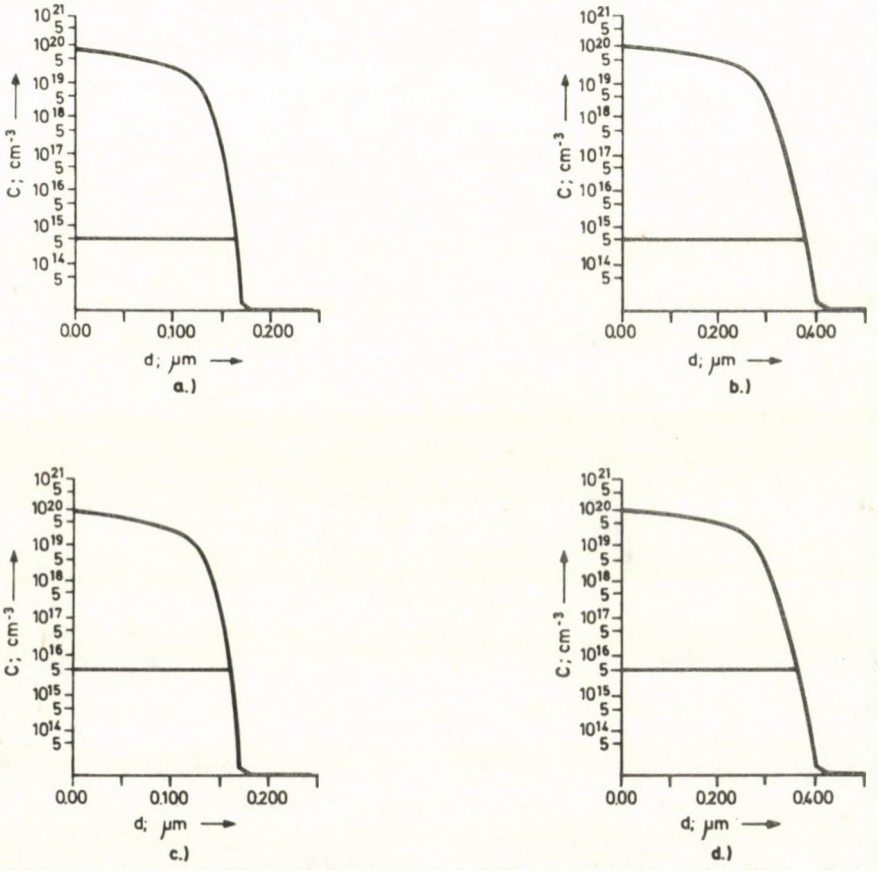


Fig. 5. Distribution profile of additive density depending on *T* temperature and *t* time of diffusion, in the case of Si wafers with different specific resistances
a) 4...6 ohm cm, $T = 890$ °C, $t = 60$ min.
b) 4...6 ohm cm, $T = 950$ °C, $t = 60$ min.
c) 1.2 ohm cm, $T = 890$ °C, $t = 60$ min.
d) 1.222 ohm cm, $T = 950$ °C, $t = 60$ min.

The most critical operation during the process of making photocells is the boron diffusion. This was performed by a modern, planar diffusion source. BN plates were used for this purpose—arranged as one plate plus two pieces of silicon wafers. The BN plate—Si wafer arrangement is shown in Fig. 6.

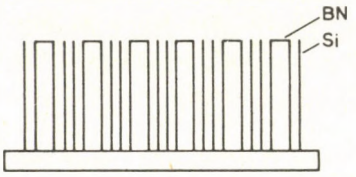


Fig. 6. BN plate–Si wafer arrangement

Great attention was paid when placing the quartz boats containing the BN plates and the silicon wafers into the reactor (the diffusion tube). If the penetration of the boat takes place too quickly, the carefully adjusted temperature and thermal profile changes and the silicon wafers are suddenly warmed up; there is a thermal shock. The rapid placing of the boat deteriorates the reproductibility of the diffusional parameters and the structural deficiencies caused by the thermal shock manifest themselves as a recombination center thus reducing the lifetime of minority charge–carriers. Life time and the L_n, L_p diffusion lengths proportional to them influence responsivity according to equations (9) and (12).

In order to avoid the above mentioned undesirable effects, we applied an automatic wafer–motion, choosing the speed of the movement to be 12 cm/min, relying on Kiss' data [4] and our own results [2].

4. Results of the investigations

The application of the photocells for precision measuring technique purposes necessitates a thorough and careful exploration of the further parameters—apart from the data sheet properties—as well as the separate, individual investigation of the different types, sample by sample. At the same time, it is important, from the point of view of the manufacturing process as well, to what extent the circumstances of manufacture effect the optical and electrical parameters of the photocells.

The photocells forming the object of our investigations were classified into various groups, and from now on we are going to refer to the different varieties according to this classification.

Photocells made of silicon wafers with a 4. . . 6 Ωcm specific resistance, of *n*-type and with ⟨100⟩ crystal orientation were classified into group–*a*. Samples made of silicon wafers with 1. . . 2 Ωcm specific resistance, of *n*-type and with ⟨111⟩ crystal orientation belong to group–*b*. Photocells in these two groups possess a 7 sq. mm. light-sensitive surface. The raw material of those belonging to group–*A* is identical with

the raw material of the photocells in group–*a*, but their light–sensitive surface is 100 sq. mm. in size. In all groups, there are samples with a penetration depth of 0.1 μm, 0.2 μm and 0.5 μm. For the sake of comparison we have also examined some photocells made by Telefunken. We listed the BPW–12 type photocells with 3.8 sq. mm. photo-sensitive surface, with a TO 18 metal encasement and closed off with a plate-glass sheet into group–*c* and we marked the BPW–35 type that has a photo-sensitive surface of 94 sq.mm. and is made without an encasement as group–*B*. For the sake of clarity, we summarized the properties of the various groups in Table I.

Table I

Properties of the inspected model groups

Model group	Photosensitive surface, mm^2	Property
a	7	4...6 Ωcm, ⟨100⟩
b	7	1...2 Ωcm, ⟨111⟩
c	3.8	BPW 12, Telefunken
A	100	4...6 Ωcm, ⟨100⟩
B	94	BPW 35, Telefunken

4.1. Impedance investigation of photocells

Based on the equivalent circuit of Fig. 2., the z_x internal impedance of the photocells may be described by the equation

$$z_x = \frac{R_b}{\sqrt{1+(\omega C R_b)^2}} \tag{23}$$

where

R_b is the internal resistance,
$C = C_t + C_D$ the sum of the junction capacity and the diffusion capacity and
ω is the angular frequency of the measuring signal.

During our investigations, we measured the ohm and the capacitive member separately as a function of the illumination and temperature.

The arrangement of the R_b measurement is indicated in Fig. 7. We applied a sine signal from a signal generator onto the photocell to be measured and connected the photocell to the input of the operational amplifier with a negative feedback. On rectifying the voltage signal of the preamplifier, we measured a voltage signal on the output proportional to the impedance. In the case of a 10 Hz sine signal, as long as the junction capacity is not very large, the condition $(\omega C R_b)^2 \ll 1$ from expression (23) is fulfilled and

$$Z_x \cong R_b$$

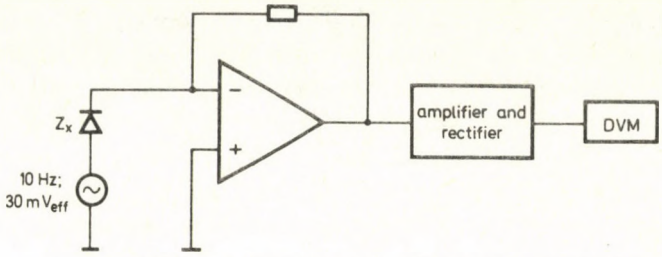


Fig. 7. A block diagram of internal resistance measurements

is obtained. Let us remark at this point that, in the case of photocells with a large photosensitive surface, where the junction capacity is large, the results were corrected with the capacitive impedance. When performing the measurements, we mounted the preamplifier in a separate shielding box, and previously scaled the measuring unit with the help of resistances of familiar values. The upper limit of measuring resistance can be marked at 10 MΩ.

We measured the capacity of the photocell according to a similar principle, choosing 100 kHz as the measuring frequency and applying a fast μA715-type operational amplifier. The circuit diagram is shown in Fig. 8. If ω is sufficiently large, in equation (23) we get $(\omega C R_b)^2 \gg 1$, and

$$Z_x \cong \frac{1}{\omega C}$$

is fulfilled.

It is ensured with both measurement arrangements that the open–loop amplification of the operational amplifier should be sufficiently large at the given measuring frequency for the inversing input to behave as a virtual earthpoint. Thus the photocell connected to the inversing input is short–circuited; therefore, during the measurements, the photocells were examined at a work point corresponding to the short circuit.

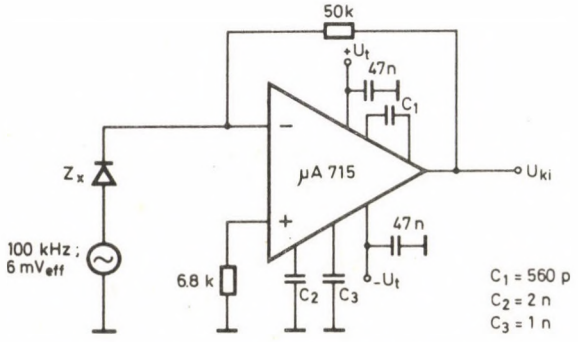


Fig. 8. Connection diagram of pre-amplifier used for photocell–capacity measurements

Results of the R_b-measurement. We examined the temperature dependency of the internal resistance at between room temperature and +80 °C in darkness (with zero illumination). On increasing the temperature, we could observe an exponential reduction of the internal resistance.

If the function

$$R_b(T) = R_{b0} \exp\left(-\frac{T}{T_c}\right) \tag{24}$$

is related to the measured values, the T_c parameter characteristic of the extent of the decrease and the values of the r^2 correlational coefficient indicating correlation are shown in Table II.

Table II

Temperature coefficient of the internal resistance

Sample number	T_c, °C	Correlational coefficient, r^2
a–1	328	0.982
a–2	27	0.989
a–3	61	0.803
b–1	34	0.991
b–2	28	0.998
b–3	37	0.995
b–4	37	0.998
A–1	25	0.993
A–2	33	0.988
B–1	32	0.985

Measurements taken as functions of illumination were performed at room temperature. A 250 W iodine-halogenous lamp set for 2850 K colour temperature was used for this purpose. The accuracy of the lamp's current was maintained with stability for three numbers figures. The maximum value of the applied illumination was 8 . . . 10 klx in the case of smaller photocells, whilst, in the case of samples with large photo–sensitive surfaces, we took measurements up to approximately 1 klx only. Some characteristic data of the inspected photocells are summarized in Table III. For the sake of comparison, we have shown the measured values of a photocell belonging to group–*B*.

Figure 9. displays the illumination dependency of the internal resistance of certain typical photocells.

After surveying the results, we may see that the diverse models/samples have essential differences as to how their internal resistance depends on the illumination. Internal resistances measured as functions of both darkness and illumination are substantially greater in the case of cells belonging to group–a as compared to group–*b*. Furthermore, you may notice that on increasing the photo–sensitive surface, the

Table III

Illumination dependency of internal resistance

Sample number	Illumination			
	0 lx	100 lx	1 klx	5 klx
a–1	>10 MΩ	3.5 MΩ	1.7 MΩ	500 kΩ
a–2	>10 MΩ	8.0 MΩ	1.8 MΩ	500 kΩ
a–3	>10 MΩ	>10 MΩ	2.2 MΩ	450 kΩ
b–1	1.6 MΩ	1.6 MΩ	950 kΩ	90 kΩ
b–2	4.2 MΩ	2.8 MΩ	1.2 MΩ	50 kΩ
b–3	1.4 MΩ	1.3 MΩ	830 kΩ	95 kΩ
b–4	430 kΩ	360 kΩ	170 kΩ	19 kΩ
A–1	8.3 MΩ	800 kΩ	6 kΩ	—
A–2	2 MΩ	650 kΩ	5 kΩ	—
B–1	1 MΩ	350 kΩ	<5 kΩ	—

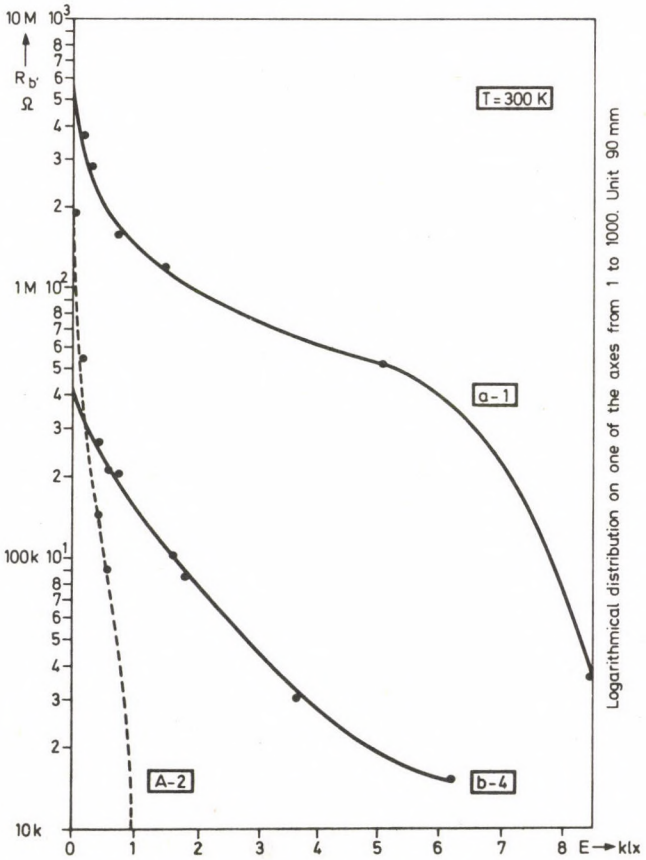


Fig. 9. Illumination dependency of internal resistance

internal resistance decreases to a great extent. This was manifest in the case of both home–made and Telefunken–made photocells.

However, if we take the fluctuation of the internal resistance as function of the changes of light current (lm) per sample, we find that about 14 times as big a light current was characteristic of the large photocells as a result of the 100 sq.mm. and 7 sq.mm. proportion. As seen in Table 2., the internal resistance of photocells belonging to group–*A* is approximately of the same value when measured at 100 lx as the internal resistance of group–a at 1 klx. This allows us to conclude that the above mentioned two models do not essentially differ from each other as far as the value of the internal resistance per surface-unit is concerned.

As for the temperature dependency of the internal resistance, we did not experience such a characteristic distinction between the various photocell groups.

The degradation of the internal resistance is rather important when utilizing photocells in practice. In several territories, for instance in the case of colorimeters or lux-meters where the photocell may substitute for photodiodes with reverse bias voltage, the photocell is used as a linear element connected via direct current, that is in short–circuit state. Figure 10. presents a basic connection of this kind, where the photocell is short–circuited and the amplifier performs a current–voltage transformation. However, as it is well known, the output offset–voltage of the operational amplifier depends on the resistance of the generator as well. Therefore, the behaviour of the amplifier will be effected by changes in the internal resistance of the photocell, in its capacity as generator resistance. Therefore, when connecting it to a current amplifier, the internal resistance of the photocell operating on short-circuit is going to gain significance as a heat and illumination dependent parameter.

Let us have a look at the simple diagram in Fig. 10. and check what has been said by way of calculations. The R_b internal resistance is a function of T temperature and of H illumination, and of I_{fot} photo-current indirectly,

$$R_b(T; I_{fot}).$$

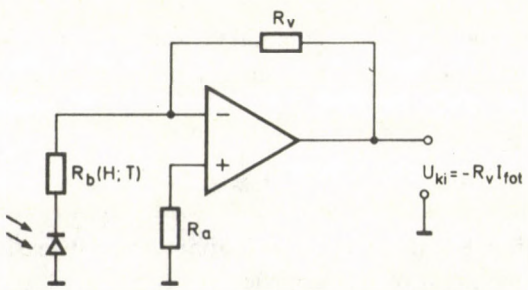


Fig. 10. Fitting a photocell to an operational amplifier

Setting up the total differential of U_{ki0} offset voltage,

$$\Delta U_{ki0} = \frac{\partial U}{\partial I_{fot}} \Delta I_{fot} + \frac{\partial U}{\partial T} \Delta T. \tag{25}$$

The output fault voltage is

$$U_{ki0} = R_v \left[I_{I0} + \frac{I_{0f}}{2} + i_d \right] - R_a \left(1 + \frac{R_v}{R_b} \right) \left[I_{I0} - \frac{I_{0f}}{2} + i_d \right] + \\ + \left(1 + \frac{R_v}{R_b} \right) [U_{0f} + U_d], \tag{26}$$

where I_{I0} is the static input current,

I_{0f} the input offset current,

U_{0f} the input offset voltage

i_d and u_d are the drift current and drift voltage respectively arising as an effect of the 1 °C change in temperature and

R_v is the feedback resistance.

R_a the resistance connected to the non inverting input.

Forming the value of U_{ki0},

$$\Delta U_{ki0} = - \frac{R_v}{R_b^2(I_{fot}; T)} \left[R_a \left(I_{I0} - \frac{I_{0f}}{2} + i_d \right) - (U_{0f} + u_d) \right] \cdot \\ \cdot \left[\frac{\partial R_b(I_{fot}; T)}{\partial T} \Delta T + \frac{\partial R_b(I_{fot}; T)}{\partial I_{fot}} \Delta I_{fot} \right] \tag{27}$$

$\Delta U_{ki0} = 0$, if

$$Ra = \frac{U_{0f} + u_d}{I_{I0} - \frac{I_{0f}}{2} + i_d}. \tag{28}$$

It becomes evident that ΔU_{ki0} can be made zero by properly setting R_a. However, there ceases to be a compensation and the internal resistance changes in equation (27) may become significant with the changes of u_d and i_d. For a complete compensation of U_{ki0}, the member

$$R_v \left[I_{I0} + \frac{I_{of}}{2} + i_d \right]$$

also has to vanish from equation (26). This is approximately feasible with an auxiliary voltage applied to the input of the amplifier.

The effect of the significant internal resistance fluctuations may be reduced by choosing exigent operational amplifier types. What is necessary, above all, is an amplifier with small offset and drift current and, as far as possible, you should make

sure that the operational amplifier and the photocell should have a steady temperature in the measuring device [11].

Measuring capacities. The temperature dependency of the photocells' capacity was inspected at temperatures ranging from +22 °C... +80 °C, with zero illumination. Measurements have shown that capacity increases to a small extent when the temperature is increased. In Table IV, we have marked the capacities taken at room

Table IV

The capacity and temperature coefficient of the photocell

Sample number	C (22 °C), nF	Temperature coefficient pF/°C
a–1	1.20	+2.97
a–2	0.99	+2.05
a–3	0.99	+1.82
b–1	1.73	+1.20
b–2	1.71	+1.86
b–3	1.63	+1.91
b–4	1.63	+0.87

temperature and the capacity changes per 1 °C as a mean value for the above indicated range of temperature. As you may see from the table, the capacities of the cells belonging to group–*a* are smaller, but the amount of the change is greater than in the case of the cells belonging to group–*b*.

The illumination dependency of the photocell–capacity was measured up to about 3...4 klx at room temperature. In the case of greater illuminations, where the capacity increase is accompanied by a significant decrease in the internal resistance, the internal impedance of the photocells can no longer be modelled with the simple substitution diagram seen in Fig. 2. and our measurement produced no appreciable results. According to our investigations, photocells with 7 sq.mm. of photo-sensitive surface are not advised for more exigent measuring purposes in cases when they are illuminated beyond the above mentioned limits (approximately above 200...300 μA photo–current).

The percentage of the capacity's change is displayed in Table V. for certain values of illumination. In the table, the value belonging to zero illumination was defined as 100%. The illumination dependency of the capacity of two cells belonging to group–*a* and to group–*b* respectively can be seen in Fig. 11.

Investigations concerning capacity may mainly become important when photocells are applied with alternating signs. Although the photocell belongs to the slow photodetectors, types with small junction surface are occasionally used for sensing modulated radiation. Publication [3], where the author acquaints us with a digital amplitude-modulated system with a photocell sensor is an example of this. The contacts of the photocell are closed off with an inductance and the value of the

Table V

Illumination dependency of the photocell's capacity

Sample number	Change of junction capacity in percent with respect to 0 lx taken as 100%		
	1 klx	2 klx	3.5 klx
a–1	102	108	134
a–2	103	116	139
a–3	102	104	124
b–1	104	116	145
b–2	102	105	115
b–3	102	104	111
b–4	102	105	122

inductance is selected in a way so that, together with the capacity of the diode, it would form an oscillating circuit tuned to a carrier frequency. Therefore, the change of the capacity is going to result in a distuning of the oscillating circuit while the change in the internal resistance is going to influence the factor of the quality (Q) of the circuit.

A well–known advantage of applying modulated radiation is the elimination of the effect of background illumination. This is particularly important in the case of

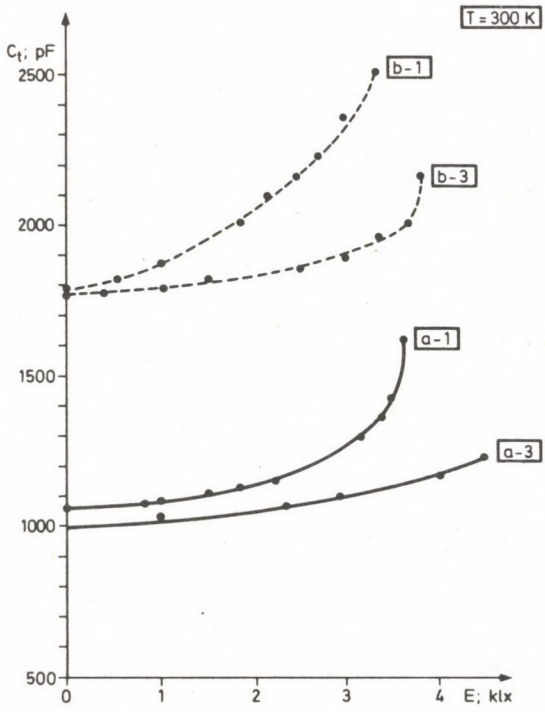


Fig. 11. Illumination dependency of photocell–capacity

optical signal transmission applied in open air. However, if the influence exerted by the background light on the photocell's impedance is significant, it indirectly effects the sign transmission as well.

4.2. Investigating the temperature dependency of the responsivity

A sketch of the arrangement used for measuring the temperature dependency of the responsivity is seen in Fig. 12. The light source presented in chapter *4.1.* was used for the purpose of illumination. The cooling and heating of the photocell were performed with the help of a Peltier-cell whose reference side was cooled with flowing water. The weak pre–vacuum (cca. 5 mbar) applied around the sample served the purpose of

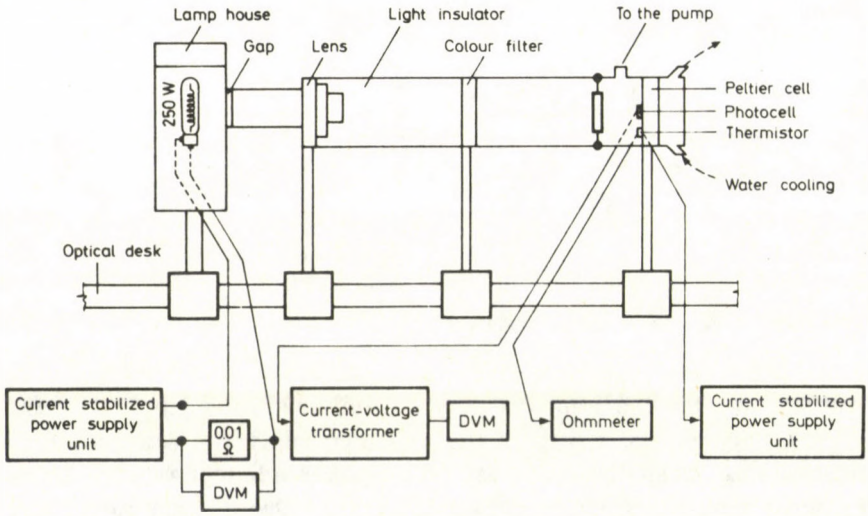


Fig. 12. Measuring arrangement set up for determining the temperature dependency of short circuit current

preventing the vapour from precipitation. The temperature of the photocell under investigation was taken by a flat, Siemens-made thermistor mounted next to it. This had previously been calibrated. The short circuit current of the photocells was measured with the help of an LM 308 operational amplifier. The open circuit voltage was directly measured with a digital voltmeter (internal resistance: 10 MΩ). For investigations with wavelength-dependency, a 620 nm red and a 854 nm infrared interference band-pass filters were used.

From the measurements we determined the relative change of the electrical response signal as a function of temperature. The percentage of change was calculated with the help of the formula

$$\frac{U_{fot}(T)-U_{fot}(T_1)}{U_{fot}(T_1)} \times 100 \equiv \%, \tag{29}$$

where $U_{fot}(T_1)$ is the value of the response signal measured at a reference temperature (room temperature) and the $U_{fot}(T)$ response signal stands for the value measured at the T temperature in question. In the case without the filters, and 1 klx illumination, the measurements were performed at temperatures ranging from $-10\,°C$ to $+80\,°C$.

Short-circuit current. In the first group of our measurements, we investigated the effect exerted on the responsivity by the penetration depth. For this purpose samples with penetration depth of 0.1 μm, 0.2 μm and 0.5 μm, belonging to models of group–*a*, were selected. All the measurements were carried out with red and infrared filters as well as without filters. Figure 13 shows the temperature dependency of the

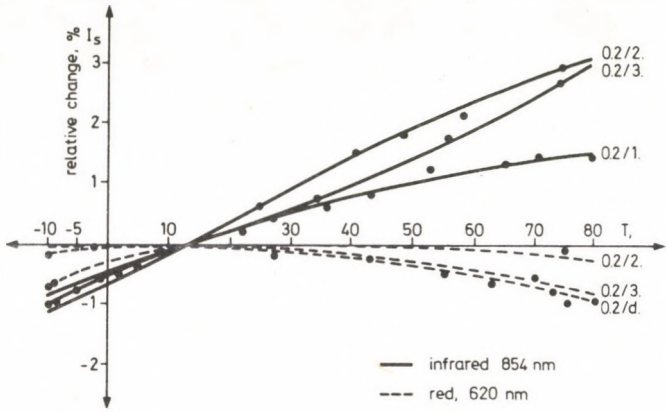


Fig. 13. Change of short circuit current in percent as a function of temperature measured with the help of red and infrared filters

responsiveness of 3 pieces of photocells with 0.2 μm penetration depth as an example. Table VI summarizes the results of the measurements performed, indicating the percentage change arising between the two temperature limits. As seen, the rate of temperature dependency is influenced by the penetration depths of the models to a very

Table VI

Measured values of the temperature dependency of responsiveness

Sample property, number	Difference in % between +80 °C and −10 °C	
	with infrared filter	with red filter
0.1/1.	+1.9	−3.5
0.1/2.	+4.0	+0.6
0.1/3.	+3.6	+0.4
0.2/1.	+2.5	−1.2
0.2/2.	+4.3	+0.7
0.2/3.	+4.0	−0.9
0.5/1.	+3.2	−1.0
0.5/2.	+5.0	−1.0

small extent, whereas a significant deviance may be experienced when measurements are taken with red or infrared filters. At this point, let us comment that the results obtained by measurements performed without a filter approach those obtained with the infrared filter. The fact that the decisive part of the radiation energy of the lamp with 2850 K colour temperature falls into the domain of the nearby infrared is due to this explanation.

So as to evaluate the results, we made calculations using correlations 3. . . 12 in estimating the extent of temperature dependency. The value of the temperature dependent parameters was defined from References [5] and [8] and from the semiconductor material properties of the photocells. Table VII contains the data used for the calculations and the results.

Table VII

Data and result used when calculating the change in percent

Data

Wavelength of filter, nm	Absorptional factor, cm^{-1}	
	T_a	T_m
854	720	940
620	4500	5750

Diffusion length, cm	T_a	T_m
	temperatures	
L_p	0.02	0.05
L_n	0.003	0.003

Results

Wavelength of filter, nm	Differences in % in the cases of diverse penetration depths		
	0.1 μm	0.2 μm	0.5 μm
854	+4.5	+4.4	+4.1
620	+0.73	+0.69	+0.60

If the data seen in Table 6. and those obtained by measurements are compared with those of Table VII and obtained by way of calculations, the weaker temperature dependency experienced in the red filter–range seems to be proved. Considering that the calculations are, to a great extent, approximative, for instance the diffusion length values may vary even in the case of samples taken from one piece, the concordance is satisfactory and testifies that the short-circuit current of the manufactured photocells follows the temperature dependency corresponding to what was expected.

In the second phase of our measurements, we compared photocells belonging to the different groups. Some characteristic results are shown in Fig. 14. On the whole, what we experienced was that the short–circuit current of the samples belonging to group–*a* showed a stronger temperature dependency than that of group–*b*. Measurements seen in the figure were taken with infrared filters.

Open circuit voltage. Figure 15. presents some results from measurements of the temperature dependency of open circuit voltage. The percent rate of change in the open

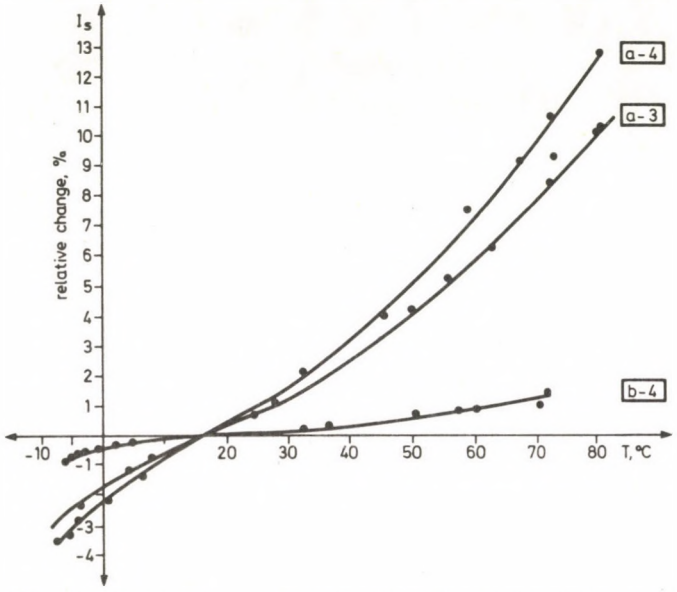


Fig. 14. Change of short circuit current in percent as a function of temperature in the case of some photocells belonging to groups *a* and *b* measured with the help of an infrared filter

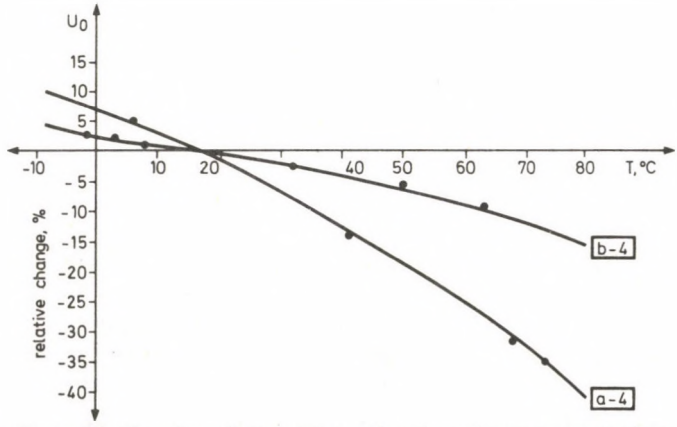


Fig. 15. Change of open circuit voltage in percent as a function of temperature in the case of photocells belonging to groups *a* and *b* measured without filters

circuit voltage can approximately be justified with correlation (19). On comparing figures 14 and 15, we may state that the extent of temperature dependency of the short circuit current and open circuit voltage are in concordance with each other.

The temperature dependency of the open circuit voltage was measured without colour filters.

4.3. Noise investigations

Noise made by the photocells was measured with the help of a LOCK-IN amplifier (Princeton Appl. Res. make, Model 124 A., type 184 pre–amplifier). We previously measured the noise current of the amplifier (I_{na}) with an input left empty and then took the complete noise current (I_{nd}) after connecting a photocell to the input. From the two values, the noise current of the photocells was determined with the help of

$$I_n = \sqrt{I_{nd}^2 - I_{na}^2}. \tag{30}$$

The measurements were performed at room temperature, with zero illumination on 10 Hz, 100 Hz, 1 kHz and 10 kHz centre frequencies. The equivalent noise bandwidth of the amplifier corresponded to 10% of the intermediate frequencies. The effective value of the noise current referring to a unit of bandwidth was determined from the measurements taken. It was expressed in $A \times Hz^{-1/2}$.

The responsivity of the photocells was measured with monochromatic light with reference to UV–444 calibrated silicon photocells. The half-bandwidth of the light wave emitted by the monochromator was about 1 nm. During our investigations we determined the wavelength corresponding to the maximum responsivity (λ_{max}) and the responsivity measurable at this wavelength in

$$A \cdot W^{-1} \cdot cm^2.$$

If we are familiar with the photo-sensitive surface of the photocells, the responsivity can be expressed in $A \cdot W^{-1}$ dimension.

Knowing the noise current and the responsivity and based on correlation (22), we calculated the NEP data characteristic of the threshold sensitivity of the photocells.

The results of the calculations and measurements were summarized in Table VIII. The value of the noise current measured at 1 kHz and the value of NEP calculated at 1 kHz and λ_{max} were listed in it.

The NEP values of the photocells belonging to group–*a* are about the same as those for group–*c*. However, the threshold sensitivity of the photocells belonging to group–*b* is worse by more than one in order of magnitude compared to them.

It is more advisable to compare the photocells made according to the same technology but with significantly differing photosensitive surfaces with the help of the

Table VIII

Spectral and threshold sensitivity properties of the inspected photocells

Sample number	λ_{max}, nm	Responsiveness at λ_{max}, $A \times W^{-1}$	Photosensitive sur face, mm^2	Noise current at 1 kHz $\times 10^{-13}$ $A \times Hz^{-1/2}$	NEP (1 kHz, 1 Hz λ_{max}), $\times 10^{-13}$ $W \times Hz^{-1/2}$
a – 1	870	0.637	7	0.092	0.14
a – 2	875	0.549	7	0.11	0.20
a – 3	860	0.546	7	0.092	0.17
b – 1	890	0.645	7	2.1	3.3
b – 2	890	0.603	7	12.0	20.0
b – 3	860	0.635	7	1.5	2.4
c – 1	880	0.421	3.8	0.097	0.23
c – 2	880	0.365	3.8	0.13	0.36
A – 1	895	0.507	100	13.0	26.0
A – 2	910	0.552	100	13.0	24.0

NEI data (Noise Equivalent Input) instead of with the NEP data as

$$\mathrm{NEI} \equiv \frac{\mathrm{NEP}}{\mathrm{A}}, \tag{31}$$

that is, NEI is equivalent to NEP referring to a unit detector surface [10]. As an example, let us compare samples marked a–2 and A–2,

$$\mathrm{NEI}_{(a-2)} = \frac{2 \times 10^{-14}}{0.07} \cong 2.8 \times 10^{-12} \quad W \times Hz^{-1/2} \times cm^{-2};$$

and

$$\mathrm{NEI}_{(A-2)} = \frac{2.4 \times 10^{-12}}{1.00} = 2.4 \times 10^{-12} \quad W \times Hz^{-1/2} \times cm^{-2}.$$

The agreement of the NEI data proves that the threshold sensitivity of the MEV made photocells with a large photosensitive surface falls into the expected order of magnitude.

The values of the noise current measured as a function of frequency did not show significant deviations. The smallest noise current arose generally at 100 Hz and 1 kHz. As an example, Table IX presents the results of noise measurements on some characteristic samples. Let us note that the accuracy of our noise measurements may be estimated at about 10%.

Table IX

Values of noise currents measured as function of centre frequencies

Frequency, Hz / Sample number	Noise current $\times 10^{-13}$ $A \times Hz^{-1/2}$			
	a – 2	*b* – 3	*c* – 2	*A* – 2
10	0.13	2.9	0.04	22.0
100	0.14	1.2	0.05	11.0
10^3	0.11	1.5	0.13	13.0
10^4	0.33	8.4	0.79	66.0

5. Conclusions

During our research we investigated the more exacting application opportunities of photocells in relation to manufacturing technology. We have been dealing with measuring temperature and illumination dependency of the internal impedance of photocells, with the temperature dependent and wavelength dependent properties of sensitivity and with investigations on noise and threshold sensitivity.

From the circuitry's point of view, the internal impedance of photocells is important when fitting it to the front amplifier. It is the significant temperature and illumination dependency of the internal impedance of the photodetectors operating in the photocell operating mode that constitutes its greatest disadvantage over the short circuited photodiodes. The rather significant illumination dependency of the ohm–member, the internal resistance, seems to be the most critical. From this aspect, the photocells belonging to group–*a* proved to be the best, for in their case, both in darkness and in an illuminated state, the internal resistance produced a relatively high value when measured.

When used in devices and applied amidst varied environmental temperatures, the temperature and wavelength dependency of the responsivity may be very important. From this respect, the behavior of the photocells belonging to groups *a* and *b* were quite different. In the case of photocells belonging to group–*b* made of raw materials with a small specific resistance, the temperature dependency of both the short circuit current and the open circuit voltage is of a much smaller value than that of samples belonging to group–*a*. Among these latter ones, we may find samples whose temperature factor of their short circuit current reached the 0.2%/K value. The temperature factor of the open circuit voltage never exceeded the 1%/K value with any of the samples.

We examined the temperature factor of the short circuit current as a function of the penetration depth and of the wavelength of the illuminating light as well. In agreement with theoretical considerations, we found that the temperature factor was smaller when applying red light than when using infrared illumination.

When measuring small light levels, we must also be familiar with the noise current and the threshold sensitivity of the photocells. The noise current was measured in the frequency range of 10 Hz to 10 kHz and the Noise Equivalent Power values were determined from the measured noise data and the $A \times W^{-1}$ sensitivity measured at the maximum responsivity wavelength. The photocells manufactured at home were compared with samples made by Telefunken.

We may state that the responsivity of the MEV made photocells is very good, it approaches the maximum, theoretically attainable value. At the same time there are significant differences between the various photocell groups from the point of view of noise and NEP as well.

Surveying our achievements, we may say that for precision measuring technical purposes it is the photocells belonging to group–*a* primarily that we would advise for

use. Its high internal resistance, the relatively small extent of illumination dependency of the internal resistance, its small noise current and great threshold sensitivity are the advantages that may be listed. The only real advantage the photocells belonging to group–*b* have is the weaker temperature dependency of the responsivity.

Photocells with a 100 sq.mm. photosensitive surface also satisfy the requirements of fine quality photodetectors, though we would mainly advise them for use as solar cells.

The applicability of home made photocells is greatly promoted by their favourable parameters.

References

1. Ambroziak, A.: Semiconductor Photoelectric Devices. Iliffe Books LTD, London, 1968.
2. Berkecz, J.–Kiss, T.: Fokozott érzékenységű fényelemek előállítása (Manufacturing photocells with higher sensitivity). Korszerű technológiák (1983) **1,** 13–21.
3. Gyárfás, A.: Az optoelektronikai eszközökkel megvalósítható jelzésátvitel néhány elméleti és gyakorlati kérdése (Some theoretical and practical issues of signal transmission feasible with optoelectronic devices) Híradástechnika **29,** (1978) 1–12.
4. Kiss, T.: Diffúziós folyamatok reprodukálhatóságának növelése (Increasing the reproductibility of diffusion processes). HIKI közlemények 15/3 (1975) 20–33.
5. Moss, T. S.: "Optical Properties of Semi-conductors. London Butterworths Sci. Pub. 1959.
6. Motchenbacher, C. D.–Fitchen, F. C.: Low-noise electronic design. John Wiley et Sons, Inc. New York Hungarian translation. Műszaki Könyvkiadó, Budapest 1977
7. Phillips, A. B.: Transistor Engineering. McGraw-Hill Book Co. New York. 1962.
8. Rivkin, S. M.: Poluprovodnyiki v Nauke i Tyekhnike (Semiconductors in Science and Technology). Soviet Union, Academy of Sciences Publishing House, Moscow–Leningrad, (1958), 463–515.
9. Sah, C. T.: The equivalent circuit model in solid-state devices. The simple energy level defect centers. Proc. IEEE **55.** (1967), 654–671.
10. Szentiday, K.: Félvezető fotodetektorok (Semiconducting Photodetectors). Műszaki Könyvkiadó, Budapest 1977.
11. Szentiday, K.: On the measurement of the internal resistance and linearity of photovoltaic cells. 9th International Symposium of the Technical Committee of Photon-detectors, Visegrád, Hungary (1980) 297–309.
12. Trutz, S.: HIPREM–1 egydimenziós technológia modellező program felhasználó dokumentációja (The users' documentation of the modelling program of HIPREM–1 one-dimensional technology). 1981 MEV (for home use)

Acta Technica Academiae Scientiarum Hungaricae, 98 (3—4), pp. 181—196 (1985)

LATERAL BUCKLING OF ELASTICALLY RESTRAINED ARCHES WITH BUILT-IN SUPPORTS

I. BÓDI*

[Received: 5 February 1985]

An iterative procedure is presented for the computation of the critical load of arches elastically restrained against lateral translation and rotation. The procedure is applicable to arches with arbitrary support systems.

1. Introduction

Several papers have dealt with the lateral buckling of centrally compressed circular arches [1], [2], [3].

The stability of cantilever arches was analyzed in [3], the critical load of arches with fork–like supports, elastically restrained against lateral translation was presented in [1] and the stability analysis of arches with fork–like supports, elastically restrained against lateral translation *and* rotation was carried out in [2].

The aim of this paper is to present a method for the computation of the critical load for arches with built–in supports, elastically restrained against lateral translation and rotation. It will be shown that the method is also applicable to arches with arbitrary lateral restraint and with arbitrary support systems (boundary conditions). For tent structures, the characteristics of the elastic restraint were determined in [4].

2. Assumptions and approximations

The assumptions and approximations are identical with those made in [2] and [3] so that we only present the most important ones:

The *material of the arch* is homogeneous, isotropic and linearly elastic.

The *cross section of the arch* is constant and has at least one axis of symmetry which lies in the plane of the arch.

The *arch* is subjected to a radially directed, uniformly distributed conservative *load system* in the plane of the cross section (Fig. 1a).

* I. Bódi, H-1052 Budapest, Petőfi S. u. 5, Hungary.

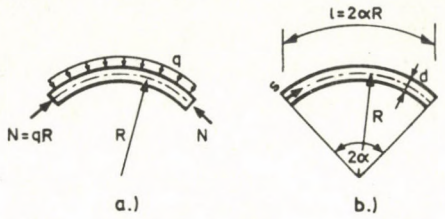


Fig. 1. Arch. with built–in supports: *1a* Loading; *1b* Geometrical characteristics of the arch

The *curvature of the arch* is constant and not too great (Fig. 1b).

It follows from this latter assumption that

a) the arch is only subjected to central compression which can be computed from the formula

$$N = qR;$$

b) even the greatest vertical dimension of the cross section can be neglected in comparison with the radius of the arch, i.e. the approximation

$$1 \pm \frac{d}{R} \approx 1$$

holds (Fig. 2b).

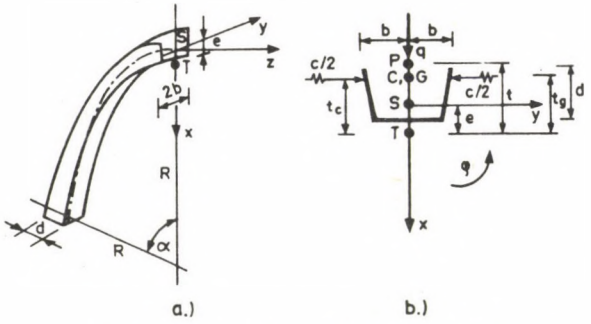


Fig. 2. Notations: *2a* Arc; *2b* Cross section

3. Notations

S	centroid of the cross section,
T	shear centre of the cross section,
P	point of application of the load,
C	point of application of the lateral elastic restraint,
G	point of application of the lateral shear restraint,
e	distance between T and S,
s	arc length measured along the centroidal line of the arch,

L length of the centroidal line of the arch,
t distance between P and T,
t_c distance between T and C,
t_g distance between T and G,
R radius of the arc measured at the centroidal line of the arch,
$v_T(s)$ function of the lateral translation (y) of the shear centre,
$\varphi(s)$ function of lateral rotation,
g constant of elastic shear restraint referred to unit length of the arc (kN),
c constant of elastic restraint referred to unit length of the arc (kN/m^2)
EI_x bending rigidity of the cross section of the arch,
GI_t torsional rigidity of the cross section of the arch,
EI_ω warping rigidity of the cross section of the arch,
N normal force,
i_x, i_y radiuses of gyration of the cross section.

4. Differential equations of lateral buckling

The differential equation system of the arch in the state of bifurcation of equilibrium derived on the basis of the assumptions and approximations given in Section *2* was presented in [2]. The two differential equations of the forth order with constant coefficients for lateral translation v_T and lateral rotation φ assume the form

$$GI_T\left(\frac{d^2\varphi}{ds^2}+\frac{1}{R}\frac{d^2v_T}{ds^2}\right)-EI_\omega\left(\frac{d^4\varphi}{ds^4}+\frac{1}{R}\frac{d^4v_T}{ds^4}\right)+$$

$$+eN\left(\frac{d^2v_T}{ds^2}-e\frac{d^2\varphi}{ds^2}\right)-\frac{EI_x}{R}\left(\frac{\varphi}{R}-\frac{d^2v_T}{ds^2}\right)+N\left(\frac{t-e}{R}-\frac{i_x^2}{R^2}\right)\varphi+$$

$$+(i_x^2+i_y^2)N\frac{d^2\varphi}{ds^2}+ct_c(v_T-t_c\varphi)-gt_g\left(\frac{d^2v_T}{ds^2}-t_g\frac{d^2\varphi}{ds^2}\right)=0 \tag{1}$$

$$EI_x\left(\frac{1}{R}\frac{d^2\varphi}{ds^2}-\frac{d^4v_T}{ds^4}\right)+\frac{GI_T}{R}\left(\frac{d^2\varphi}{ds^2}+\frac{1}{R}\frac{d^2v_T}{ds^2}\right)-\frac{EI_\omega}{R}\left(\frac{d^4\varphi}{ds^4}+\frac{1}{R}\frac{d^4v_T}{ds^4}\right)+$$

$$+N\left(e\frac{d^2\varphi}{ds^2}-\frac{d^2v_T}{ds^2}\right)-c(v_T-t_c\varphi)+g\left(\frac{d^2v_T}{ds^2}-t_g\frac{d^2\varphi}{ds^2}\right)=0 \tag{2}$$

with the boundary conditions

$$\varphi(s=0)=\varphi(s=L)=0, \tag{3)–(4)}$$

$$v_T(s=0)=v_T(s=L)=0, \tag{5)–(6)}$$

$$\frac{d\varphi}{ds}(s=0)=\frac{d\varphi}{ds}(s=L)=0, \tag{7)–(8)}$$

$$\frac{dv_T}{ds}(s=0)=\frac{dv_T}{ds}(s=L)=0 \tag{9)–(10)}$$

valid for arches with built–in ends.

We shall first determine the particular solutions which satisfy differential equations (1) and (2) then, from these particular solutions, we shall choose those which also satisfy boundary conditions (3)–(10). If there are more solutions satisfying both the system of differential equations and the boundary conditions, then the one which yields the smallest critical compressive force N_{CR} is considered as the solution to the problem.

5. Solution of the system of differential equations

For the solution of the system of differential equations, we shall rely on the method presented in [5].

Let us rearrange the system of differential equations by introducing the differential operators

$$\mathscr{L}_{11}(\ldots)=-EI_\omega\frac{\mathrm{d}^4(\ldots)}{\mathrm{d}s^4}+[GI_T-N(e^2+i_x^2+i_y^2)+gt_g^2]\frac{\mathrm{d}^2(\ldots)}{\mathrm{d}s^2}+$$

$$+\left[N\left(\frac{t-e}{R}-\frac{i_x^2}{R^2}\right)-ct_c^2-\frac{EI_x}{R^2}\right](\ldots), \tag{11}$$

$$\mathscr{L}_{12}(\ldots)=\mathscr{L}_{21}(\ldots)=-\frac{EI_\omega}{R}\frac{\mathrm{d}^4(\ldots)}{\mathrm{d}s^4}+$$

$$+\left[\frac{GI_T}{R}+\frac{EI_x}{R}+Ne-gt_g\right]\frac{\mathrm{d}^2(\ldots)}{\mathrm{d}s^2}+ct_c(\ldots), \tag{12}$$

$$\mathscr{L}_{22}(\ldots)=-\left[EI_x+\frac{EI_\omega}{R^2}\right]\frac{\mathrm{d}^4(\ldots)}{\mathrm{d}s^4}+$$

$$+\left[\frac{GI_T}{R^2}-N+g\right]\frac{\mathrm{d}^2(\ldots)}{\mathrm{d}s^2}-c(\ldots). \tag{13}$$

By doing so, we obtain the system of differential equations as

$$\begin{bmatrix}\mathscr{L}_{11} & \mathscr{L}_{12}\\ \mathscr{L}_{12} & \mathscr{L}_{22}\end{bmatrix}\begin{bmatrix}\varphi\\ v_T\end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix}. \tag{14}$$

Let us determine the determinant of operator matrix **L** which is also a linear operator:

$$\mathscr{D}\doteq\det\mathbf{L}=\mathscr{L}_{11}\mathscr{L}_{22}-\mathscr{L}_{12}^2. \tag{15}$$

In performing the operations, after some rearrangement, we arrive at the symbolic sum for operator $\mathscr{D}$:

$$\mathscr{D}(\ldots)=a_1\frac{\mathrm{d}^8(\ldots)}{\mathrm{d}s^8}+[a_2+Nb_1]\frac{\mathrm{d}^6(\ldots)}{\mathrm{d}s^6}+[a_3+Nb_2+N^2c_1]\frac{\mathrm{d}^4(\ldots)}{\mathrm{d}s^4}+$$

$$+[a_4+Nb_3+N^2c_2]\frac{\mathrm{d}^2(\ldots)}{\mathrm{d}s^2}+[a_5+b_4](\ldots), \tag{16}$$

where we have

N compressive force in the arch

$$a_1=EI_\omega EI_x, \tag{17}$$

$$a_2=-\left[gEI_\omega+EI_x(GI_T+gt_g^2)-\frac{2EI_\omega EI_x}{R^2}\right], \tag{18}$$

$$a_3=gGI_T+cEI_\omega+EI_x\left(\frac{EI_\omega}{R^4}+ct_c^2-\frac{2GI_T}{R^2}+\frac{2gt_g}{R}\right), \tag{19}$$

$$a_4=-\left[cg(t_g+t_c)^2+cGI_T+EI_x\left(\frac{GI_T}{R^4}+\frac{g}{R^2}+\frac{2ct_c}{R}\right)\right], \tag{20}$$

$$a_5=c\frac{EJ_x}{R^2}, \tag{21}$$

$$b_1=\left(EI_x+\frac{EI_\omega}{R^2}\right)(e^2+i_x^2+i_y^2)+EI_\omega, \tag{22}$$

$$b_2=-\left\{\frac{GI_T}{R^2}(e^2+i_x^2+i_y^2)+\left(EI_x+\frac{EI_\omega}{R^2}\right)\left(\frac{t-e}{R}-\frac{i_x^2}{R^2}\right)+\right.$$

$$\left.+g[i_x^2+i_y^2+(e-t_g)^2]+GI_T\right\}, \tag{23}$$

$$b_3=\frac{EI_x}{R^2}+\left(\frac{t-e}{R}-\frac{i_x^2}{R^2}\right)\left(\frac{GI_T}{R^2}+g\right)+c[i_x^2+i_y^2+(e-t_c)^2], \tag{24}$$

$$b_4=-c\left(\frac{t-e}{R}-\frac{i_x^2}{R^2}\right), \tag{25}$$

$$c_1=i_x^2+i_y^2, \tag{26}$$

$$c_2=-\left(\frac{t-e}{R}-\frac{i_x^2}{R^2}\right). \tag{27}$$

Let us now try to find a scalar function $H=f(s)$ which satisfies the condition

$$\mathscr{D}(H)=\det \mathbf{L}(H)=0, \tag{28}$$

i.e. if the determinant of operator matrix **L** vanishes, then function H also satisfies the matrix equation

$$\mathbf{L} \cdot \operatorname{adj} \mathbf{L}(H) = 0$$

as is shown in [5].

Having determined scalar function $H(s)$, we substitute it to a row of the matrixarithmetic adjoint matrix of the operator matrix in (14). By doing so, we arrive at displacement functions $\varphi(s)$ and $v_T(s)$:

$$v_T(s) = \mathscr{L}_{11}[H(s)], \tag{29}$$

$$\varphi(s) = -\mathscr{L}_{12}[H(s)]. \tag{30}$$

Function H satisfying condition (28) can be determined by making use of expression (16) which requires the solution of the ordinary, linear homogeneous differential equation of the eighth order with constant coefficients:

$$a_1 \frac{\mathrm{d}^8 H}{\mathrm{d}s^8} + (a_2 + Nb_1) \frac{\mathrm{d}^6 H}{\mathrm{d}s^6} + (a_3 + Nb_2 + Nc_1^2) \frac{\mathrm{d}^4 H}{\mathrm{d}s^4} +$$

$$+ (a_4 + Nb_3 + Nc_2^2) \frac{\mathrm{d}^2 H}{\mathrm{d}s^2} + (a_5 + b_4) H = 0. \tag{31}$$

We are looking for the solution in the form

$$H = e^{\lambda s}$$

with which the characteristic equation of the differential equation assumes the form

$$a_1 \lambda^8 + (a_2 + Nb_1) \lambda^6 + (a_3 + Nb_2 + N^2 c_1) \lambda^4 +$$

$$+ (a_4 + Nb_3 + N^2 c_2) \lambda^2 + a_5 + b_4 = 0. \tag{32}$$

In determining roots $\lambda_1, \lambda_2, \ldots, \lambda_8$, of the algebraic equation (32), two cases must be considered.

Case 1.

If all roots are single—and this is the usual case—, then the general solution to differential equation (31) assumes the form

$$H(s) = C_1 e^{\lambda_1 s} + C_2 e^{\lambda_2 s} + \ldots + C_8 e^{\lambda_8 s}, \tag{33}$$

or in a more concise form

$$H(s) = \sum_{n=1}^{8} C_n e^{\lambda_n s}, \tag{34}$$

where $C_1, C_2, \ldots, C_8$ are arbitrary constants. By making use of operators (11) and

(12), we obtain the functions

$$v_T(s)=\mathscr{L}_{11}H(s)=\sum_{n=1}^{8} C_n e^{\lambda_n s}\cdot A(\lambda_n, N) \tag{35}$$

and

$$\varphi(s)=-\mathscr{L}_{12}H(s)=\sum_{n=1}^{8} C_n e^{\lambda_n s}\cdot B(\lambda_n, N), \tag{36}$$

where we have

$$A(\lambda_n, N)=-EI_\omega\lambda_n^4+[GI_T-N(e^2+i_x^2+i_y^2)+gt_g^2]\lambda_n^2+$$
$$+N\left(\frac{t-e}{R}-\frac{i_x^2}{R^2}\right)-ct_c^2-\frac{EI_x}{R^2}; \tag{37}$$

$$B(\lambda_n, N)=\frac{EI_\omega}{R}\lambda_n^4-\left[\frac{GI_T+EI_x}{R}+Ne-gt_g\right]\lambda_n^2-ct_c. \tag{38}$$

Case 2.

Multiple roots only emerge in special cases. If λ_i is a k_i-fold root, the general solution assumes the form

$$H(s)=(C_{10}+C_{11}s+\ldots+C_{1k_1-1}s^{k_1-1})e^{\lambda_1 s}+$$
$$+(C_{20}+C_{21}s+\ldots+C_{2k_2-1}s^{k_2-1})e^{\lambda_2 s}+\ldots$$
$$+(C_{m0}+C_{m1}s+\ldots C_{mk_m-1}s^{k_m-1})e^{\lambda_m s},$$

where we have

$$k_1\lambda_1+k_2\lambda_2+\ldots k_m\lambda_m=8,$$

and

$$C_{10}, C_{11}\ldots C_{1k_1-1} \qquad \text{ill.} \qquad C_{10}, C_{11}\ldots C_{mk_1-1}$$

are arbitrary constants.

The general solution in concise form reads

$$H(s)=\sum_{n=1}^{m} e^{\lambda_n s}\cdot\sum_{j=0}^{k_n-1} C_{nj}\cdot s^j,$$

and the displacement functions emerge as

$$v_T(s)=\mathscr{L}_{11}\left[\sum_{h=1}^{m} e^{\lambda_n s}\cdot\sum_{j=0}^{k_n-1} C_{nj}s^j\right],$$

$$\varphi(s)=-\mathscr{L}_{12}\left[\sum_{n=1}^{m} e^{\lambda_n s}\cdot\sum_{j=0}^{k_n-1} C_{nj}s^j\right],$$

where $\mathscr{L}_{11}$ and $\mathscr{L}_{12}$ are the linear differential operators defined by (11) and (12).

6. Boundary conditions

We shall only discuss the case of single roots in detail.

The boundary conditions (3)–(10) of the arch with built-in ends are obtained from formulae (35) and (36) by derivation:

$$v_T(s=0)=\sum_{n=1}^{8} C_n \cdot A(\lambda_n, N)=0, \tag{39}$$

$$v_T(s=L)=\sum_{n=1}^{8} C_n e^{\lambda_n L} A(\lambda_n, N)=0, \tag{40}$$

$$\frac{\mathrm{d}v_T}{\mathrm{d}s}(s=0)=\sum_{n=1}^{8} C_n \lambda_n A(\lambda_n, N)=0, \tag{41}$$

$$\frac{\mathrm{d}v_T}{\mathrm{d}s}(s=L)=\sum_{n=1}^{8} C_n \lambda_n e^{\lambda_n L} A(\lambda_n, N)=0, \tag{42}$$

$$\varphi(s=0)=\sum_{n=1}^{8} C_n B(\lambda_n, N)=0, \tag{43}$$

$$\varphi(s=L)=\sum_{n=1}^{8} C_n e^{\lambda_n L} B(\lambda_n, N)=0, \tag{44}$$

$$\frac{\mathrm{d}\varphi}{\mathrm{d}s}(s=0)=\sum_{n=1}^{8} C_n \lambda_n B(\lambda_n, N)=0, \tag{45}$$

$$\frac{\mathrm{d}\varphi}{\mathrm{d}s}(s=L)=\sum_{n=1}^{8} C_n \lambda_n e^{\lambda_n L} B(\lambda_n, N)=0. \tag{46}$$

The above equations for the boundary conditions can be transformed into the system of linear, homogeneous equations of the eighth order based on the undetermined coefficients C_n:

$$\begin{bmatrix} A(\lambda_1,N) & A(\lambda_2,N) & \dots & A(\lambda_8,N) \\ e^{\lambda_1 L}A(\lambda_1,N) & e^{\lambda_2 L}A(\lambda_2,N) & \dots & e^{\lambda_8 L}A(\lambda_8,N) \\ \lambda_1 A(\lambda_1,N) & \lambda_2 A(\lambda_2,N) & \dots & \lambda_8 A(\lambda_8,N) \\ \lambda_1 e^{\lambda_1 L}A(\lambda_1,N) & \lambda_2 e^{\lambda_2 L}A(\lambda_2,N) & \dots & \lambda_8 e^{\lambda_8 L}A(\lambda_8,N) \\ B(\lambda_1,N) & B(\lambda_2,N) & \dots & B(\lambda_8,N) \\ e^{\lambda_1 L}B(\lambda_1,N) & e^{\lambda_2 L}B(\lambda_2,N) & \dots & e^{\lambda_8 L}B(\lambda_8,N) \\ \lambda_1 B(\lambda_1,N) & \lambda_2 B(\lambda_2,N) & \dots & \lambda_8 B(\lambda_8,N) \\ \lambda_1 e^{\lambda_1 L}B(\lambda_1,N) & \lambda_2 e^{\lambda_2 L}B(\lambda_2,N) & \dots & \lambda_8 e^{\lambda_8 L}B(\lambda_8,N) \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ C_8 \end{bmatrix} = 0, \tag{47}$$

or in concise form

$$\mathbf{K} \cdot \mathbf{C} = \mathbf{0} . \tag{48}$$

Apart from the given geometrical and rigidity characteristics of the arch, the elements of coefficient matrix $\mathbf{K}$ only depend on compressive force N—formula (29) shows that roots $\lambda_1, \lambda_2, \ldots, \lambda_8$ are also functions of N.

Hence the equation

$$\mathbf{K}(N) \cdot \mathbf{C} = 0 \tag{49}$$

holds. Solutions different from the trivial solutions can be obtained only if the determinant of the system of the equation vanishes, i.e. if

$$\det \mathbf{K}(N) = 0 \tag{50}$$

holds. Every critical compressive force $N_{CR}^{(k)}$ belonging to buckling half wave k satisfies condition (50) and the minimum of this series yields the solution to the problem:

$$N_{CR} = \min [N_{CR}^{(k)}]; \qquad k = 1, 2, 3 \ldots \tag{51}$$

7. Solution process

The solution of the problem is obtained by iteration for which we need the approximate value of the critical compressive force $\tilde{N}_{CR}$ causing lateral buckling:

$$\tilde{N}_{CR} = \min [\tilde{N}_{CR}^{(k)}] ,$$

where

$$\tilde{N}_{CR}^{(k)} = N_{CR,0}^{(k)} + g + \frac{L^2 c}{k^2 \pi^2} . \tag{52}$$

In this formula $N_{CR,0}^{(k)}$ represents the critical compressive force which belongs to buckling half wave k of an arch with built-in ends but without elastic restraint. According to [3], its value is given by the formula

$$N_{CR,0}^{(k)} = \frac{EI_x}{R^2} \cdot \frac{\left(k^2 - \frac{\alpha^2}{\pi^2}\right)^2}{\frac{\alpha^2}{\pi^2}\left(k^2 + \frac{EI_x}{GI_T}\frac{\alpha^2}{\pi^2}\right)} , \tag{53}$$

where α stands for the half central angle of the arc.

According to the example of a beam with built-in ends and with elastic restraint presented in [6], formula (52) gives a good approximation for critical compressive force $N_{CR}^{(k)}$.

The derivation in [8] for beams with a straight axis shows that, for $c = 0$, we obtain the exact value.

By making use of formulae (17)–(27) and (32), first we compute the roots of the characteristic equation, then we determine the general solution to the characteristic differential equation by formula (33) and the displacement functions and their derivatives by formulae (35)–(46). Finally, we obtain the coefficient matrix for the boundary conditions and its determinant from formula (47). If the determinant vanishes, our approximation at the beginning of the process was correct, i.e. we used the exact value of the critical force. If the determinant does not vanish, we repeat the process by using the reduced value.

$$\tilde{N}_{CR}^{i+1} = \tilde{N}_{CR}^{i} - \Delta N \tag{54}$$

of the critical force. The process has to be repeated until determinant (47) vanishes or changes sign.

In the latter case the exact value of the critical force is obtained by the principle of halving intervals.

8. Application to different boundary conditions

It is advantageous with the method presented in the foregoing that, contrary to common methods, it is not necessary to choose a system of basis functions which satisfies the boundary conditions since, knowing characteristic roots, we can directly determine the form of the functions representing the lateral displacements.

In this way, apart from the fact that the value of N_{CR} has to be chosen at the beginning, the effect of the boundary conditions only appears at end of the process in constructing the boundary condition matrix and in the value of its determinant. It follows that other types of boundary conditions can easily be analized; it is only the boundary condition matrix which has to be modified according to the conditions related to the lateral displacements at the supports. Accordingly, we can easily carry out the lateral stability analysis of arches with different support systems; fork–like supports at both ends, a fork–like support and a built–in end, two elastic supports or even diafragms at arbitrary points.

In the case of more complicated support systems when it is more difficult to give a good approximation for the critical compressive force we may need corresponding values of the compressive force and the determinant. In the neighbourhood of the smallest root we have to use smaller increments for the step–by–step iteration to obtain the exact value of the critical compressive force. We mention here that we computed the critical force of the arch with fork–like supports presented in [2] by using the above process and arrived at the same value as in [2].

9. The role of the warping rigidity

The formulae presented above for the lateral stability analysis of arches, especially formulae (37) and (38) show that the warping rigidity and the St. Venant torsional rigidity can always be added to produce a "resultant torsional rigidity" as

$$D_{Tk}=GI_T+\lambda_k^2 EI_\omega,$$

where λ_k is the k-th characteristic root.

This relationship can immediately recognized in the case of arches with fork–like supports at both ends since the system of eigenfunctions is very simple in their case [2]:

$$U_i(s)=\sin k\frac{\pi s}{L}; \qquad k=1,2,3\ldots$$

The "resultant torsional rigidity" in this case assumes the form

$$D_{Tk}=GI_T+k^2\frac{\pi^2}{L^2}EI_\omega.$$

The value of the first term is generally much greater than that of the second one and consequently the latter term is often neglected, i.e. it is assumed that $EI_\omega=0$ holds. This approximation is only acceptable for cross sections whose primary warping function with respect to the middle line of the cross section vanishes and the warping rigidity only comprises the secondary warping functions belonging to points outside the middle line.

In neglecting the warping rigidity, coefficient (17) of the eighth order term of characteristic equation (32) vanishes reducing the order of the equation and the number of the linearly independent solutions to characteristic differential equation (16) to six. Consequently, displacement functions $v_T(s)$ and $\varphi(s)$ only consist of six terms, respectively. The neglect of EI_ω also results in the loss of the two boundary conditions related to warping so that the coefficient matrix of the boundary conditions is now of 6×6 dimensions making the solution of the problem easier.

10. Remarks and suggestions for the numerical analysis

To help to solve some numerical problems emerging through the procedure, we shall now give some simple programming tricks and methods.

As we have seen, during the procedure we have to determine the determinant of boundary condition matrix $\mathbf{K}(N)$ defined by (47). Since matrix $\mathbf{K}$ is nearly singular even in the neighbourhood of critical force N_{CR}, mainly to avoid rounding errors, it is expedient to compute the value of the determinant by using a programming language (e.g. FORTRAN) in which variables with double accuracy can be defined.

It is also expedient to factor out appropriate quantities of the columns of the determinant so that the remaining elements be in the same order of magnitude. We have to be careful because these reduced values must still remain in the input range of the computer.

Greater accuracy can also be achieved by setting the origin of the arc length in the middle of the arch instead of at the built–in end because in this way the values of the hyperbolic functions in the boundary conditions are in the same order of magnitude.

The singularity of the boundary condition matrix transformed according to the foregoing can be examined by applying the QR decomposition.

As is known, any real matrix, and matrix $\mathbf{K}$ is real, can be decomposed as

$$\mathbf{K}=\mathbf{Q}\times\mathbf{R},$$

where $\mathbf{Q}$ is a uniterian matrix (i.e. $\mathbf{Q}\times\mathbf{Q}'=\mathbf{I}$)
and $\mathbf{R}$ is an upper triangular matrix whose main diagonal only contains non–negative elements.

This decomposition is numerically stable and mathematically equivalent to the Gram–Schmidt orthogonalization process applied to the linearly independent columns of matrix $\mathbf{K}$.

In the case of singular matrix $\mathbf{K}$, being $\mathbf{Q}$ a uniterian matrix, matrix $\mathbf{R}$ must be singular, i.e. the equation

$$\det\mathbf{R}=r_{11}\cdot r_{22}\ldots r_{88}=0$$

must hold. It follows that one of the elements of the main diagonal of matrix $\mathbf{R}$ must vanish.

In summary, we can conclude that, using simple transformations, first we have to produce a matrix with elements of the same order of magnitude but within the input range of the computer. Then, by applying the QR decomposition, we have to compute the value of the boundary condition matrix in a numerically stable way.

11. Numerical example

Let us compute the the critical compressive force of the arch with built-in supports shown in Fig. 3. (The critical force of an arch with the same characteristics but with fork-like supports is given in Fig. 2.)

The geometrical and rigidity characteristics of the arch are

$$R=9.30\text{ m},$$

$$L=24.48\text{ m},$$

$$E=10^7\text{ kN/m}^2,$$

$$G\approx 0.4E=4\times 10^6\text{ kN/m}^2,$$

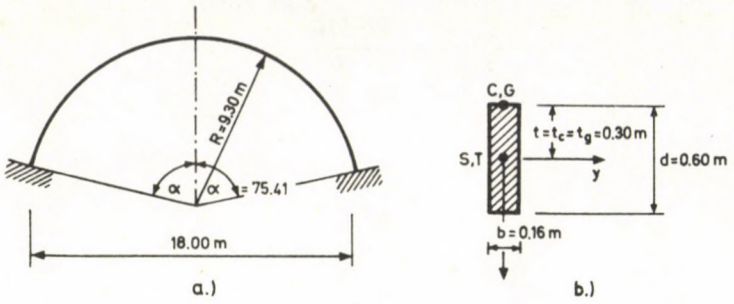


Fig. 3. Arch for the numerical example: *3a* Elevation, *3b* Cross section

$$EI_x = 2048 \text{ kNm}^2,$$

$$GI_T = 2726.5 \text{ kNm}^2,$$

$$EI_\omega = 61.4 \text{ kNm}^4,$$

$$t = t_c = t_g = 0.30 \text{ m},$$

$$e = 0,$$

$$i_x^2 \approx 0; \qquad i_y^2 = 0.030 \text{ m}^2.$$

The specific characteristics of the elastic restraint referred to unit length of the arc are

$$c = 4.54 \text{ kN/m}^2,$$

$$g = 216 \text{ kN}.$$

The coefficients of characteristic equation (32) come from formulae (17)–(27) as

$$a_1 = \quad 1.2575 \times 10^5, \qquad b_1 = \quad 1.2286 \times 10^2, \qquad c_1 = 3.0 \times 10^{-2},$$

$$a_2 = -5.6340 \times 10^6, \qquad b_2 = -2.8194 \times 10^3, \qquad c_2 = 3.23 \times 10^{-2}.$$

$$a_3 = \quad 4.8947 \times 10^5, \qquad b_3 = \quad 3.2208 \times 10^1,$$

$$a_4 = -1.9192 \times 10^4, \qquad b_4 = -1.465 \times 10^{-1},$$

$$a_5 = \quad 1.0750 \times 10^2,$$

The upper bound of the critical compressive force for $k = 1$ is obtained from (53):

$$\tilde{N}_{CR} \approx 81.01 + 216 + 275.67 = 572.7 \text{ kN}.$$

The values of determinant **K** of the boundary condition matrix belonging to compressive force N in the iteration steps are compiled in Table 1.

The characteristic equation after the ninth, final step assumes the form

$$1.2583 \times 10^5 \lambda^8 - 5.5689 \times 10^6 \lambda^6 - 9.2639 \times 10^5 \lambda^4 -$$

$$- 1.1615 \times 10^4 \lambda^2 + 3.1573 \times 10^1 = 0,$$

Table I

Number of steps	N	det **K**
1.	538.4	−1088.9
2.	452.2	1667.3
3.	495.3	1053.9
4.	516.8	104.12
5.	527.6	−474.61
6.	522.2	−179.11
8.	518.2	34.636
9.	518.9	−0.4518

and its roots take on the values

$$\lambda_{1-2} = \pm 0.127\,361i,$$

$$\lambda_{3-4} = \pm 0.389\,679i,$$

$$\lambda_{5-6} = \pm 6.665\,058,$$

$$\lambda_{7-8} = \pm 0.047\,886.$$

The complex pairs of roots only have an imaginary part and the application of the Euler formula for the exponential function yields scalar function (33) as

$$H(s) = C_1 \sin(0.127\,361s) + C_2 \cos(0.127\,361s) +$$
$$+ C_3 \sin(0.389\,679s) + C_4 \cos(0.387\,679s) +$$
$$+ C_5 \operatorname{sh}(6.665\,058s) + C_6 \operatorname{ch}(6.665\,058s) +$$
$$+ C_7 \operatorname{sh}(0.047\,886s) + C_8 \operatorname{sh}(0.047\,886s).$$

The functions of lateral displacements can be determined from formulae (35) and (36). Taking into consideration the remarks and suggestions made in Section 10, we obtain the boundary condition matrix from (39)–(46). The elements of the 8×8 matrix are compiled in Table II. After the ninth step the determinant assumes the value

$$\det \mathbf{K} = -0.4518 \approx 0$$

so we consider the compressive force obtained in the ninth step as the critical compressive force:

$$N_{CR} = 518.9 \text{ kN}.$$

Table II

The elements of the 8×8 matrix

.32388E+02	−.29424E+01	.10612E+08	−.96409E+06	−.32388E+02	−.29424E+01	−.10612E+08	−.96409E+06
−.61445E+02	.15510E+01	−.20133E+08	.50818E+06	−.61445E+02	−.15510E+01	−.20133E+08	−.50818E+06
−.71872E+02	−.16059E+01	−.20188E+08	−.44909E+06	.71872E+02	−.16059E+01	.20100E+08	−.44909E+06
−.41218E+01	.28007E+02	−.11525E+07	.78324E+07	−.41210E+01	−.28007E+02	−.11525E+07	−.78324E+07
.61576E+02	−.93136E−01	.20131E+08	−.30449E+05	−.61576E+02	−.93136E−01	−.20131E+08	−.30449E+05
−.73126E+00	−.78425E+01	−.23907E+06	−.25640E+07	−.73126E+00	.76425E+01	−.23907E+06	.25640E+07
.88538E+06	−.59011E+07	.14747E+08	−.98290E+08	−.88538E+06	−.59011E+07	−.14747E+08	−.98290E+08
−.88538E+06	.59011E+07	−.14747E+08	.98290E+08	−.88538E+06	−.59011E+07	−.14747E+08	−.98290E+08

References

1. Kollár, L.–Gyurkó, I.: Lateral buckling of elastically supported arches. Acta Technica Hung. **94** (1982), 37–45.
2. Kollár, L.–Bódi, .: Lateral buckling of arches with fork–like supports, elastically restrained along their entire length against lateral displacement and rotation. Acta Technica Hung. **95** (1982), 99–106.
3. Timoshenko, S. P.–Gere, J. M.: Theory of Elastic Stability. McGraw Hill, New York, 1961.
4. Kollár, L.: The supporting effect of the fabric of tent structures streched onto an arch row on the lateral stability of the arches. Acta Technica Hung. **94** (1982), 197–214.
5. Hegedüs, I.: Influence function of skew circular rings on elastic bedding. Periodica Polytechnica Civil Engineering. **25** (1981), 29–45.
6. Pflüger, A.: Stabilitätsprobleme der Elastostatik. 2nd Ed. Springer-Verlag Berlin–Heidelberg–New York, p 364.
7. Zurmühl, R.: Matrizen und ihre technischen Anwendungen 3. Aufl. Springer-Verlag. Berlin–Göttingen–Heidelberg 1961.
8. Csonka, P.: Buckling of Bars Elastically Built-in along their Entire Length. Acta Technica Hung. **32** (1961) 424–427.

Acta Technica Academiae Scientiarum Hungaricae, 98 (3—4), pp. 197—204 (1985)

SHEAR DESIGN OF REINFORCED AND PRESTRESSED CONCRETE ELEMENTS BY THE NEW CANADIAN CODE

M. P. COLLINS*–P. LENKEI**

[Received: October 1984]

The shear design of reinforced and prestressed concrete members using the "compression field theory" has been incorporated in the 1984 new Canadian code. The article introduces the theory and the method of design. A comparison is made between the new Canadian code and other international and national codes, including the existing and the past Hungarian codes.

Symbols

b — width of the member
b_v — width of sheared part of the member
d — depth of the member to the centroid of compression reinforcement
d_v — depth of sheared part of the member
f_2 — principal compressive stress of concrete
$f_{2,\max}$ — possible maximum of the principal compressive stresses
f'_c —concrete cylinder compressive strength
f_y — yield stress of steel
s — spacing of shear reinforcement
v — external shear stress

A_v — area of shear reinforcement
C — compressive load acting on the compressed part
D — diagonale compressive force
E_s — Young's modulus for steel
M — applied moment
M_f — factored applied moment
N_f — factored axial force
N_v — axial tensile force due to shear (equivalent axial force)
T — tensile load acting on the tensioned part
V — shear force
V_f — factored shear force

ε — normal strain
ε_1 — principal tensile strain
ε_2 — principal compressive strain
ε_t — tensile strain in transverse steel direction
ε_x — tensile strain in x-longitudinal direction
θ — angle of inclination of principal compressive stress
λ — to account for low density concrete ($\lambda = 1.00$ for normal density concrete)
Φ_c — resistance factor for concrete ($\Phi_c = 0.6$)
Φ_s — resistance factor for steel

* M. P. Collins, Department of Civil Engineering, University of Toronto, Toronto, Ontario, Canada
** P. Lenkei, H-1119 Budapest, Szakasits Á. u. 4, Hungary

1. Introduction

There was—and actually is—a wide and comprehensive discussion in North America about the improvement of the shear design of reinforced and prestressed concrete elements. An important contribution to this process is the shear chapter of the new Canadian Code "Concrete Structures for Buildings" CSA-A23.3-84, which was approved in 1984.

This code, besides giving a simplified method which is essentially the well known and traditional ACI $V = V_c + V_s$ procedure (ACI 318-83), introduces a new general method. The general method, called the *Compression Field Theory*, is based upon 15 years of extensive research, carried out at the University of Toronto [Mitchell and Collins (1974), Collins (1978), Collins and Mitchell (1980), Vecchio and Collins (1982)], and also uses concepts from the plasticity models [Thuerlimann et al (1982), Marti (1984) and (1985), Mueller (1978)].

The *compression field theory* considers shear as influencing the design of both the transverse and the longitudinal reinforcement. In its most general form the method permits the resistance and behaviour of members in shear to be investigated in detail by performing a sectional analysis which considers the equilibrium, compatibility and stress-strain requirements for different portions of the section. Such an analysis (see Fig. 1) would show that the shear stress distribution is not uniform, that the direction of

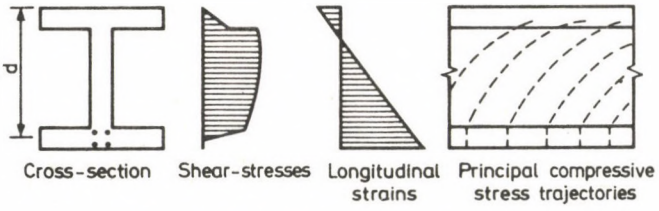


Fig. 1. Detailed analysis of a beam in shear

principal stresses changes over the depth of the beam and that tensile stresses in the concrete between the cracks contribute to the shear resistance of the member.

In lieu of the detailed analysis outlined above, the Code permits a more direct procedure, which concentrates on the conditions at mid-depth of the beam. In this procedure (see Fig. 2) the shear stresses are assumed to be uniformly distributed over an

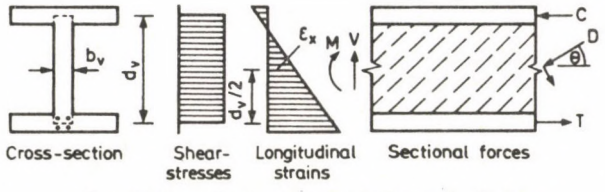


Fig. 2. More direct analysis of a beam in shear

area b_v wide and d_v deep, the direction of the principal compressive stresses (defined by angle Θ) is assumed to remain constant over d_v and tensile stresses in cracked concrete are ignored.

2. Stresses and Strains at Mid-Depth of the Beam

The cross-sectional dimensions of the member must be sufficiently large to ensure that the diagonally cracked concrete is capable of resisting the imposed inclined compressive stresses (i.e., $f_2 < f_{2\max}$).

If the principal tensile stress in the concrete is zero, then the principal compressive stress in the concrete, f_2, will be related to the shear stress on the concrete, $v = V_f/(b_v d_v)$, by the following equilibrium equation which can be derived from the Mohr's circle in Fig. 3.

$$f_2 = \left(\tan\Theta + \frac{1}{\tan\Theta}\right)\left(\frac{V_f}{b_v d_v}\right) \tag{1}$$

If the concrete at mid-depth is severely deformed (large principal tensile strain ε_1) its ability to resist compressive stresses will be substantially reduced. In the Code the failure value of f_2 is related to ε_1 by the following:

$$\frac{f_{2\max}}{\lambda \Phi_c f'_c} = \frac{1}{0.8 + 170\,\varepsilon_1} \leq 1.0 \tag{2}$$

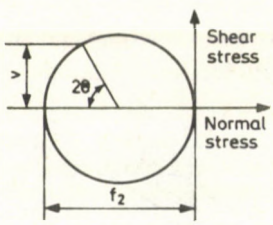


Fig. 3. Concrete stresses at mid-depth of the beam in web

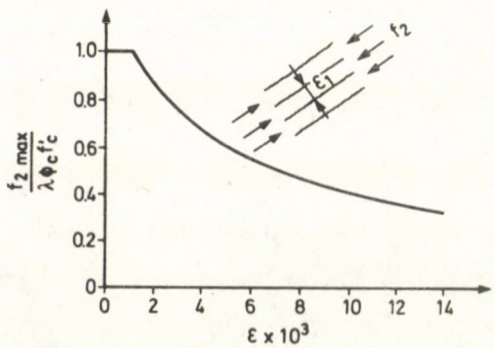


Fig. 4. Relating $f_{2\max}$ to ε_1

This expression was derived from the experimental data obtained from testing panels in pure shear, [Vecchio and Collins (1982)].

The principal tensile strain, ε_1, is related to the longitudinal strain at mid-depth, ε_x, the principal compressive strain (assumed to be -0.002) and the principal strain direction (assumed to coincide with the principal stress direction) by the following compatibility equation which can be derived from the Mohr's circle in Fig. 5.

$$\varepsilon_1 = \varepsilon_x + (\varepsilon_x + 0.002)/\tan^2 \Theta \tag{3}$$

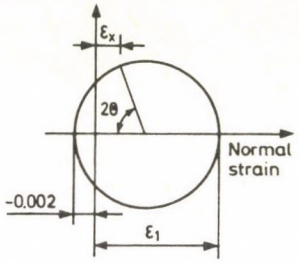


Fig. 5. Strain at mid-depth of the beam

3. Design of Reinforcement

Transverse reinforcement must be provided to equilibrate the outwards thrust of the diagonal compressive stresses in the concrete. The free body diagram in Fig. 6(a) demonstrates that, for uniformly loaded beams, the transverse reinforcement within

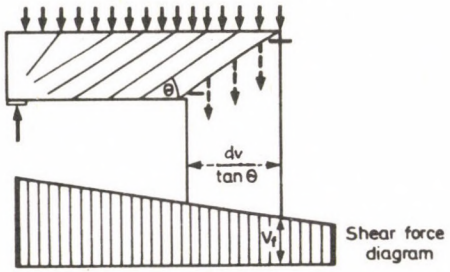


Fig. 6/a. The "staggering concept" for design of transverse shear reinforcement

the length of $d_v/\tan \Theta$ may be designed to resist the lowest shear within this length. This has become known as the "staggering concept" for shear design.

$$\frac{A_v \Phi_s f_y}{s} \frac{d_v}{\tan \Theta} \geq V_f \tag{4}$$

Fig. 6(b) illustrates the way in which the distribution of transverse reinforcement is determined from the shear envelope. Each step of the required resistance diagram results in a zone of equally spaced stirrups.

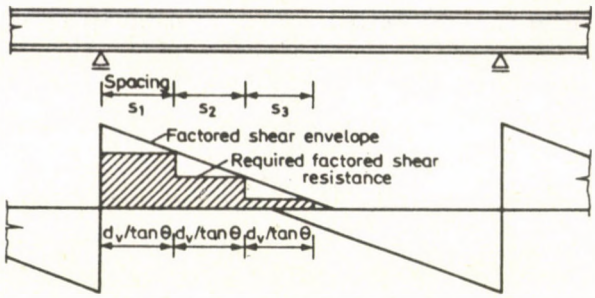


Fig. 6/b. The design of transverse steel

The shear force on the section is resisted by diagonal stresses in the concrete. Thus in Fig. 2 it is the vertical component of force D which is carrying the shear. The horizontal component of force D is equivalent to an axial compression on the concrete of $V/\tan\Theta$. This unwanted compression needs to be cancelled out by tensile forces in the longitudinal reinforcement (see Fig. 7). Thus shear

$$N_v = \frac{V}{\tan\Theta} \tag{5}$$

causes compressive stresses in the concrete and tensile stresses in the longitudinal reinforcement. In terms of the tension in the longitudinal reinforcement the shear is equivalent to an axial tensile load of $V/\tan\Theta$.

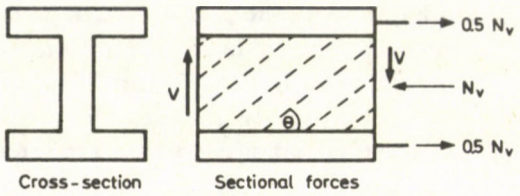


Fig. 7. Longitudinal forces due to shear

4. Choice of Θ

In designing a section to resist shear the engineer may choose any value of Θ between 15° and 75°, however the same value must be used in satisfying all of the requirements of a section.

Choosing values of Θ less than 45° will result in less transverse reinforcement but more longitudinal reinforcement being required which is usually an economical trade-off. As Θ is made smaller f_2 becomes larger. For a constant ε_x making Θ smaller increases ε_1 and hence decreases $f_{2\max}$. The lower limit for Θ is set when f_2 reaches $f_{2\max}$. The lower the shear stress on the concrete the lower the value of Θ at which f_2 will equal $f_{2\max}$ (see Fig. 8).

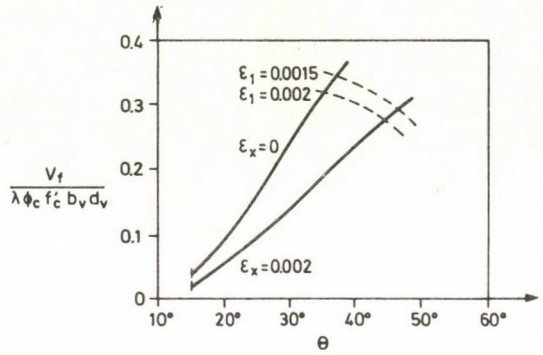


Fig. 8. Values of Θ at which $f_2 = f_{2\max}$

The longitudinal strain at mid-depth, ε_x, (see Fig. 2) can be found by performing a plane sections analysis for the section under the applied moment, M_f, axial force, N_f, and the equivalent axial force N_v, Eq. (5). Sections with high axial compression, prestress or low values of moment will have small web deformations (low ε_x and thus low ε_1) and hence will be able to tolerate higher shear stresses.

In lieu of determining ε_x from a plane sections analysis, the Code permits ε_x to be taken conservatively as -0.002.

If the cross-sectional dimensions are adequate it will be possible to choose a value of Θ which ensures that the concrete does not crush prematurely and that the transverse reinforcement yields before failure ($\varepsilon_t > f_y/E_s$). Note from Fig. 5 that $\varepsilon_t = \varepsilon_1 - \varepsilon_x - 0.002$. A chart such as that shown in Fig. 8 can assist in the choice of Θ.

After Θ has been chosen the transverse reinforcement is designed to satisfy Eq. (4) while the longitudinal reinforcement is designed to resist the equivalent axial tension N_v in addition to the applied moments, M_f, and axial loads, N_f. Because the influence of the shear on the longitudinal reinforcement is accounted for directly, the traditional detailing rules intended for this purpose can be waived.

5. Comparison

To compare the general method of the new Canadian Code with other code prescriptions an analytical study was performed. It was assumed that no bending moments, no axial forces and no prestressing occur at the section investigated and that

the longitudinal reinforcement is adequate for the shear. The spalling of the concrete cover was neglected. In the calculations no safety factors were included and material strengths were taken as characteristic strengths. Only stirrups as shear reinforcement were taken into account.

The results of the comparison calculations between the ACI 318-83 method, the general method of the new Canadian Code, the CEB MC-78 accurate method, the USSR SNIP-II-21-75 and the existing Hungarian Code MSz 15022/1—86 and the previous MSz 15022/1—71 are shown in Fig. 9. The truss angle of Θ for the general method was chosen as the average of the possible values, approximately corresponding to a longitudinal strain ε_x of -0.001 at mid-depth. For the CEB method the values of Θ was chosen of the lowest possible ($\tan \Theta = 3/5$).

From Fig. 9. it is clear that the new general method permits higher shear stresses than the traditional ACI approach and covers all the possible domains of shear behaviour.

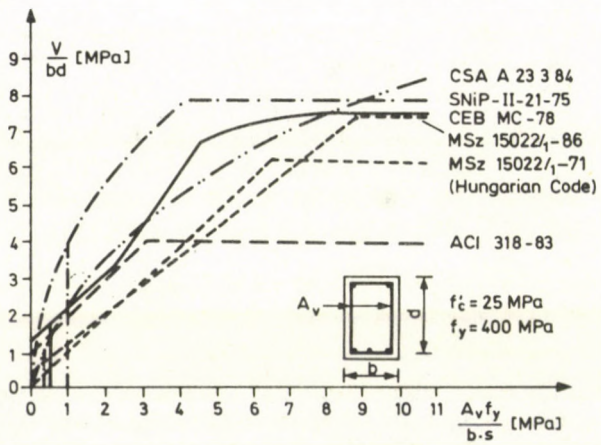


Fig. 9. Comparison of shear provisions of several codes (The vertical lines correspond to minimal shear reinforcement)

The different codes give a large scatter in predicted shear capacities, the ACI being the most conservative and the SNIP the highest. In comparing these values one should take into account, that the CSA permits the use of the staggering concept, which for a given beam results in designing for lower values of shear.

6. Concluding Remarks

Prestressed concrete, deep beams, corbels, control of diagonal cracking, spalling of the concrete cover, spacing limits for transverse reinforcement and truss models for design regions adjacent to supports, concentrated loads or abrupt changes in cross

section, are additional topics covered in the new Canadian Code with the help of the compression field theory and truss models.

In dealing with all these topics, the aim has been to develop regulations which are integrated but not complicated, having a clear physical explanation.

References

1. Collins, M. P.: Towards a Rational Theory for RC Members in Shear, Journal of the Structural Division, ASCE, V. **104,** April 1978, 649–666
2. Collins, M. P.–Mitchell, D.: Shear and Torsion Design of Prestressed and Non-Prestressed Concrete Beams, Journal of the Prestressed Concrete Institute, V. **25,** No. 5, Sept./Oct. 1980, 32–100
3. Mitchell, D.–Collins, M. P.: Diagonal Compression Field Theory—A Rational Model for Structural Concrete in Pure Torsion, Journal of the American Concrete Institute, V. **71,** Aug. 1974, 396–408
4. Marti, P.: Basic Tools of Reinforced Concrete Beam Design, ACI Journal, Nov.–Dec. 1984
5. Marti, P.: The Use of Truss Models in Detailing, to be published in ACI Journal, Nov.–Dec. 1985
6. Mueller, P.: Plastische Berechnung von Stahlbetonscheiben und Balken, Juli 1978, Bericht Nr. 83, Institut für Baustatik und Konstruktion ETH, Zürich, 160
7. Thuerlimann, B.–Marti, P.–Pralong, J.–Ritz, P.–Zimmerli, B.: Anwendung der Plastizitaetstheorie auf Stahlbeton, (Application of the Theory of Plasticity to Reinforced Concrete), Institute of Structural Engineering, ETH Zürich, 1983, 252
8. Vecchio, F.–Collins, M. P.: The Response of Reinforced Concrete to in-Plane Shear and Normal Stresses, University of Toronto, Dept. of Civil Engineering, Publication No. 82–03, Mar. 1983, 332

Acta Technica Academiae Scientiarum Hungaricae, 98 (3—4), pp. 205—219 (1985)

TRANSFORMATION OF TIME-VARYING MULTIVARIABLE LINEAR DISCRETE-TIME SYSTEMS INTO A PHASE-VARIABLE BLOCK OF CANONICAL FORM

S. CSAPÓ*

[Received: 30 August 1984]

Taken as a basis for investigation in this work is a class of time variant multivariable linear discrete-time systems of order n and input r where $r \leqq n$ and quotient n/r are not whole numbers. Specified for this class of systems are the necessary and sufficient conditions of transformability into a phase-variable block canonic (PVBC) form. Transformation of the state equation into PVBC form has been shown by Fahmy and O'Reilly for the aforementioned class of multivariable linear discrete-time systems but limited to the case constant in time. Consequently, this work is intended to extend the relationships deduced by Fahmy and O'Reilly to the case varying with time.

1. Introduction

The state equation of time variant multivariable linear discrete-time systems over set R of real numbers can be determined by vectorial difference equation

$$\mathbf{x}(k+1)=\mathbf{A}(k)\mathbf{x}(k)+\mathbf{B}(k)\mathbf{u}(k) \tag{1}$$

the initial state being $\mathbf{x}(k_0)$

where $k \in Z$	whole numbers
$\mathbf{x}(k) \in \chi$	$n \times 1$ state vector
$\mathbf{u}(k) \in U$	$r \times 1$ control vector
$\mathbf{A}(k) \in R^{n \times n}$ and	
$\mathbf{B}(k) \in R^{n \times r}$	properly dimensioned state matrices
$\chi = R^n$ and	
$U = R^r$	Eucledian spaces,

and accordingly, R^n and R^r are the state space and control space respectively. Assume that system matrix $\mathbf{A}(k)$ is non-singular and/or that column vectors of number r ($r \leqq n$) of input matrix $\mathbf{B}(k)$ are linearly independent for each $k \in Z$.

Let us introduce on the basis of the work of Fahmy and O'Reilly [1] the definition of positive whole numbers λ, β and γ. Accordingly, λ is the greatest whole

* S. Csapó, H-5130 Jászapáti, Vöröshadsereg u. 57, Hungary

number which is less than n/r, and

$$\beta = n - \lambda r, \qquad 0 < \beta < r, \tag{2a}$$

$$\gamma = r - \beta, \qquad 0 < \gamma < r. \tag{2b}$$

Let α be the smallest whole number for which controllability matrix

$$\mathbf{Q}_{c,\alpha}(k_0) = [\mathbf{B}(k_0+\alpha-1), \mathbf{A}(k_0+\alpha-1)\mathbf{B}(k_0+\alpha-2), \ldots$$
$$\ldots, \mathbf{A}(k_0+\alpha-1)\mathbf{A}(k_0+\alpha-2) \ldots \mathbf{A}(k_0+1)\mathbf{B}(k_0)]$$

for time k_0 is of rank n that means that at least one non-singular quadratic matrix of dimensions $n \times n$ could be selectable from matrix $\mathbf{Q}_{c,\alpha}(k_0) \in R^{n \times \alpha r}$. In this case, α is the controllability index [2]. Controllability index for system (1) characterized by relationships (2): $\alpha = \lambda + 1$. Thus, the state equation of PVBC form associated with system (1) is defined for time interval $k_0 \leqq k \leqq k_0 + \lambda$.

Let us introduce a variable parameter linear transformation by means of relationship

$$\mathbf{z}(k) = \mathbf{T}(k)\mathbf{x}(k) \tag{3}$$

and assume that there exist an inverse $\mathbf{T}^{-1}(k)$ of transformation matrix $\mathbf{T}(k) \in R^{n \times n}$ for each k within time interval $k_0 \leqq k \leqq k_0 + \lambda$ that is

$$\text{rank } \mathbf{T}(k) = n, \qquad k_0 \leqq k \leqq k_0 + \lambda. \tag{4}$$

With linear transformation in (3) substituted into the state equation given in (1) we obtain state equation

$$\mathbf{z}(k+1) = \bar{\mathbf{A}}(k)\mathbf{z}(k) + \bar{\mathbf{B}}\mathbf{u}(k), \qquad k_0 \leqq k \leqq k_0 + \lambda \tag{5}$$

of PVBC form, where

$$\bar{\mathbf{A}}(k) = \mathbf{T}(k+1)\mathbf{A}(k)\mathbf{T}^{-1}(k), \tag{6a}$$

$$\bar{\mathbf{B}} = \mathbf{T}(k+1)\mathbf{B}(k). \tag{6b}$$

Configuration of system matrix $\bar{\mathbf{A}}(k) \in R^{n \times r}$ variable with time and input matrix $\bar{\mathbf{B}} \in R^{n \times n}$ constant in time:

$$\bar{\mathbf{A}}(k) = \mathbf{T}(k+1)\mathbf{A}(k)\mathbf{T}^{-1}(k) =$$

$$= \begin{bmatrix} \mathbf{0}_{\beta,\beta} & \mathbf{0}_{\beta,\gamma} & \mathbf{I}_{\beta,\beta} & \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} & \ldots & \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} \\ \mathbf{0}_{\gamma,\beta} & \mathbf{0}_{\gamma,\gamma} & \mathbf{0}_{\gamma,\beta} & \mathbf{I}_{\gamma,\gamma} & \mathbf{0}_{\gamma,\beta} & \ldots & \mathbf{0}_{\gamma,\gamma} & \mathbf{0}_{\gamma,\beta} \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0}_{\beta,\beta} & \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} & \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} & \ldots & \mathbf{0}_{\beta,\gamma} & \mathbf{I}_{\beta,\beta} \\ \mathbf{M}_1(k) & \mathbf{M}_2(k) & \mathbf{M}_3(k) & \mathbf{M}_4(k) & \mathbf{M}_5(k) & \ldots & \mathbf{M}_{2\lambda}(k) & \mathbf{M}_{2\lambda+1}(k) \\ \mathbf{N}_1(k) & \mathbf{N}_2(k) & \mathbf{N}_3(k) & \mathbf{N}_4(k) & \mathbf{N}_5(k) & \ldots & \mathbf{N}_{2\lambda}(k) & \mathbf{N}_{2\lambda+1}(k) \end{bmatrix} \tag{7a}$$

$$\bar{\mathbf{B}} = \mathbf{T}(k+1)\mathbf{B}(k) = \begin{bmatrix} \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} \\ \mathbf{0}_{\gamma,\gamma} & \mathbf{0}_{\gamma,\beta} \\ \vdots & \vdots \\ \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} \\ \mathbf{I}_{\gamma,\gamma} & \mathbf{0}_{\gamma,\beta} \\ \mathbf{0}_{\beta,\gamma} & \mathbf{I}_{\beta,\beta} \end{bmatrix} \tag{7b}$$

respectively

where $\mathbf{0}_{p,q}$ zero matrix,
$\mathbf{I}_{p,p} \in R^{p \times p}$ unit matrix, and

$$\mathbf{M}_i(k) \in R^{\gamma \times \beta},\ \mathbf{N}_i(k) \in R^{\beta \times \beta}, \qquad i = 1, 3, \ldots, 2\lambda+1 \tag{8a}$$

$$\mathbf{M}_i(k) \in R^{\gamma \times \gamma},\ \mathbf{N}_i(k) \in R^{\beta \times \gamma}, \qquad i = 2, 4, \ldots, 2\lambda. \tag{8b}$$

Configuration of transformation hypermatrix in (3):

$$\mathbf{T}(k) = [\mathbf{T}_1(k), \mathbf{T}_2(k), \ldots, \mathbf{T}_{2\lambda-1}(k), \mathbf{T}_{2\lambda}(k), \mathbf{T}_{2\lambda+1}(k)]^B \tag{9}$$

where the superscript B denotes the block transpose, and

$$\mathbf{T}_i(k) \in R^{\beta \times n}, \qquad i = 1, 3, \ldots, 2\lambda+1 \tag{10a}$$

$$\mathbf{T}_i(k) \in R^{\gamma \times n}, \qquad i = 2, 4, \ldots, 2\lambda. \tag{10b}$$

Note that the relationships outlined so far differ from Fahmy and O'Reilly's relationships 1983a, 1983b only in that the corresponding matrices are now considered to be variable with time.

2. Main results

As seen, transformation matrix $\mathbf{T}(k) \in R^{n \times n}$ (9) shall be determined in order to produce the state equation of PVBC form (5). On the basis of relationship (6a), matrix equation

$$\bar{\mathbf{A}}(k)\mathbf{T}(k) = \mathbf{T}(k+1)\mathbf{A}(k), \qquad k_0 \leqq k \leqq k_0 + \lambda \tag{11}$$

can be written. On the basis of (11), taking into consideration system matrix $\bar{\mathbf{A}}(k)$ given in (7a), relationships

$$\mathbf{T}_i(k) = \mathbf{T}_{i-2}(k+1)\mathbf{A}(k) \in R^{\beta \times n}, \qquad i = 3, 5, \ldots, 2\lambda+1 \tag{12a}$$

$$\mathbf{T}_i(k) = \mathbf{T}_{i-2}(k+1)\mathbf{A}(k) \in R^{\gamma \times n}, \qquad i = 4, 6, \ldots, 2\lambda \tag{12b}$$

can be written. After repeated substitution, the following relationships are obtained on the basis of (12):

$$\mathbf{T}_i(k)=\mathbf{T}_1\left(k+\frac{i-1}{2}\right)\prod_{j=1}^{(i-1)/2}\mathbf{A}\left(k-j+\frac{i-1}{2}\right), \qquad i=1,3,\ldots,2\lambda+1\,, \tag{13a}$$

$$\mathbf{T}_i(k)=\mathbf{T}_2\left(k+\frac{i-2}{2}\right)\prod_{j=1}^{(i-2)/2}\mathbf{A}\left(k-j+\frac{i-2}{2}\right), \qquad i=2,4,\ldots,2\lambda\,. \tag{13b}$$

Taking into consideration definition

$$\mathbf{F}(k,k_0)\triangleq\begin{cases}\prod_{j=k_0}^{k-1}\mathbf{A}(j)=\mathbf{A}(k-1)\mathbf{A}(k-2)\ldots\mathbf{A}(k_0) & \text{if} \quad k>k_0\\ \mathbf{I} & \text{if} \quad k=k_0\\ \text{non-defined} & \text{if} \quad k<k_0\end{cases} \tag{14}$$

of fundamental matrix $\mathbf{F}(k,k_0)\in R^{n\times n}$ of the controlled system given in (1), expressions (13) can be written also as

$$\mathbf{T}_i(k)=\mathbf{T}_1\left(k+\frac{i-1}{2}\right)\mathbf{F}\left(k+\frac{i-1}{2},k\right)=E^{(i-1)/2}\mathbf{T}_1(k)\,, \tag{15a}$$

$$\mathbf{T}_i(k)=\mathbf{T}_2\left(k+\frac{i-2}{2}\right)\mathbf{F}\left(k+\frac{i-2}{2},k\right)=E^{(i-2)/2}\mathbf{T}_2(k)\,. \tag{15b}$$

For the sake of convenience, a matrix-operator E has been introduced to (15), which means for any matrix $\mathbf{L}(k)\in R^{p\times n}$

$$E\mathbf{L}(k)\triangleq\mathbf{L}(k+1)\mathbf{F}(k+1\,,\mathrm{k}) \qquad \text{for each} \quad k\,. \tag{16}$$

In case matrix $\mathbf{L}(k)$ falls under repetition of the operation determined by operator E, j-times, we arrive at a result

$$E^j\mathbf{L}(k)=E(E^{j-1}\mathbf{L}(k))=E(\mathbf{L}(k+j-1)\mathbf{F}(k+j-1,k))=$$
$$=\mathbf{L}(k+j)\mathbf{F}(k+j,k)\,, \qquad j=0,1,\ldots. \tag{17}$$

The definition of abbreviated symbols (15) is thus obvious on the basis of (17). According to (15), matrix $\mathbf{T}(k)$ (9) will be thus

$$\mathbf{T}(k)=\begin{bmatrix}\mathbf{T}_1(k)\\ \mathbf{T}_2(k)\\ \vdots\\ E^{\lambda-1}\mathbf{T}_1(k)\\ E^{\lambda-1}\mathbf{T}_2(k)\\ E^{\lambda}\mathbf{T}_1(k)\end{bmatrix}=\begin{bmatrix}\mathbf{T}_1(k)\\ \mathbf{T}_2(k)\\ \vdots\\ \mathbf{T}_1(k+\lambda-1)\mathbf{F}(k+\lambda-1,k)\\ \mathbf{T}_2(k+\lambda-1)\mathbf{F}(k+\lambda-1,k)\\ \mathbf{T}_1(k+\lambda)\mathbf{F}(k+\lambda,k)\end{bmatrix}. \tag{18}$$

A conditions for the existence of the state equation of PVBC form (5) is that, according to (6) and in relation with (18), partial matrices

$$\mathbf{T}_1(k_0), \mathbf{T}_1(k_0+1), \ldots, \mathbf{T}_1(k_0+2\lambda+1), \tag{19a}$$

$$\mathbf{T}_2(k_0), \mathbf{T}_2(k_0+1), \ldots, \mathbf{T}_2(k_0+2\lambda) \tag{19b}$$

should exist. Now, the definition of partial matrices $\mathbf{T}_1(k) \in R^{\beta \times n}$ and $\mathbf{T}_2(k) \in R^{\gamma \times n}$ follows. If $(k-1)$ is written in place of k in (6b), then we will obtain the following relationships, taking into consideration the configuration of the input matrix given in (7b) and of the transformation matrix given in (9):

$$\mathbf{T}_i(k)\mathbf{B}(k-1) = \mathbf{\Delta}_i^{\beta,r}, \qquad i = 1, 3, \ldots, 2\lambda+1 \tag{20a}$$

where the definition of matrix $\mathbf{\Delta}_i^{\beta,r} \in R^{\beta \times r}$

$$\mathbf{\Delta}_i^{\beta,r} = \begin{cases} \mathbf{0}_{\beta,r} & \text{if} \quad i < 2\lambda+1 \\ [\mathbf{0}_{\beta,\gamma} \;\; \mathbf{I}_{\beta,\beta}] & \text{if} \quad i = 2\lambda+1 \end{cases}$$

or

$$\mathbf{T}_i(k)\mathbf{B}(k-1) = \mathbf{\Delta}_i^{\gamma,r}, \qquad i = 2, 4, \ldots, 2\lambda \tag{20b}$$

where the definition of matrix $\mathbf{\Delta}_i^{\gamma,r} \in R^{\gamma \times r}$

$$\mathbf{\Delta}_i^{\gamma,r} = \begin{cases} \mathbf{0}_{\gamma,r} & \text{if} \quad i < 2\lambda \\ [\mathbf{I}_{\gamma,\gamma} \;\; \mathbf{0}_{\gamma,\beta}] & \text{if} \quad i = 2\lambda \end{cases}$$

respectively.

Let us multiply equations (12) from the right side by input matrix $\mathbf{B}(k-1)$ and write $(k-1)$ in place of k. Then we obtain relationships

$$\mathbf{T}_i(k-1)\mathbf{B}(k-2) = \mathbf{T}_{i-2}(k)\mathbf{A}(k-1)\mathbf{B}(k-2) = \mathbf{\Delta}_i^{\beta,r}, \qquad i = 3, 5, \ldots, 2\lambda+1, \tag{21a}$$

$$\mathbf{T}_i(k-1)\mathbf{B}(k-2) = \mathbf{T}_{i-2}(k)\mathbf{A}(k-1)\mathbf{B}(k-2) = \mathbf{\Delta}_i^{\gamma,r}, \qquad i = 4, 6, \ldots, 2\lambda. \tag{21b}$$

Leaving now the relationships for $i=3$ and $i=4$ in (21) out of consideration, substituting in accordance with (12) in the expressions left, then writing $(k-1)$ in place of k we obtain:

$$\mathbf{T}_i(k-2)\mathbf{B}(k-3) = \mathbf{T}_{i-4}(k)\mathbf{A}(k-1)\mathbf{A}(k-2)\mathbf{B}(k-3) = \mathbf{\Delta}_i^{\beta,r},$$

$$i = 5, 7, \ldots, 2\lambda+1,$$

$$\mathbf{T}_i(k-2)\mathbf{B}(k-3) = \mathbf{T}_{i-4}(k)\mathbf{A}(k-1)\mathbf{A}(k-2)\mathbf{B}(k-3) = \mathbf{\Delta}_i^{\gamma,r},$$

$$i = 6, 8, \ldots, 2\lambda.$$

It is than easy to see that there exist relationships

$$\mathbf{T}_i(k-\mathrm{j})\mathbf{B}(k-j-1) = \mathbf{T}_{i-2j}(k)\mathbf{A}(k-1) \ldots \mathbf{A}(k-j)\mathbf{B}(k-j-1) = \mathbf{\Delta}_i^{\beta,r}, \tag{22a}$$

$$j = 1, 2, \ldots, \lambda; \qquad i = 2j+1, \ldots, 2\lambda+1.$$

and

$$\mathbf{T}_i(k-j)\mathbf{B}(k-j-1)=\mathbf{T}_{i-2j}(k)\mathbf{A}(k-1)\ \ldots\ \mathbf{A}(k-j)\mathbf{B}(k-j-1)=\mathbf{\Delta}_i^{\gamma,r} \tag{22b}$$

$$j=1,2,\ldots,\lambda-1\,;\qquad i=2j+2,\ldots,2\lambda$$

of general form. On the basis of (22b), relationship

$$\mathbf{T}_i(k-\mathrm{j})=\mathbf{T}_{i-2j}(k)\mathbf{A}(k-1)\ \ldots\ \mathbf{A}(k-j)\,,\quad \begin{matrix} j=1,2,\ldots,\lambda-1 \\ i=2j+2,\ldots,2\lambda \end{matrix} \tag{23}$$

follows, which will be used later. Let us now put $i=2j+1$ in place of i in (22a) and $i=2j+2$ in place of i in (22b), taking into consideration the equation for $i=1$ and $i=2$ in (20) respectively. Then, using (14), we obtain the following relationships:

$$\mathbf{T}_{2j+1}(k-j)\mathbf{B}(k-j-1)=\mathbf{T}_1(k)\mathbf{F}(k,k-j)\mathbf{B}(k-j-1)=\mathbf{\Delta}_{2j+1}^{\beta,r}\,. \tag{24a}$$

where $j=0,\ 1,\ \ldots,\ \lambda$ and

$$\Delta_{2j+1}^{\beta,r}=\begin{cases}\mathbf{0}_{\beta,r}, & \text{if}\quad j<\lambda,\\ [\mathbf{0}_{\beta,\gamma},\ \mathbf{I}_{\beta,\beta}], & \text{if}\quad j=\lambda\end{cases}$$

and

$$\mathbf{T}_{2j+2}(k-j)\mathbf{B}(k-j-1)=\mathbf{T}_2(k)\mathbf{F}(k,k-j)\mathbf{B}(k-j-1)=\mathbf{\Delta}_{2j+2}^{\gamma,r} \tag{24b}$$

where $j=0,1,\ldots,\lambda-1$

$$\Delta_{2j+2}^{\gamma,r}=\begin{cases}\mathbf{0}_{\gamma,r}, & \text{if}\quad j<\lambda-1,\\ [\mathbf{I}_{\gamma,\gamma},\ \mathbf{0}_{\gamma,\beta}], & \text{if}\quad j=\lambda-1\end{cases}$$

respectively.

Let us introduce an operator $\tilde{E}$ which, for matrix $\mathbf{B}(k-1)$, means

$$\tilde{E}\mathbf{B}(k-1)\triangleq\mathbf{F}(k,k-1)\mathbf{B}(k-2). \tag{25}$$

In case matrix $\mathbf{B}(k-1)$ falls under repetition of the operation (25) determined by operator $\tilde{E}$ j-times, we arrive at the result

$$\tilde{E}^j\mathbf{B}(k-1)=\tilde{E}(\tilde{E}^{j-1}\mathbf{B}(k-1))=\tilde{E}(\mathbf{F}(k,k-j+1)\mathbf{B}(k-j))=$$

$$=\mathbf{F}(k,k-j)\mathbf{B}(k-j-1),\qquad j=0,1,\ldots. \tag{26}$$

Using the abbreviated symbols according to (26), the equations of number $\lambda+1$ (24a) and the equation of number λ of (24b) can be combined to obtain one single matrix equation, as follows:

$$\mathbf{T}_1(k)\,[\mathbf{B}(k-1),\,\tilde{E}\mathbf{B}(k-1),\,\ldots,\,\tilde{E}^{\lambda-1}\mathbf{B}(k-1),\,\tilde{E}^{\lambda}\mathbf{B}(k-1)]=$$

$$=[\mathbf{0}_{\beta,\lambda r},\ \mathbf{0}_{\beta,\gamma},\ \mathbf{I}_{\beta,\beta}]\in R^{\beta\times(\lambda+1)r}, \tag{27a}$$

and

$$\mathbf{T}_2(k)\,[\mathbf{B}(k-1),\ \tilde{E}\mathbf{B}(k-1),\ \ldots,\ \tilde{E}^{\lambda-2}\mathbf{B}(k-1),\ \tilde{E}^{\lambda-1}\mathbf{B}(k-1)] =$$

$$= [\mathbf{0}_{\gamma,(\lambda-1)r},\ \mathbf{I}_{\gamma,\gamma},\ \mathbf{0}_{\gamma,\beta}] \in R^{\gamma\times\lambda r} \tag{27b}$$

respectively, where the value of k according to (19) $k = k_0, k_0+1, \ldots, k_0+2\lambda+1$. It can be seen that the attainability matrix for time k appears in (27a) according to attainability index $\lambda+1$:

$$\mathbf{Q}_r(k) = [\mathbf{B}(k-1),\ \tilde{E}\mathbf{B}(k-1),\ \ldots,\ \tilde{E}^{\lambda-1}\mathbf{B}(k-1),\ \tilde{E}^{\lambda}\mathbf{B}(k-1)] =$$

$$= [\mathbf{B}(k-1),\ \mathbf{F}(k,k-1)\mathbf{B}(k-2),\ \ldots,\ \mathbf{F}(k,k-\lambda+1)\mathbf{B}(k-\lambda),$$

$$\mathbf{F}(k,k-\lambda)\mathbf{B}(k-\lambda-1)] \in R^{n\times(\lambda+1)r}. \tag{28}$$

As can be seen, matrix equation (27a) cannot be inverted at all while matrix equation (27b) is non-invertible in an ordinary sense for partial matrices $\mathbf{T}_1(k) \in R^{\beta\times n}$ and $\mathbf{T}_2(k) \in R^{\gamma\times n}$ respectively. To overcome this difficulty, let us decompose matrix $\mathbf{B}(k-1) \in R^{n\times r}$ in the following form:

$$\mathbf{B}(k-1) \triangleq [\mathbf{B}_1(k-1),\ \mathbf{B}_2(k-1)] \tag{29}$$

where $\mathbf{B}_1(k-1) \in R^{n\times\gamma}$ and $\mathbf{B}_2(k-1) \in R^{n\times\beta}$. Accordingly, also relationships

$$\tilde{E}^{\lambda}\mathbf{B}(k-1) = [\tilde{E}^{\lambda}\mathbf{B}_1(k-1),\ \tilde{E}^{\lambda}\mathbf{B}_2(k-1)] \tag{30}$$

exist where, according to (26),

$$\tilde{E}^{\lambda}\mathbf{B}_1(k-1) = \mathbf{F}(k,k-\lambda)\mathbf{B}_1(k-\lambda-1) \in R^{n\times\gamma} \tag{31}$$

in case $j=\lambda$.

Let us now take into consideration the decomposition according to (30) in attainability matrix $\mathbf{Q}_r(k)$ (28). Then, in case column vectors of number γ of (31) depend linearly on column vectors of number λr of matrix $\mathbf{Q}_r(k) \in R^{n\times(\lambda+1)r}$ for each k within time interval $k_0 \leqq k \leqq k_0+2\lambda+1$ that is, if relationship

$$\text{range}\,[\mathbf{B}(k-1),\ \tilde{E}\mathbf{B}(k-1),\ \ldots,\ \tilde{E}^{\lambda-1}\mathbf{B}(k-1)] \supseteq \text{range}\,[\tilde{E}^{\lambda}\mathbf{B}_1(k-1)], \qquad k = k_0, k_0+1, \ldots, k_0+2\lambda+1 \tag{32}$$

exists, then matrix equation

$$\mathbf{T}_1(k)\tilde{E}^{\lambda}\mathbf{B}_1(k-1) = \mathbf{0}_{\beta,\gamma}, \qquad k_0 \leqq k \leqq k_0+2\lambda+1 \tag{33}$$

can be omitted from (27a). Note that relationship (32) is a generalization of Ramar and Ramaswami's hypothesis formulated for systems continuous in time in 1971 for discrete-time systems.

According to (33), the matrix equation given in (27a) can be written as

$$\mathbf{T}_1(k)\,[\mathbf{B}(k-1),\ \tilde{E}\mathbf{B}(k-1),\ \ldots,\ \tilde{E}^{\lambda-1}\mathbf{B}(k-1),\ \tilde{E}^{\lambda}\mathbf{B}_2(k-1)] =$$

$$= [\mathbf{0}_{\beta,\lambda r},\ \mathbf{I}_{\beta,\beta}] \in R^{\beta\times n}, \qquad k_0 \leqq k \leqq k_0+2\lambda+1 \tag{34}$$

where now the attainability matrix, truncated as compared with (28), appears in the form of an $n \times n$ quadratic matrix:

$$\mathbf{Q}_{rt}(k) = [\mathbf{B}(k-1), \tilde{E}\mathbf{B}(k-1), \ldots, \tilde{E}^{\lambda-1}\mathbf{B}(k-1), \tilde{E}^{\lambda}\mathbf{B}_2(k-1)]. \tag{35}$$

Then, taking into consideration the symbols used in (35), the matrix equation given in (34) can be written in a simpler form:

$$\mathbf{T}_1(k)\mathbf{Q}_{rt}(k) = [\mathbf{0}_{\beta,\lambda r}, \mathbf{I}_{\beta,\beta}], \qquad k_0 \leqq k \leqq k_0 + 2\lambda + 1. \tag{36}$$

It goes now without saying that, for the matrix equation given in (27b), the missing equation of number $(\lambda+1)$ shall be assumed in such a way that quadratic matrix $\mathbf{Q}_{rt}(k) \in R^{n \times n}$ given in (35) will appear also in (27b). According to (11), matrix equation

$$\mathbf{M}_1(k)\mathbf{T}_1(k) + \mathbf{M}_2(k)\mathbf{T}_2(k) + \ldots + \mathbf{M}_{2\lambda}(k)\mathbf{T}_{2\lambda}(k) +$$
$$+ \mathbf{M}_{2\lambda+1}(k)\mathbf{T}_{2\lambda+1}(k) = \mathbf{T}_{2\lambda}(k+1)\mathbf{A}(k), \qquad k_0 \leqq k \leqq k_0 + 2\lambda \tag{37}$$

can be written. Let us postmultiply (37) by matrix $\mathbf{B}(k-1)$ and take the conditions of (20) into consideration. The following matrix equation can then be derived:

$$\mathbf{M}_{2\lambda}(k)\mathbf{T}_{2\lambda}(k)\mathbf{B}(k-1) + \mathbf{M}_{2\lambda+1}(k)\mathbf{T}_{2\lambda+1}(k)\mathbf{B}(k-1) =$$
$$= \mathbf{T}_{2\lambda}(k+1)\mathbf{A}(k)\mathbf{B}(k-1), \qquad k_0 \leqq k \leqq k_0 + \lambda. \tag{38}$$

Let us now substitute matrix $\mathbf{B}(k-1)$ as decomposed in (29) into (38), taking into consideration the conditions given in (20) for values $i = 2\lambda$ and $i = 2\lambda + 1$. Thus, the existence of equation (38) is possible also if conditions

$$\mathbf{M}_{2\lambda}(k) = \mathbf{T}_{2\lambda}(k+1)\mathbf{A}(k)\mathbf{B}_1(k-1), \qquad k_0 \leqq k \leqq k_0 + \lambda \tag{39}$$

$$\mathbf{M}_{2\lambda+1}(k) = \mathbf{T}_{2\lambda}(k+1)\mathbf{A}(k)\mathbf{B}_2(k-1), \tag{40}$$

are fulfilled. On the other hand, derivation of number $(\lambda+1)$ matrix equation mentioned earlier can obviously be expected only on the basis of (40). The condition given in (40) can be expressed by means of partial matrix $\mathbf{T}_2(k) \in R^{\gamma \times n}$ only if the matrix on the left side of (40) is invariable with time that is if condition

$$'\mathbf{M}_{2\lambda+1}(k) = \mathbf{M}_{2\lambda+1} = \text{const.}, \qquad k_0 \leqq k \leqq k_0 + \lambda \tag{41}$$

is fulfilled. Taking (41) into consideration, $(k-\lambda)$ can therefore be put in place of k in (40):

$$\mathbf{M}_{2\lambda+1} = \mathbf{T}_{2\lambda}(k-\lambda+1)\mathbf{A}(k-\lambda)\mathbf{B}_2(k-\lambda-1), \qquad k_0 \leqq k \leqq k_0 + \lambda. \tag{42}$$

According to the expression given in (23), for $j = \lambda - 1$,

$$\mathbf{T}_{2\lambda}(k-\lambda+1) = \mathbf{T}_2(k)\mathbf{A}(k-1)\ldots\mathbf{A}(k-\lambda+1), \qquad k_0 \leqq k \leqq k_0 + \lambda \tag{43}$$

can be written. After substitution of (43) into (42), and taking into consideration the abbreviated symbols in (26), we obtain the wanted number $(\lambda+1)$ matrix equation:

$$\mathbf{M}_{2\lambda+1}=\mathbf{T}_2(k)\mathbf{F}(k,k-\lambda)\mathbf{B}_2(k-\lambda-1)=\mathbf{T}_2(k)\tilde{E}^{\lambda}\mathbf{B}_2(k-1). \tag{44}$$

Now the matrix equation (27b) including (44) is

$$\mathbf{T}_2(k)\,[\mathbf{B}(k-1),\,\tilde{E}\mathbf{B}(k-1),\,\ldots,\,\tilde{E}^{\lambda-1}\mathbf{B}(k-1),\,\tilde{E}^{\lambda}\mathbf{B}_2(k-1)]=$$

$$=[\mathbf{0}_{\gamma,(\lambda-1)r},\,\mathbf{I}_{\gamma,\gamma},\,\mathbf{0}_{\gamma,\beta},\,\mathbf{M}_{2\lambda+1}]\in R^{\gamma\times n},\qquad k_0\leqq k\leqq k_0+2\lambda, \tag{45}$$

where matrix $\mathbf{Q}_{rt}(k)$ given in (35) appears. Thus, taking into consideration the symbols given in (35), matrix equation (45) can be written, as follows:

$$\mathbf{T}_2(k)\mathbf{Q}_{rt}(k)=[\mathbf{0}_{\gamma,(\lambda-1)r},\,\mathbf{I}_{\gamma,\gamma},\,\mathbf{0}_{\gamma,\beta},\,\mathbf{M}_{2\lambda+1}]. \tag{46}$$

Then, if there exists now the inverse of matrix $\mathbf{Q}_{rt}(k)\in R^{n\times n}$ for each k in time interval $k_0 \leqq k\leqq k_0+2\lambda+1$, that is if relationship

$$\operatorname{rank}\mathbf{Q}_{rt}(k)=n,\qquad k_0\leqq k\leqq k_0+2\lambda+1 \tag{47}$$

is fulfilled, then the matrix equation given in (36) and (46) will be invertible for partial matrix $\mathbf{T}_1(k)\in R^{\beta\times n}$ and $\mathbf{T}_2(k)\in R^{\gamma\times n}$ respectively, and consequently, matrices (19) can be calculated in the following form:

$$\mathbf{T}_1(k)=[\mathbf{0}_{\beta,\lambda r},\,\mathbf{I}_{\beta,\beta}]\mathbf{Q}_{rt}^{-1}(k),\qquad k_0\leqq k\leqq k_0+2\lambda+1; \tag{48}$$

$$\mathbf{T}_2(k)=[\mathbf{0}_{\gamma,(\lambda-1)r},\,\mathbf{I}_{\gamma,\gamma},\,\mathbf{0}_{\gamma,\beta},\,\mathbf{M}_{2\lambda+1}]\mathbf{Q}_{rt}^{-1}(k),\qquad k_0\leqq k\leqq k_0+2\lambda. \tag{49}$$

Transformation matrix $\mathbf{T}(k)$ (18) for time interval $k_0\leqq k\leqq k_0+\lambda+1$ can be considered known on the basis of (48) and (49).

Now we show that in case $\mathbf{Q}_{rt}(k)$ is not singular for any k within time interval $k_0 \leqq k\leqq k_0+\lambda$, then inverse $\mathbf{T}^{-1}(k)$ of transformation matrix $\mathbf{T}(k)$ will exist for each k within time interval $k_0\leqq k\leqq k_0+\lambda$.

On the other hand, it is easy to see that truncated attainability matrix $\bar{\mathbf{Q}}_{rt}(k)\in R^{n\times n}$ for the system of PVBC form (5) can be produced in the following form:

$$\bar{\mathbf{Q}}_{rt}(k)\triangleq\mathbf{T}(k)\mathbf{Q}_{rt}(k), \tag{50}$$

where

$$\bar{\mathbf{Q}}_{rt}(k)=[\bar{\mathbf{B}},\,\tilde{E}\bar{\mathbf{B}},\,\tilde{E}^2\mathbf{B},\,\ldots,\,\tilde{E}^{\lambda-1}\mathbf{B},\,\tilde{E}^{\lambda}\bar{\mathbf{B}}_2]=$$

$$=[\bar{\mathbf{B}},\,\bar{\mathbf{F}}(k,k-1)\bar{\mathbf{B}},\,\bar{\mathbf{F}}(k,k-2)\bar{\mathbf{B}},\,\ldots$$

$$\ldots,\,\bar{\mathbf{F}}(k,k-\lambda+1)\bar{\mathbf{B}},\,\bar{\mathbf{F}}(k,k-\lambda)\bar{\mathbf{B}}_2]\in R^{n\times n}. \tag{51}$$

Obviously, also the matrix product on the right side of (50) will result in a matrix of a

configuration according to (51):

$$\mathbf{T}(k)\mathbf{Q}_{rt}(k)=$$

$$=\begin{bmatrix} \mathbf{0}_{\beta,r} & \mathbf{0}_{\beta,r} & \mathbf{0}_{\beta,r} & \dots & \mathbf{0}_{\beta,r} & \mathbf{0}_{\beta,r} & \mathbf{I}_{\beta,\beta} \\ \mathbf{0}_{r,r} & \mathbf{0}_{r,r} & \mathbf{0}_{r,r} & \dots & \mathbf{0}_{r,r} & \mathbf{I}_{r,r} & \mathbf{L}_{1,\lambda+1} \\ \mathbf{0}_{r,r} & \mathbf{0}_{r,r} & \mathbf{0}_{r,r} & \dots & \mathbf{I}_{r,r} & \mathbf{L}_{2,\lambda} & \mathbf{L}_{2,\lambda+1} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ \mathbf{0}_{r,r} & \mathbf{I}_{r,r} & \mathbf{L}_{\lambda-1,3} & \dots & \mathbf{L}_{\lambda-1,\lambda-1} & \mathbf{L}_{\lambda-1,\lambda} & \mathbf{L}_{\lambda-1,\lambda+1} \\ \mathbf{I}_{r,r} & \mathbf{L}_{\lambda,2} & \mathbf{L}_{\lambda,3} & \dots & \mathbf{L}_{\lambda,\lambda-1} & \mathbf{L}_{\lambda,\lambda} & \mathbf{L}_{\lambda,\lambda+1} \end{bmatrix} \tag{52}$$

where partial matrices $\mathbf{L}_{i,j}(k) \in R^{r\times r}$

$$\mathbf{L}_{i,j}(k)=E^{i-1}\left\{\begin{bmatrix}\mathbf{T}_2(k)\\ E\mathbf{T}_1(k)\end{bmatrix}\right\}\tilde{E}^{j-1}\mathbf{B}(k-1), \quad \begin{matrix} i=2,3,\dots,\lambda \\ j=\lambda-i+2,\dots,\lambda\end{matrix} \tag{53}$$

and partial matrices $\mathbf{L}_{i,\lambda+1}(k) \in R^{r\times\beta}$

$$\mathbf{L}_{i,\lambda+1}(k)=E^{i-1}\left\{\begin{bmatrix}\mathbf{T}_2(k)\\ E\mathbf{T}_1(k)\end{bmatrix}\right\}\tilde{E}^{\lambda}\mathbf{B}_2(k-1), \qquad i=1,2,\dots,\lambda. \tag{54}$$

As seen, matrix $\bar{\mathbf{Q}}_{rt}(k)$ is non-singular, moreover, its determinant value is $|\bar{\mathbf{Q}}_{rt}(k)|=1$. In case relationship

$$\text{rank } \mathbf{Q}_{rt}(k)=n, \qquad k_0\leqq k\leqq k_0+\lambda \tag{55}$$

is fulfilled, there exists an inverse matrix $\mathbf{Q}_{rt}^{-1}(k)$ for each k within time interval $k_0\leqq k \leqq k_0+\lambda$ and thus the inverse of both sides of (50) can be taken:

$$\mathbf{Q}_{rt}^{-1}(k)\mathbf{T}^{-1}(k)=\bar{\mathbf{Q}}_{rt}^{-1}(k), \qquad k_0\leqq k\leqq k_0+\lambda. \tag{56}$$

From (56), inverse matrix $\mathbf{T}^{-1}(k)$ can be expressed as

$$\mathbf{T}^{-1}(k)=\mathbf{Q}_{rt}(k)\bar{\mathbf{Q}}_{rt}^{-1}(k), \qquad k_0\leqq k\leqq k_0+\lambda. \tag{57}$$

It follows then that (4) is not an additional condition for change-over to PVBC form (5), since the fulfilment of relationship (4) that is the condition for the existence of inverse matrix $\mathbf{T}^{-1}(k)$ has already been specified through relationship (55), according to (47).

According to what has been said so far, we might be right in saying that system (1) can be transformed into PVBC form (5) for time interval $k_0\leqq k\leqq k_0+\lambda$ only if relationships

$$\text{range}\,[\mathbf{B}(k-1), \tilde{E}\mathbf{B}(k-1), \dots, \tilde{E}^{\lambda-1}\mathbf{B}(k-1)] \supseteq \text{range}\,[\tilde{E}^{\lambda}\mathbf{B}_1(k-1)], \tag{58}$$

$$\text{rank}\,[\mathbf{B}(k-1), \tilde{E}\mathbf{B}(k-1), \dots, \tilde{E}^{\lambda-1}\mathbf{B}(k-1), \tilde{E}^{\lambda}\mathbf{B}_2(k-1)]=n \tag{59}$$

are fulfilled for each k within time interval $k_0\leqq k\leqq k_0+2\lambda+1$.

On the other hand, it can also be seen that partial matrix $\mathbf{M}_{2\lambda+1}(k)R \in {}^{\gamma\times\beta}$ assumed to be variable with time in system matrix $\bar{\mathbf{A}}(k)$ must actually be constant in time if matrix $\mathbf{T}_2(k) \in R^{\gamma\times n}$ had to be calculated on the basis of (49). Therefore, the state equation of PVBC form (5) is non-unique, because elements of number $\gamma \cdot \beta$ of matrix $\mathbf{M}_{2\lambda+1}$, constant in time, can be selected optionally, independently of each other.

Now we discuss the special case when quotient n/r is a whole number. The relevant relationships can be easily derived on the basis of what has been said so far.

3. The special case

When positive whole numbers β and γ for system (1) are defined as

$$\beta = n - \lambda r, \qquad 0 < \beta \leqq r, \tag{2*a}$$

$$\gamma = r - \beta, \qquad 0 \leqq \gamma < r \tag{2*b}$$

then the possibility of quotient n/r being a whole number is not excluded. Accordingly, in the special case where n/r is a whole number, there exist relationships $\gamma = 0$ and $r = \beta$.

Therefore,

$$n = (\lambda + 1)r$$

can be written in accordance with (2*), the value of controllability index being $\alpha = n/r = \lambda + 1$. To explain demonstratively: partial matrices of dimensions $\gamma \times p$ in the earlier relationships will all disappear because of $\gamma = 0$ while partial matrices of dimensions $\beta \times q$ will increase to a size of $r \times q$. Accordingly, transformation matrix (9) will have the form of

$$\mathbf{T}_0(k) = [\mathbf{T}_{01}(k),\ \mathbf{T}_{02}(k),\ \ldots,\ \mathbf{T}_{0\lambda}(k),\ \mathbf{T}_{0\lambda+1}(k)]^B \tag{9*}$$

where the expression of partial matrices $\mathbf{T}_{0i}(k) \in R^{r\times n}$ is

$$\mathbf{T}_{0i}(k) = \mathbf{T}_{01}(k+i-1)\mathbf{F}(k+i-1, k) = \mathbf{E}^{i-1}\mathbf{T}_{01}(k), \qquad i = 1, \ldots, \lambda+1\,. \tag{10*a}$$

Now, linear transformation (3):

$$\mathbf{z}(k) = \mathbf{T}_0(k)\mathbf{x}(k)\,. \tag{3*}$$

With transformation (3*) substituted in state equation (1) we obtain a state equation of PVBC form:

$$\mathbf{z}(k+1) = \bar{\mathbf{A}}_0(k)\mathbf{z}(k) + \bar{\mathbf{B}}_0\mathbf{u}(k), \qquad k_0 \leqq k \leqq k_0 + \lambda \tag{5*}$$

Configuration of time variant system matrix $\bar{\mathbf{A}}_0(k) \in R^{n\times n}$:

$$\bar{\mathbf{A}}_0(k)=\mathbf{T}_0(k+1)\mathbf{A}(k)\mathbf{T}_0^{-1}(k)=$$

$$=\begin{bmatrix} \mathbf{0}_{r,r} & \mathbf{I}_{r,r} & \mathbf{0}_{r,r} & \dots & \mathbf{0}_{r,r} & \mathbf{0}_{r,r} \\ \mathbf{0}_{r,r} & \mathbf{0}_{r,r} & \mathbf{I}_{r,r} & \dots & \mathbf{0}_{r,r} & \mathbf{0}_{r,r} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0}_{r,r} & \mathbf{0}_{r,r} & \mathbf{0}_{r,r} & \dots & \mathbf{0}_{r,r} & \mathbf{I}_{r,r} \\ \mathbf{A}_1(k) & \mathbf{A}_2(k) & \mathbf{A}_3(k) & \dots & \mathbf{A}_\lambda(k) & \mathbf{A}_{\lambda+1}(k) \end{bmatrix} \tag{7*a}$$

and expression of input matrix $\bar{\mathbf{B}}_0 \in R^{n\times r}$

$$\bar{\mathbf{B}}_0=\mathbf{T}_0(k+1)\mathbf{B}(k)=\begin{bmatrix} \mathbf{0}_{r,r} \\ \mathbf{0}_{r,r} \\ \vdots \\ \mathbf{0}_{r,r} \\ \mathbf{I}_{r,r} \end{bmatrix} \tag{7*b}$$

where $\mathbf{0}_{r,r}$ zero matrix,

$\mathbf{I}_{r,r} \in R^{r\times r}$ unit matrix, and

$\mathbf{A}_i(k) \in R^{r\times r}$, $\quad i=1, 2, \dots, \lambda+1$.

Using (10*a), configuration of transformation matrix $\mathbf{T}_0(k) \in R^{n\times n}$ (9*) is

$$\mathbf{T}_0(k)=\begin{bmatrix} \mathbf{T}_{01}(k) \\ E\mathbf{T}_{01}(k) \\ \vdots \\ E^{\lambda-1}\mathbf{T}_{01}(k) \\ E^{\lambda}\mathbf{T}_{01}(k) \end{bmatrix}=\begin{bmatrix} \mathbf{T}_{01}(k) \\ \mathbf{T}_{01}(k+1)\mathbf{F}(k+1,k) \\ \vdots \\ \mathbf{T}_{01}(k+\lambda-1)\mathbf{F}(k+\lambda-1,k) \\ \mathbf{T}_{01}(k+\lambda)\mathbf{F}(k+\lambda,k) \end{bmatrix} \tag{18*}$$

where $k=k_0, k_0+1, \dots, k_0+\lambda$. A condition for the existence of state equation of PVBC form (5*) is that partial matrices

$$\mathbf{T}_{01}(k_0), \mathbf{T}_{01}(k_0+1), \dots, \mathbf{T}_{01}(k_0+2\lambda+1) \tag{19*a}$$

exist according to (7*), in relation with (18*). Now, the necessary conditions (58) and (59) of transformability into PVBC form (5*) are combined in one single condition:

$$\text{rank}\,[\mathbf{B}(k-1), \tilde{E}\mathbf{B}(k-1), \dots, \tilde{E}^{\lambda}\mathbf{B}(k-1)]=n\,, \qquad k_0 \leqq k \leqq k_0+2\lambda+1\,, \tag{59*}$$

where, accordingly, attainability matrix given in (28), but being in this case just of dimensions $n \times n$, appears:

$$Q_r(k) = [\mathbf{B}(k-1), \tilde{E}\mathbf{B}(k-1), \ldots, \tilde{E}^{\lambda}\mathbf{B}(k-1)] \in R^{n \times n} \tag{28*}$$

Using the symbols given in (28*), condition (59*) can be given also as

$$\text{rank } \mathbf{Q}_r(k) = n, \qquad k_0 \leqq k \leqq k_0 + 2\lambda + 1 . \tag{47*}$$

Then, in case of fulfilment of condition (47*), the expression for partial matrices $\mathbf{T}_{01}(k) \in R^{r \times n}$ (19*a) is

$$\mathbf{T}_{01}(k) = [\mathbf{0}_{r,\lambda r}, \mathbf{I}_{r,r}]\mathbf{Q}_r^{-1}(k), \qquad k_0 \leqq k \leqq k_0 + 2\lambda + 1 . \tag{48*}$$

As is obvious on the basis of what has been said above, inverse $\mathbf{T}_0^{-1}(k)$ of transformation matrix $\mathbf{T}_0(k)$ (18*) will exist for each k within time interval $k_0 \leqq k \leqq k_0 + \lambda$ if the condition given in (47*) is fulfilled for values $k = k_0, k_0 + 1, \ldots, k_0 + \lambda$ that is, if relationship

$$\text{rank } \mathbf{T}_0(k) = n, \qquad k_0 \leqq k \leqq k_0 + \lambda \tag{4*}$$

will exist. Consequently, fulfilment of relationship (47*) is the necessary and sufficient condition for transformability into PVBC form (5*). Thus, it is also obvious that in case quotient n/r is a whole number, there exists only one single state equation of PVBC form (5*) according to the uniqueness of linear transformation (3*).

Finally, let us discuss the case constant in time.

4. The case constant in time

Taking the results obtained for time variant systems as a starting point, it is easy to change over to the case constant in time, that means that the relationships first given by Fahmy and O'Reilly (1983a, 1983b) [1] can be directly derived. It has been assumed by Fahmy and O'Reilly that the state matrices of the controlled system (1) are invariable with time. Therefore, Fahmy and O'Reilly determined the conditions of transformability into PVBC form only for systems constant in time, described by state equation

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k) \tag{60}$$

Discussed in the following are only the most important relationships. In the present case, the linear transformation given in (3) will have constant parameters:

$$\mathbf{z}(k) = \mathbf{T}\mathbf{x}(k) \tag{61}$$

where the configuration of transformation matrix $\mathbf{T} \in R^{n \times n}$ according to (9) is

$$\mathbf{T} = [\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_{2\lambda-1}, \mathbf{T}_{2\lambda}, \mathbf{T}_{2\lambda+1}]^B , \tag{62}$$

in which partical matrices

$$\mathbf{T}_i \in R^{\beta \times n}, \qquad i=1, 3, \ldots, 2\lambda+1, \tag{63a}$$

$$\mathbf{T}_i \in R^{\gamma \times n}, \qquad i=2, 4, \ldots, 2\lambda \tag{63b}$$

are included accordingly. Now, the state equation of PVBC form (5) can be written as

$$\mathbf{z}(k+1)=\bar{\mathbf{A}}\mathbf{z}(k)+\bar{\mathbf{B}}\mathbf{u}(k) \tag{64}$$

where the expressions of state matrices on the basis of (7) are

$$\bar{\mathbf{A}}=\mathbf{T}\mathbf{A}\mathbf{T}^{-1}, \tag{65a}$$

$$\bar{\mathbf{B}}=\mathbf{T}\mathbf{B}. \tag{65b}$$

For the case constant in time, fundamental matrix (14) has the form

$$\mathbf{F}(k-k_0) \triangleq \begin{cases} \mathbf{A}^{k-k_0} & \text{if} \quad k>k_0 \\ \mathbf{I}, & \text{if} \quad k=k_0 \\ \text{non-defined}, & \text{if} \quad k<k_0. \end{cases} \tag{66}$$

Since now the expression given in (17) is

$$E^j\mathbf{L} \triangleq \mathbf{L}\mathbf{F}(j)=\mathbf{L}\mathbf{A}^j, \qquad j=0, 1, \ldots, \lambda \tag{67}$$

therefore, partial matrices for the constant case (15) are

$$\mathbf{T}_i=E^{(i-1)/2}\mathbf{T}_1=\mathbf{T}_1\mathbf{F}\left(\frac{i-1}{2}\right)=\mathbf{T}_1\mathbf{A}^{(i-1)/2}, \qquad i=1, 3, \ldots, 2\lambda+1, \tag{68a}$$

$$\mathbf{T}_i=E^{(i-2)/2}\mathbf{T}_2=\mathbf{T}_2\mathbf{F}\left(\frac{i-2}{2}\right)=\mathbf{T}_2\mathbf{A}^{(i-2)/2}, \qquad i=2, 4, \ldots, 2\lambda. \tag{68b}$$

Then the expression of transformation matrix (62):

$$\mathbf{T}=[\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_1\mathbf{A}^{\lambda-1}, \mathbf{T}_2\mathbf{A}^{\lambda-1}, \mathbf{T}_1\mathbf{A}^{\lambda}]^B. \tag{69}$$

The necessary conditions (58) and (59) of transformability into PVBC form are now independent of time:

$$\text{range}\,[\mathbf{B}, \tilde{E}\mathbf{B}, \ldots, \tilde{E}^{\lambda-1}\mathbf{B}] \supseteq \text{range}\,[\tilde{E}^{\lambda}\mathbf{B}_1], \tag{70a}$$

$$\text{rank}\,[\mathbf{B}, \tilde{E}\mathbf{B}, \ldots, \tilde{E}^{\lambda}\mathbf{B}_2]=n. \tag{70b}$$

Since expression (26) takes now the form of

$$\tilde{E}^j\mathbf{B}=\mathbf{F}(j)\mathbf{B}=\mathbf{A}^j\mathbf{B}, \qquad j=0, 1, \ldots, \lambda,$$

conditions (70) can be written in the form given by Fahmy and O'Reilly (1983a, 1983b):

$$\text{range}\,[\mathbf{B}, \mathbf{A}\mathbf{B}, \ldots, \mathbf{A}^{\lambda-1}\mathbf{B}] \supseteq \text{range}\,[\mathbf{A}^{\lambda}\mathbf{B}_1], \tag{71a}$$

$$\text{rank}\,[\mathbf{B}, \mathbf{AB}, \ldots, \mathbf{A}^{\lambda-1}\mathbf{B}, \mathbf{A}^{\lambda}\mathbf{B}_2] = n\,. \tag{71b}$$

And finally, the expression of partial matrices $\mathbf{T}_1 \in R^{\beta \times n}$ and $\mathbf{T}_2 \in R^{\gamma \times n}$, changi)ng over from relationships (48) and (49) to the case constant in time:

$$\mathbf{T}_1 = [\mathbf{0}_{\beta,\lambda r}, \mathbf{I}_{\beta,\beta}]\,[\mathbf{B}, \mathbf{AB}, \ldots, \mathbf{A}^{\lambda}\mathbf{B}_2]^{-1}\,, \tag{72a}$$

$$\mathbf{T}_2 = [\mathbf{0}_{\gamma,(\lambda-1)r}, \mathbf{I}_{\gamma,\gamma}, \mathbf{0}_{\gamma,\beta}, \mathbf{M}_{2\lambda+1}]\,[\mathbf{B}, \mathbf{AB}, \ldots, \mathbf{A}^{\lambda}\mathbf{B}_2]^{-1}\,. \tag{72b}$$

Obviously, inverse $\mathbf{T}^{-1}$ of transformation matrix (69) will exist if the necessary and sufficient conditions given in (71) are fulfilled.

5. Conclusions

Determined in this work are the necessary and sufficient conditions of transformation into phase-variable block canonical form (PVBC) for a class of time variant multivariable linear discrete-time systems where quotient n/r is not a whole number. The relationships given here are essentially the theorem of Fahmy and O'Reilly (1983a) extended to the time variant case. It has been shown that partial matrix $\mathbf{M}_{2\lambda+1}(k) \in R^{\gamma \times \beta}$, assumed to be variable with time, of time variant system matrix $\bar{\mathbf{A}}(k)$ shall necessarily be constant in time $-\mathbf{M}_{2\lambda+1}(k) = \mathbf{M}_{2\lambda+1} = \text{const.}$—if in changing over to the case constant in time, we want to derive the relevant relationships of Fahmy and O'Reilly (1983a). Consequently, it can be seen that linear transformation into PVBC form is non-unique in the time variant case like in the case constant in time, because elements of number $\gamma \cdot \beta$ of partial matrix $\mathbf{M}_{2\lambda+1}$ can be selected optionally, independently of each other. Also, it was seen that the linear transformation unique only if quotient n/r had been a whole number. Hence, in this case, only one single state equation of PVBC form is associated with the system. Finally, it was shown that the results developed here for the case constant in time complied with the relationships are given by Fahmy and O'Reilly.

References

1. M. M. Fahmy, J. O'Reilly: Organisation of the non-uniqueness of a canonical structure of linear multivariable systems. Int. J. Syst. Sci., Vol. **14,** pp. 585–601, June 1983a Comments on design of optimal dead-beat controllers. IEEE. Trans. Automat. Contr., Vol. **AC-28,** pp. 125–127, January 1983b
2. R. E. Kalman, P. L. Falb, M. A. Arbib: Topics in Mathematical System Theory. New York: McGraw Hill, 1969
3. K. Ramar, B. Ramaswami: Transformation of time-variable multi-input systems to a canonical form. IEEE Trans. Automat. Contr., Vol. **AC-16,** pp. 371–374, August 1971

Acta Technica Academiae Scientiarum Hungaricae, 98 (3—4), pp. 221—232 (1985)

APPLICATION OF MINIMUM-TIME DEAD-BEAT CONTROL LAW TO A CLASS OF MULTIVARIABLE LINEAR SYSTEMS VARIABLE WITH TIME

S. CSAPÓ*

[Received: 3 September 1984]

This paper is intended to show that M. M. Fahmy's and J. O'Reilly's theorem can be conveniently applied to the area of time-varying systems. For a class of such systems of discrete-time the ratio n/r is not an integer, where n is the number of states and r is the number of controlled inputs. It can be seen that the time-varying minimum-time dead beat-beat (MTDB) control law appears in the parameters of a matrix which is constant in time. An example related to the third-order system shows that the MTDB control law can be determined by means of a scalar parameter.

1. Introduction

The state equation of time variant multivariable linear systems of discrete time over set R of real numbers can be given as a vectorial difference equation

$$\mathbf{x}(k+1)=\mathbf{A}(k)\mathbf{x}(k)+\mathbf{B}(k)\mathbf{u}(k)\,, \tag{1}$$

the initial state being $\mathbf{x}(k_0)$,

where $k \in Z$ whole numbers,
$\mathbf{x}(k) \in \chi$ $n \times 1$ state vector,
$\mathbf{u}(k) \in U$ $r \times 1$ control vector,
$\mathbf{A}(k) \in R^{n \times n}$ and
$\mathbf{B}(k) \in R^{n \times r}$ are properly dimensioned state matrices,
$\chi = R^n$ and
$U = R^r$ Euclidean spaces and,

consequently, R^n the state space and R^r the control space. Assume that system matrix $\mathbf{A}(k)$ can be inverted and/or the column vectors of number $r(r \leqq n)$ of input matrix $\mathbf{B}(k)$ are linearly independent for any $k \in Z$. The solution of state equation (1) for initial state

* Csapó, S.: H-5130 Jászapáti, Vöröshadsereg út 57, Hungary

$\mathbf{x}(k_0)$

$$\mathbf{x}(k)=\mathbf{F}(k,k_0)\mathbf{x}(k_0)+\sum_{i=k_0}^{k-1}\mathbf{F}(k,i+1)\mathbf{B}(i)\mathbf{u}(i) \tag{2}$$

where matrix $\mathbf{F}(k,k_0)\in R^{n\times n}$ is the fundamental matrix of system (1) according to definition [3]

$$\mathbf{F}(k,k_0)\hat{=}\begin{cases}\prod_{i=k_0}^{k-1}\mathbf{A}(i)=\mathbf{A}(k-1)\mathbf{A}(k-2)\ldots\mathbf{A}(k_0), & \text{if } k>k_0\\ \mathbf{I}, & \text{if } k=k_0\\ \text{non-defined}, & \text{if } k<k_0\end{cases} \tag{3}$$

Let α be the smallest positive whole number for which controllability matrix

$$\mathbf{Q}_{c,\alpha}(k_0)=[\mathbf{B}(k_0+\alpha-1),\ \mathbf{F}(k_0+\alpha,k_0+\alpha-1)\mathbf{B}(k_0+\alpha-2),\ \ldots \\ \ldots,\mathbf{F}(k_0+\alpha,k_0+1)\mathbf{B}(k_0)] \tag{4}$$

for time k_0 is of rank n that is relationship

$$\text{rank}\ \mathbf{Q}_{c,\alpha}(k_0)=n \tag{5}$$

exists. In case (5) is fulfilled, α will be the controllability index [3] for system (1). Assume that relationship (5) exists for any initial state $\mathbf{x}(k_0)\in R^n$ at any time $k_0\geqq 0$. In this case, system (1) is uniformly controllable within time interval $k_0\leqq k\leqq k_0+\alpha$ that is state vector $\mathbf{x}(k)$ (2) must be a zero vector for $k=k_0+\alpha$ in the form of

$$\mathbf{x}(k_0+\alpha)=\mathbf{0}=\mathbf{F}(k_0+\alpha,k_0)\mathbf{x}(k_0)+\sum_{i=k_0}^{k_0+\alpha-1}\mathbf{F}(k_0+\alpha,i+1)\mathbf{B}(i)\mathbf{u}(i), \tag{6}$$

according to the definition [3] of controllability.

In case (5) is fulfilled, the control series

$$\mathbf{u}(k_0),\mathbf{u}(k_0+1),\ldots,\mathbf{u}(k_0+\alpha-1) \tag{7}$$

occurring in equation (6) is called minimum–time dead-beat (MTDB) control series. Hence, MTDB control is defined for controllability index α. With control series (7) produced by state feedback, the MTDB control law can be defined as

$$\mathbf{u}(k)=\mathbf{K}(k)\mathbf{x}(k),\qquad k=k_0,k_0+1,\ldots,k_0+\alpha-1 \tag{8}$$

where $\mathbf{K}(k)\in R^{r\times n}$ are time variant state feedback matrices. To produce matrices $\mathbf{K}(k)$, controlled system (1) shall be transformed into a phase-variable block canonical (PVBC) form for time interval $k_0\leqq k\leqq k_0+\alpha-1$.

2. Phase-variable form for a class of multivariable systems variable with time

For the investigated class of time variant systems, quotient n/r is typically not a whole number. Thus, following the considerations of [1], [2], let us introduce the definition of positive whole numbers λ, β, and γ also here. Accordingly, λ is the greatest whole number less than n/r and

$$\beta = n - \lambda r\,, \qquad 0 < \beta < r\,, \tag{9a}$$

$$\gamma = r - \beta\,, \qquad 0 < \gamma < r\,. \tag{9b}$$

Now we can say that $\alpha = \lambda + 1$ is the controllability index of system (1) characterized by the relationships in (9).

Let us introduce a linear transformation of variable parameter by means of relationship

$$\mathbf{z}(k) = \mathbf{T}(k)\mathbf{x}(k) \tag{10}$$

and assume that there exists inverse $\mathbf{T}^{-1}(k)$ of transformation matrix $\mathbf{T}(k) \in R^{n \times n}$ for any k within interval $k_0 \leqq k \leqq k_0 + \lambda$ that is, there exists a relationship

$$\operatorname{rank} \mathbf{T}(k) = n\,, \qquad k_0 \leqq k \leqq k_0 + \lambda\,. \tag{11}$$

By substituting linear transformation (10) into equation (1) we obtain state equation

$$\mathbf{z}(k+1) = \bar{\mathbf{A}}(k)\mathbf{z}(k) + \bar{\mathbf{B}}\mathbf{u}(k)\,, \qquad k_0 \leqq k \leqq k_0 + \lambda \tag{12}$$

of PVBC form, the configuration of system matrix $\bar{\mathbf{A}}(k) \in R^{n \times n}$ variable with time, and input matrix $\bar{\mathbf{B}} \in R^{n \times r}$ being given by (13) and (14), respectively:

$$\bar{\mathbf{A}}(k) = \mathbf{T}(k+1)\mathbf{A}(k)\mathbf{T}^{-1}(k) =$$

$$= \begin{bmatrix} \mathbf{0}_{\beta,\beta} & \mathbf{0}_{\beta,\gamma} & \mathbf{I}_{\beta,\beta} & \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} & \dots & \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} \\ \mathbf{0}_{\gamma,\beta} & \mathbf{0}_{\gamma,\gamma} & \mathbf{0}_{\gamma,\beta} & \mathbf{I}_{\gamma,\gamma} & \mathbf{0}_{\gamma,\beta} & \dots & \mathbf{0}_{\gamma,\gamma} & \mathbf{0}_{\gamma,\beta} \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0}_{\beta,\beta} & \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} & \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} & \dots & \mathbf{0}_{\beta,\gamma} & \mathbf{I}_{\beta,\beta} \\ \mathbf{M}_1(k) & \mathbf{M}_2(k) & \mathbf{M}_3(k) & \mathbf{M}_4(k) & \mathbf{M}_5(k) & \dots & \mathbf{M}_{2\lambda}(k) & \mathbf{M}_{2\lambda+1}(k) \\ \mathbf{N}_1(k) & \mathbf{N}_2(k) & \mathbf{N}_3(k) & \mathbf{N}_4(k) & \mathbf{N}_5(k) & \dots & \mathbf{N}_{2\lambda}(k) & \mathbf{N}_{2\lambda+1}(k) \end{bmatrix} \tag{13}$$

$$\bar{\mathbf{B}} = \mathbf{T}(k+1)\mathbf{B}(k) = \begin{bmatrix} \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} \\ \mathbf{0}_{\gamma,\gamma} & \mathbf{0}_{\gamma,\beta} \\ \vdots & \vdots \\ \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} \\ \mathbf{I}_{\gamma,\gamma} & \mathbf{0}_{\gamma,\beta} \\ \mathbf{0}_{\beta,\gamma} & \mathbf{I}_{\beta,\beta} \end{bmatrix} \tag{14}$$

where $\mathbf{0}_{p,q}$ zero matrix,
$\mathbf{I}_{p,p} \in R^{p \times p}$ unit matrix,

and

$$\mathbf{M}_i(k) \in R^{\gamma \times \beta}, \quad \mathbf{N}_i(k) \in R^{\beta \times \beta}, \qquad i = 1, 3, \ldots, 2\lambda + 1, \tag{15a}$$

$$\mathbf{M}_i(k) \in R^{\gamma \times \gamma}, \quad \mathbf{N}_i(k) \in R^{\beta \times \gamma}, \qquad i = 2, 4, \ldots, 2\lambda. \tag{15b}$$

Let the configuration of time variant transformation hypermatrix $\mathbf{T}(k)$ in (10) be

$$\mathbf{T}(k) = [\mathbf{T}_1(k), \mathbf{T}_2(k), \ldots, \mathbf{T}_{2\lambda-1}(k), \mathbf{T}_{2\lambda}(k), \mathbf{T}_{2\lambda+1}(k)]^B \tag{16}$$

where B designated the block transpose, and

$$\mathbf{T}_i(k) \in R^{\beta \times n}, \qquad i = 1, 3, \ldots, 2\lambda + 1, \tag{17a}$$

$$\mathbf{T}_i(k) \in R^{\gamma \times n}, \qquad i = 2, 4, \ldots, 2\lambda. \tag{17b}$$

Note that the structure of relationships (10) through (17) complies with what has been presented in [1], however, the matrices dealt with now shall be considered to be variable with time.

3. Main results

Given in this section are the necessary and sufficient conditions for the existence of state equation of PVBC form (12) as well as the method to calculate state feedback matrices $\mathbf{K}(k) \in R^{r \times n}(k = k_0, k_0 + 1, \ldots, k_0 + \lambda)$ appearing in MTDB control law (8). In deriving the partial matrices (17) of transformation matrix $\mathbf{T}(k)$ (16), matrix equation

$$\bar{\mathbf{A}}(k)\mathbf{T}(k) = \mathbf{T}(k+1)\mathbf{A}(k), \qquad k_0 \leqq k \leqq k_0 + \lambda \tag{18}$$

that can be written on the basis of (13) shall be taken into consideration. On the basis of (18), it is easy to realize that the partial matrices (17) can be produced [4] in the following form:

$$\mathbf{T}_i(k) = \mathbf{T}_1\left(k + \frac{i-1}{2}\right)\mathbf{F}\left(k + \frac{i-1}{2}, \mathrm{k}\right) = E^{(i-1)/2}\mathbf{T}_1(k),$$

$$i = 1, 3, \ldots, 2\lambda + 1, \tag{19a}$$

$$\mathbf{T}_i(k)=\mathbf{T}_2\left(k+\frac{i-2}{2}\right)\mathbf{F}\left(k+\frac{i-2}{2},k\right)=E^{(i-2)/2}\mathbf{T}_2(k),$$

$$i=2,4,\ldots,2\lambda. \tag{19b}$$

Operator E introduced in (19) means for any matrix $\mathbf{L}(k)\in R^{p\times n}$

$$E\mathbf{L}(k)\triangleq\mathbf{L}(k+1)\mathbf{F}(k+1,k) \tag{20}$$

for each k.

Then, by repeating operation (20) relating to matrix $\mathbf{L}(k)$ j-times, we obtain relationship

$$E^j\mathbf{L}(k)=\mathbf{L}(k+j)\mathbf{F}(k+j,k),\qquad j=0,1,\ldots. \tag{21}$$

Thus, on the basis of (21), the definition of abbreviated symbols introduced in (19) is quite obvious. Transformation matrix $\mathbf{T}(k)$ can therefore be given in accordance with (19) as

$$\mathbf{T}(k)=\begin{bmatrix}\mathbf{T}_1(k)\\ \mathbf{T}_2(k)\\ \vdots\\ E^{\lambda-1}\mathbf{T}_1(k)\\ E^{\lambda-1}\mathbf{T}_2(k)\\ E^{\lambda}\mathbf{T}_1(k)\end{bmatrix}=\begin{bmatrix}\mathbf{T}_1(k)\\ \mathbf{T}_2(k)\\ \vdots\\ \mathbf{T}_1(k+\lambda-1)\mathbf{F}(k+\lambda-1,k)\\ \mathbf{T}_2(k+\lambda-1)\mathbf{F}(k+\lambda-1,k)\\ \mathbf{T}_1(k+\lambda)\mathbf{F}(k+\lambda,k)\end{bmatrix} \tag{22}$$

Since according to (18), matrix $\mathbf{T}(k)$ must exist for each k within interval $k_0\leqq k\leqq k_0+\lambda+1$, partial matrices

$$\mathbf{T}_1(k_0),\mathbf{T}_1(k_0+1),\ldots,\mathbf{T}_1(k_0+2\lambda+1), \tag{23a}$$

$$\mathbf{T}_2(k_0),\mathbf{T}_2(k_0+1),\ldots,\mathbf{T}_2(k_0+2\lambda) \tag{23b}$$

must exist in connection with (22). Then, system (1) can be transformed into PVBC form (12) for time interval $k_0\leqq k\leqq k_0+\lambda$ only if necessary and sufficient conditions

$$\text{range}\,[\mathbf{B}(k-1),\tilde{E}\mathbf{B}(k-1),\ldots,\tilde{E}^{\lambda-1}\mathbf{B}(k-1)]\supseteq\text{range}\,[\tilde{E}^{\lambda}\mathbf{B}_1(k-1)] \tag{24}$$

$$\text{rank}\,[\mathbf{B}(k-1),\tilde{E}\mathbf{B}(k-1),\ldots,\tilde{E}^{\lambda-1}\mathbf{B}(k-1),\tilde{E}^{\lambda}\mathbf{B}_2(k-1)]=n \tag{25}$$

are fulfilled for each k within time interval $k_0\leqq k\leqq k_0+2\lambda+1$ [4]. Partial matrices $\mathbf{B}_1(k-1)\in R^{n\times\gamma}$ and $\mathbf{B}_2(k-1)\in R^{n\times\beta}$ result from a decomposition of input matrix $\mathbf{B}(k-1)\in R^{n\times r}$ in a form

$$\mathbf{B}(k-1)=[\mathbf{B}_1(k-1),\mathbf{B}_2(k-1)]\qquad\text{for each}\quad k.$$

Operator $\tilde{E}$ introduced in (24) and (25) means for input matrix $\mathbf{B}(k-1)$

$$\tilde{E}\mathbf{B}(k-1) \triangleq \mathbf{F}(k, k-1)\mathbf{B}(k-2) \qquad \text{for each} \quad k. \tag{26}$$

Then a repetition of operation (26) relating to matrix $\mathbf{B}(k-1)j$-times results in relationship

$$\tilde{E}^j\mathbf{B}(k-1) = \mathbf{F}(k, k-j)\mathbf{B}(k-j-1), \qquad j = 0, 1, \ldots, \lambda, \tag{27}$$

while partial matrices (23) can be calculated as

$$\mathbf{T}_1(k) = [0_{\beta, \lambda r}, \mathbf{I}_{\beta, \beta}] \times [\mathbf{B}(k-1), \tilde{E}\mathbf{B}(k-1), \ldots, \tilde{E}^{\lambda-1}\mathbf{B}(k-1), \tilde{E}^{\lambda}\mathbf{B}_2(k-1)]^{-1}$$

$$k = k_0, k_0+1, \ldots, k_0+2\lambda+1 \tag{28a}$$

$$\mathbf{T}_2(k) = [\mathbf{0}_{\gamma, (\lambda-1)r}, \mathbf{I}_{\gamma, \gamma}, \mathbf{0}_{\gamma, \beta}, \mathbf{M}_{2\lambda+1}] \times$$

$$\times [\mathbf{B}(k-1), \tilde{E}\mathbf{B}(k-1), \ldots, \tilde{E}^{\lambda-1}\mathbf{B}(k-1), \tilde{E}^{\lambda}\mathbf{B}_2(k-1)]^{-1} \tag{28b}$$

[4]. Note that in case necessary condition (25) is fulfilled, linear transformation (10) can be inverted because in this case there exists inverse matrix $\mathbf{T}^{-1}(k)$ for each k within time interval $k_0 \leqq k \leqq k_0 + \lambda$ [4]. Therefore, necessary conditions (24) and (25) are at the same time sufficient conditions with respect to the existence of the state equation of PVBC form given in (12).

In accordance with relationship (28b), partial matrix $\mathbf{M}_{2\lambda+1}(k) \in R^{\gamma \times \beta}$ of system matrix $\bar{\mathbf{A}}(k) \in R^{n \times n}$, assumed to be time variant earlier, must be necessarily constant in time

$$\mathbf{M}_{2\lambda+1}(k) = \mathbf{M}_{2\lambda+1} = \text{const.}, \qquad \text{for each } k. \tag{29}$$

Accordingly, linear transformation (10) is non-unique as elements of number $\gamma \cdot \beta$ of partial matrix $\mathbf{M}_{2\lambda+1}$ as the parameter matrix, constant in time, can be chosen optionally. Consequently, state equations of PVBC form (12) of infinite number can be associated with the controlled system (1).

The condition for operation of MTDB (6) for system (12) can be applied to phase space R^n in the following way:

$$\mathbf{z}(k_0+\lambda+1) = \mathbf{0} = \bar{\mathbf{F}}(k_0+\lambda+1, k_0)\mathbf{z}(k_0) +$$

$$+ \sum_{i=k_0}^{k_0+\lambda} \bar{\mathbf{F}}(k_0+\lambda+1, i+1)\bar{\mathbf{B}}\mathbf{u}(i) \tag{30}$$

where the fundamental matrix of system (12) is

$$\bar{\mathbf{F}}(k, k_0) = \mathbf{T}(k)\mathbf{F}(k, k_0)\mathbf{T}^{-1}(k_0)$$

for each k.

For system (12), the MTDB control law:

$$\mathbf{u}(k) = \mathbf{K}(k)\mathbf{z}(k), \qquad k = k_0, k_0+1, \ldots, k_0+\lambda \tag{31}$$

where $\bar{\mathbf{K}}(k) \in R^{r \times n}$ is the state feedback matrix for phase space $\bar{R}^n$. With (31) substituted

into (12), we obtain the state equation of the closed system:

$$\mathbf{z}(k+1)=[\bar{\mathbf{A}}(k)+\bar{\mathbf{B}}\bar{\mathbf{K}}(k)]\mathbf{z}(k), \qquad k_0 \leqq k \leqq k_0+\lambda . \tag{32}$$

Let us now substitute the term of control vectors (30) into equation (30). By repeated use of (32), let us express any state $\mathbf{z}(k)$ $(k_0+1 \leqq k \leqq k_0+\lambda)$ by means of initial state $\mathbf{z}(k_0)$. Then, equation (30) can be written as

$$\mathbf{0}=\mathbf{z}(k_0+\lambda+1)=[\bar{\mathbf{A}}(k_0+\lambda)+\bar{\mathbf{B}}\bar{\mathbf{K}}(k_0+\lambda)] \ldots$$
$$\ldots[\bar{\mathbf{A}}(k_0+1)+\bar{\mathbf{B}}\bar{\mathbf{K}}(k_0+1)]\,[\bar{\mathbf{A}}(k_0)+\bar{\mathbf{B}}\bar{\mathbf{K}}(k_0)]\mathbf{z}(k_0) \tag{33}$$

Now, if each of matrices $\mathbf{K}(k)$ of number $\lambda+1$ in (33) is determined for values $k=k_0, k_0+1, \ldots, k_0+\lambda$ in form

$$\bar{\mathbf{K}}(k)=\begin{bmatrix} -\mathbf{M}_1(k) & -\mathbf{M}_2(k) & \ldots & -\mathbf{M}_{2\lambda}(k) & -\mathbf{M}_{2\lambda+1} \\ -\mathbf{N}_1(k) & -\mathbf{N}_2(k) & \ldots & -\mathbf{N}_{2\lambda}(k) & -\mathbf{N}_{2\lambda+1}(k) \end{bmatrix} \tag{34}$$

then the factor matrices in (33) will become constant in time, taking into consideration the state matrix configuration in (13) and (14):

$$\bar{\mathbf{W}}=[\mathbf{A}(k)+\bar{\mathbf{B}}\bar{\mathbf{K}}(k)]=\text{const.}, \qquad k_0 \leqq k \leqq k_0+\lambda ,$$

where the expression for matrix $\bar{\mathbf{W}} \in R^{n \times n}$ constant in time:

$$\bar{\mathbf{W}}=\begin{bmatrix} \mathbf{0}_{\beta,\beta} & \mathbf{0}_{\beta,\gamma} & \mathbf{I}_{\beta,\beta} & \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} & \ldots & \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} \\ \mathbf{0}_{\gamma,\beta} & \mathbf{0}_{\gamma,\gamma} & \mathbf{0}_{\gamma,\beta} & \mathbf{I}_{\gamma,\gamma} & \mathbf{0}_{\gamma,\beta} & \ldots & \mathbf{0}_{\gamma,\gamma} & \mathbf{0}_{\gamma,\beta} \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0}_{\beta,\beta} & \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} & \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} & \ldots & \mathbf{0}_{\beta,\gamma} & \mathbf{I}_{\beta,\beta} \\ \mathbf{0}_{\gamma,\beta} & \mathbf{0}_{\gamma,\gamma} & \mathbf{0}_{\gamma,\beta} & \mathbf{0}_{\gamma,\gamma} & \mathbf{0}_{\gamma,\beta} & \ldots & \mathbf{0}_{\gamma,\gamma} & \mathbf{0}_{\gamma,\beta} \\ \mathbf{0}_{\beta,\beta} & \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} & \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} & \ldots & \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} \end{bmatrix} \tag{35}$$

Taking (35) into consideration, the condition for operation of MTDB (35) takes the shape

$$\mathbf{z}(k_0+\lambda+1)=\bar{\mathbf{W}}^{\lambda+1}\mathbf{z}(k_0)=\mathbf{0} . \tag{36}$$

The condition given in (36) is fulfilled since matrix $\bar{\mathbf{W}}$ (35) is a nilpotent matrix according to superscript $\lambda+1$:

$$\bar{\mathbf{W}}^{\lambda+1}=\mathbf{0} . \tag{37}$$

With the linear transformation according to (10) substituted into (31), the MTDB control law (8) can be written in the form

$$\mathbf{u}(k)=\bar{\mathbf{K}}(k)\mathbf{T}(k)\mathbf{x}(k)=\mathbf{K}(k)\mathbf{x}(k), \qquad k_0 \leqq k \leqq k_0+\lambda \tag{38}$$

where the expression for state feedback matrices $\mathbf{K}(k)$ for original state space R^n:

$$\mathbf{K}(k)=\bar{\mathbf{K}}(k)\mathbf{T}(k)=\begin{bmatrix}-\mathbf{M}_1(k) & -\mathbf{M}_2(k) & \dots & \mathbf{M}_{2\lambda}(k) & -\mathbf{M}_{2\lambda+1}\\ -\mathbf{N}_1(k) & -\mathbf{N}_2(k) & \dots & \mathbf{N}_{2\lambda}(k) & -\mathbf{N}_{2\lambda+1}(k)\end{bmatrix}\mathbf{T}(k) \quad (39)$$

Finally, it should be taken into consideration that, in order to produce feedback matrices $\mathbf{K}(k)$ (39) ($k=k_0, k_0+1, \dots, k_0+\lambda$), transformation matrix $\mathbf{T}(k)$ (22) and its inverse $\mathbf{T}^{-1}(k)$ and then system matrix $\bar{\mathbf{A}}(k)$ from which then matrix $\bar{\mathbf{K}}(k)$ (34) in (39) becomes known shall be determined. However, it should be mentioned that in case the sufficient conditions (24) and (25) for transformability into PVBC form (12) are fulfilled, it is not necessary to calculate transformation matrix $\mathbf{T}(k)$ (22) and its inverse $\mathbf{T}^{-1}(k)$, nor system matrix $\bar{\mathbf{A}}(k)$ (13) itself to determine feedback matrices $\mathbf{K}(k)$ (39), but, instead, a simpler way can be followed. Namely, with matrix equation (18) taken into consideration, it can be seen that a simpler formula is obtained on the basis of (18) to calculate matrix $\mathbf{K}(k)$ (39):

$$\mathbf{K}(k)=\begin{bmatrix}-\mathbf{T}_{2\lambda}(k+1)\mathbf{A}(k)\\ -\mathbf{T}_{2\lambda+1}(k+1)\mathbf{A}(k)\end{bmatrix}, \qquad k_0\leqq k\leqq k_0+\lambda\,. \quad (40)$$

By means of abbreviated symbols (19), the expression for feedback matrix (40) can be written as

$$\mathbf{K}(k)=\begin{bmatrix}-E^{\lambda} & \mathbf{T}_2(k)\\ -E^{\lambda+1} & \mathbf{T}_1(k)\end{bmatrix} \quad (41)$$

where relationships

$$\mathbf{T}_{2\lambda}(k+1)\mathbf{A}(k)=\mathbf{T}_2(k+\lambda)\mathbf{F}(k+\lambda,k)=E^{\lambda}\mathbf{T}_2(k)\,, \quad (42)$$

$$\mathbf{T}_{2\lambda+1}(k+1)\mathbf{A}(k)=\mathbf{T}_1(k+\lambda+1)\mathbf{F}(k+\lambda+1,k)=E^{\lambda+1}\mathbf{T}_1(k) \quad (43)$$

exist for any $k=k_0, k_0+1, \dots, k_0+\lambda$.

4. Example

To illustrate what has been said so far, let $n=3$ and $r=2$. Let

$$\mathbf{A}(k)=\begin{bmatrix}1 & 1 & 0\\ 0 & 1 & 0\\ -e^{-k+1} & 0 & 1\end{bmatrix}, \qquad \mathbf{B}(k)=\begin{bmatrix}0 & 1\\ 0 & -1\\ 1 & e^{-k}\end{bmatrix} \quad (44)$$

be the state matrices of system (1).

As seen, now $\lambda=\beta=\gamma=1$. Thus matrix $\mathbf{B}(k-1)$ can be decomposed as

$$\mathbf{B}_1(k-1)=\begin{bmatrix}0\\0\\1\end{bmatrix}, \qquad \mathbf{B}_2(k-1)=\begin{bmatrix}1&\\-1&\\&e^{-k+1}\end{bmatrix}. \tag{45}$$

The sufficient conditions (24) and (25) of transformability into PVBC form (12) exist because relationships

$$\text{range}\,[\mathbf{B}(k-1)]\supset\text{range}\,[\mathbf{A}(k-1)\mathbf{B}_1(k-2)]\,, \tag{46}$$

$$\text{rank}\,[\mathbf{B}(k-1),\quad \mathbf{A}(k-1)\mathbf{B}_2(k-2)]=3 \tag{47}$$

for each k, where

$$[\mathbf{B}(k-1),\mathbf{A}(k-1)\mathbf{B}_2(k-2)]=\begin{bmatrix}0&1&0\\0&-1&-1\\1&e^{-k+1}&0\end{bmatrix}. \tag{48}$$

Now expressions (28) can be written as

$$\mathbf{T}_1(k)=[0,\quad 0,\quad 1]\,[\mathbf{B}(k-1),\mathbf{A}(k-1)\mathbf{B}_2(k-2)]^{-1}\,,$$

$$\mathbf{T}_2(k)=[1,\quad 0,\quad m]\,[\mathbf{B}(k-1),\mathbf{A}(k-1)\mathbf{B}_2(k-2)]^{-1}\,,$$

of which, by the use of matrix (48)

$$\mathbf{T}_1(k)=[-1,\qquad -1,\qquad 0], \tag{49a}$$

$$\mathbf{T}_2(k)=[-e^{-k+1}-m,\quad -m,\qquad 1]. \tag{49b}$$

Transformation matrix $\mathbf{T}(k)$ given in (22) will be in this case

$$\mathbf{T}(k)=\begin{bmatrix}-1&-1&0\\-e^{-k+1}-m&-m&1\\-1&-2&0\end{bmatrix}. \tag{50}$$

Matrices (13) and (14) of the state equation of PVBC form (12) are now:

$$\bar{\mathbf{A}}(k)=\begin{bmatrix}0&0&1\\e^{-k}-m&1&m\\-1&0&2\end{bmatrix}, \qquad \bar{\mathbf{B}}=\begin{bmatrix}0&0\\1&0\\0&1\end{bmatrix}. \tag{51}$$

Expression of state feedback matrices $\bar{\mathbf{K}}(k)$, relating to phase space $\bar{R}^n$, will now be for values $k=k_0$, k_0+1

$$\bar{\mathbf{K}}(k)=\begin{bmatrix} -e^{-k}+m & -1 & -m \\ 1 & 0 & -2 \end{bmatrix}. \tag{52}$$

According to (39), feedback matrix $\mathbf{K}(k)$ for original state space R^n:

$$\mathbf{K}(k)=\bar{\mathbf{K}}(k)\mathbf{T}(k)=\begin{bmatrix} e^{-k}+e^{-k+1}+m & e^{-k}+2m & -1 \\ 1 & 3 & 0 \end{bmatrix}. \tag{53}$$

We obtain again (53) by the use of (41):

$$\mathbf{K}(k)=\begin{bmatrix} -E\mathbf{T}_2(k) \\ -E^2\mathbf{T}_1(k) \end{bmatrix}=\begin{bmatrix} -\mathbf{T}_2(k+1)\mathbf{A}(k) \\ -\mathbf{T}_1(k+2)\mathbf{A}(k+1)\mathbf{A}(k) \end{bmatrix}. \tag{54}$$

Thus we know the MTDB control law (8) for the third-order system investigated. By substituting the control law (8) into the state equation given in (1) we obtain the state equation of the closed-loop system:

$$\mathbf{x}(k+1)=\mathbf{W}(k)\mathbf{x}(k), \qquad k_0 \leqq k \leqq k_0+\lambda \tag{55}$$

where the time variant state matrix, $\mathbf{W}(k) \in R^{n \times n}$, of the system be the use of matrices (44) and (53):

$$\mathbf{W}(k)=[\mathbf{A}(k)+\mathbf{B}(k)\mathbf{K}(k)]=\begin{bmatrix} 2 & 4 & 0 \\ -1 & -2 & 0 \\ 2e^{-k}+m & 4e^{-k}+2m & 0 \end{bmatrix}. \tag{56}$$

It is also obvious that relationship

$$\mathbf{x}(k_0+2)=\mathbf{W}(k_0+1)\mathbf{x}(k_0+1)=\mathbf{W}(k_0+1)\mathbf{W}(k_0)\mathbf{x}(k_0)=0 \tag{57}$$

shall exist for some initial state $\mathbf{x}(k_0)$. This is fulfilled because

$$\mathbf{W}(k_0+1)\mathbf{W}(k_0)=0\,. \tag{58}$$

On the other hand, the course of the trajectory starting from initial state $\mathbf{x}(k_0)$ obviously depends on the value of scalar parameter m selected optionally.

5. Conclusions

For the class (n/r being not a whole number) of multivariable linear systems variable with time, we have seen a possible method to determine the MTDB control law, which assumes the existence of a state equation of PVBC form of the system. The necessary and sufficient conditions of transformability into PVBC form with reference

to the new results developed in connection with the time variant case [4]. We have seen that the state equation of PVBC form is non-unique because partial matrix $\mathbf{M}_{2\lambda+1}(k) \in R^{\gamma \times \beta}$, assumed to be time variant initially, of time variant system matrix $\bar{\mathbf{A}}(k)$ must necessarily be constant in time, $\mathbf{M}_{2\lambda+1}(k) = \mathbf{M}_{2\lambda+1} = \text{const.}$, for each k. At the same time, matrix $\mathbf{M}_{2\lambda+1}$ appears as a parameter matrix as its elements of number $\gamma \cdot \beta$ can be selected independently of each other. Nor state feedback matrices $\bar{\mathbf{K}}(k_0)$, $\bar{\mathbf{K}}(k_0 + 1), \ldots, \bar{\mathbf{K}}(k_0 + \lambda)$ associated with the system of PVBC form are non-unique because parameter matrix $\mathbf{M}_{2\lambda+1}$ appears also in these matrices. As a result, feedback matrices $\mathbf{K}(k_0), \mathbf{K}(k_0 + 1), \ldots, \mathbf{K}(k_0 + \lambda)$ also appear in the parameters of matrix $\mathbf{M}_{2\lambda+1}$. Finally, it shall be recognized that the new results given here are a generalization of the relationships given in [1] for the time variant case, a fact easily detectable when changing over to the case constant in time.

References

1. M. M. Fahmy, J. O'Reilly: Comments on design of optimal dead-beat controllers. IEEE Trans. Automat. Contr., Vol. **AC-28,** pp. 125–127, Jan. 1983
2. Organisation of the non-uniqueness of a canonical structure of linear multivariable systems. Int. Syst. Sci., Vol. **14,** pp. 585–601, June 1983
3. R. E. Kalman, P. L. Falb, M. A. Arbib: Topics in Mathematical System Theory. New York, McGraw Hill, 1969
4. S. Csapó: A phase-variable block canonical form for a class of time-varying linear multiple-input discrete time systems. Int. J. Syst. Sci., submitted for publication

Acta Technica Academiae Scientiarum Hungaricae, 98 (3—4), pp. 233—250 (1985)

MINIMUM–TIME CONTROL OF TIME–VARYING MULTIVARIABLE, LINEAR, DISCRETE–TIME SYSTEMS BY STATE VARIABLES FEEDBACK

S. CSAPÓ

[Received: 8 January 1985]

Let the time-varying, linear, discrete-time controlled system be considered of state n and input $r(r \leqq n)$ and quotient n/r be non-integer. Assume for the system a trajectory starting from an initial state and arriving at a final state determined in advance, along which time is minimum. The minimum–time control problem so defineable can be realized by feedback of state variables of number n, assumed to be measurable. It will be seen that a possible variation of the required state control law can be given also on the basis of the theorem outlined previously. Also an example is presented for a third-order system to demonstrate the conditions.

1. Introduction

A time–varying multivariable, linear, discrete–time system over set R of real numbers can be given by vector-difference equation

$$\mathbf{x}(k+1) = \mathbf{A}(k)\mathbf{x}(k) + \mathbf{B}(k)\mathbf{u}(k), \qquad \text{initial state} \quad \mathbf{x}(k_0) \tag{1}$$

where $k \in Z$ (integers), $\mathbf{x}(k) \in R^n$ – state vector
$\mathbf{u}(k) \in R^r$ – control vector
$\mathbf{A}(k) \in R^{n \times n}$ and $\mathbf{B}(k) \in R^{n \times r}$ – properly dimensioned time–varying state matrices
R^n and R^r – Eucledian spaces, R^n being space of states while R^r control space.

Let rank $\mathbf{A}(k) = n$, and rank $\mathbf{B}(k) = r$ for all $k \in Z$. Let quotient n/r be non-integer for the class of systems considered. In case of such systems, controllability index α can be given as

$$\alpha = \lambda + 1 \tag{2}$$

where λ is the largest positive integer smaller than n/r. (Note that the most important relationships referred to later in this work are summed up in Appendix.)

* S. Csapó, H-5130—Jászapáti, Vöröshadsereg u. 57, Hungary

Acta Technica Academiae Scientiarum Hungaricae 98, 1985
Akadémiai Kiadó, Budapest

According to (A–3), the movement or trajectory of system (1) for initial state $\mathbf{x}(k_0) \in R^n$, assuming some control sequence

$$\mathbf{u}_{[k_0, k-1]} = \{\mathbf{u}(k_0), \mathbf{u}(k_0+1), \ldots, \mathbf{u}(k-1)\}, \tag{3}$$

can be described as

$$\mathbf{x}(k) = \boldsymbol{\varphi}_x(k; k_0, \mathbf{x}(k_0), \mathbf{u}_{[k_0, k-1]}) =$$

$$= \mathbf{F}(k, k_0)\mathbf{x}(k_0) + \sum_{i=k_0}^{k-1} \mathbf{F}(k, i+1)\mathbf{B}(i)\mathbf{u}(i), \tag{4}$$

where matrix $\mathbf{F}(k, k_0) \in R^{n \times n}$ is the fundamental matrix of system (1) according to definition (A–4). In relation with controllability index α given in (2), a final state x^F determined in advance can be reached in minimum time along a trajectory starting from initial state $\mathbf{x}(k_0)$ by a control sequence

$$\mathbf{u}_{[k_0, k_0+\lambda]} = \{\mathbf{u}(k_0), \mathbf{u}(k_0+1), \ldots, \mathbf{u}(k_0+\lambda)\} \tag{5}$$

according to relationship

$$\mathbf{x}(k_0+\lambda+1) = \mathbf{x}^F = \boldsymbol{\varphi}_x(k_0+\lambda+1; k_0, \mathbf{x}(k_0), \mathbf{u}_{[k_0, k_0+\lambda]}) =$$

$$= \mathbf{F}(k_0+\lambda+1, k_0)\mathbf{x}(k_0) +$$

$$+ \sum_{i=k_0}^{k_0+\lambda} \mathbf{F}(k_0+\lambda+1, i+1)\mathbf{B}(i)\mathbf{u}(i). \tag{6}$$

Hence, in this work, the task is to produce the control sequence (5) that satisfies condition (6) formulated for minimum–time systems.

2. Results

Assume that, starting from an initial state $\mathbf{x}(k_0)$ we have arrived along trajectory (4) at a state $\mathbf{x}(k)$ where $k_0+1 \leqq k < k_0+\lambda+1$. Thus, for the second section of the trajectory defined by states $\mathbf{x}(k)$ and $\mathbf{x}^F$, relationship

$$\mathbf{x}(k_0+\lambda+1) = \mathbf{x}^F = \boldsymbol{\varphi}_x(k_0+\lambda+1; k, \mathbf{x}(k), \mathbf{u}_{[k, k_0+\lambda]}) =$$

$$= \mathbf{F}(k_0+\lambda+1, k)\mathbf{x}(k) + \sum_{i=k}^{k_0+\lambda} \mathbf{F}(k_0+\lambda+1, i+1)\mathbf{B}(i)\mathbf{u}(i) \tag{7}$$

can be written. Using formula (A–9) applying to the inverse of the fundamental matrix, (7) can be written as follows:

$$\mathbf{x}(k) - \mathbf{F}(k, k_0+\lambda+1)\mathbf{x}^F = \sum_{i=k}^{k_0+\lambda} -\mathbf{F}(k, i+1)\mathbf{B}(i)\mathbf{u}(i). \tag{8}$$

Let vector $\mathbf{w}(k)n \times 1$ be introduced by means of relationship

$$\mathbf{w}(k) \triangleq \mathbf{x}(k) - \mathbf{F}(k, k_0 + \lambda + 1)\mathbf{x}^F, \qquad k_0 \leqq k \leqq k_0 + \lambda + 1 \tag{9}$$

to designate the left side of (8). Obviously, zero vector will be obtained for value $k = k_0 + \lambda + 1$ in (9):

$$\mathbf{w}(k_0 + \lambda + 1) = \mathbf{x}(k_0 + \lambda + 1) - \mathbf{F}(k_0 + \lambda + 1, k_0 + \lambda + 1)\mathbf{x}^F = \mathbf{0}. \tag{10}$$

On the other hand, the initial value of vector $\mathbf{w}(k)$ (9):

$$\mathbf{w}(k_0) = \mathbf{x}(k_0) - \mathbf{F}(k_0, k_0 + \lambda + 1)\mathbf{x}^F. \tag{11}$$

Accordingly, on the basis of (9), state sequence

$$\mathbf{w}(k_0), \mathbf{w}(k_0 + 1), \ldots, \mathbf{w}(k_0 + \lambda), \mathbf{w}(k_0 + \lambda + 1) = \mathbf{0} \tag{12}$$

defined by initial state (11) and final state (10) having zero vector value can be defined. For space of states (12), we introduce space of states W^n of dimension n. It is also obvious that it is state sequence

$$\mathbf{x}(k_0), \mathbf{x}(k_0 + 1), \ldots, \mathbf{x}(k_0 + \lambda), \mathbf{x}(k_0 + \lambda + 1) = \mathbf{x}^F$$

that in space of states R^n corresponds to state sequence $\mathbf{w}(k) \in W^n$ $(k = k_0, k_0 + 1, \ldots, k_0 + \lambda + 1)$ given in (12).

For transients $\mathbf{w}(k+1)$ of vectors $\mathbf{w}(k) \in W^n$ $(k = k_0, k_0 + 1, \ldots, k_0 + \lambda)$ (12), assume a fictive system

$$\mathbf{w}(k+1) = \mathbf{A}(k)\mathbf{w}(k) + \mathbf{B}(k)\mathbf{u}(k), \qquad \text{initial state} \quad \mathbf{w}(k_0) \tag{13}$$

characterized by state matrices $\mathbf{A}(k)$ and $\mathbf{B}(k)$ of controlled system (1) where control vector $\mathbf{u}(k) \in R^r$ of system (1) is though to be the input vector. Now we show that initial state $\mathbf{w}(k_0) \in W^n$ (11) is brought by control sequence (5) carrying initial state $\mathbf{x}(k_0) \in R^n$ into a final state $\mathbf{x}^F$ determined in advance into the origin of space W^n according to discrete state equation (13).

Let the left side, then the right side of (8) be substituted into (13), and $\mathbf{F}(k+1, k) = \mathbf{A}(k)$ given in (A–7) be taken into consideration. In this way, relationships

$$\mathbf{w}(k+1) = \mathbf{A}(k)\mathbf{x}(k) - \mathbf{F}(k+1, k_0 + \lambda + 1)\mathbf{x}^F + \mathbf{B}(k)\mathbf{u}(k) =$$

$$= \mathbf{x}(k+1) - \mathbf{F}(k+1, k_0 + \lambda + 1)\mathbf{x}^F, \tag{14}$$

and

$$\mathbf{w}(k+1) = \mathbf{A}(k) \sum_{i=k}^{k_0+\lambda} - \mathbf{F}(k, i+1)\mathbf{B}(i)\mathbf{u}(i) + \mathbf{B}(k)\mathbf{u}(k) =$$

$$= \sum_{i=k+1}^{k_0+\lambda} - \mathbf{F}(k+1, i+1)\mathbf{B}(i)\mathbf{u}(i), \tag{15}$$

are obtained, respectively, where $k = k_0, k_0 + 1, \ldots, k_0 + \lambda$. (14) and (15) are obviously

equal to each other:

$$\mathbf{x}(k+1)-\mathbf{F}(k+1, k_0+\lambda+1)\mathbf{x}^F=\sum_{i=k+1}^{k_0+\lambda}-\mathbf{F}(k+1, i+1)\mathbf{B}(i)\mathbf{u}(i). \tag{16}$$

(16) can be derived also from (8) since (8) will be true also if $(k+1)$ is written in place of k in the equation. Assuming an initial state $\mathbf{w}(k_0)$ and control sequence (3), the trajectory of fictive system (13):

$$\mathbf{w}(k) \triangleq \boldsymbol{\varphi}_w(k; k_0, \mathbf{w}(k_0), \mathbf{u}_{[k_0, k-1]}) = \mathbf{F}(k, k_0)\mathbf{w}(k_0) +$$

$$+\sum_{i=k_0}^{k-1}\mathbf{F}(k, i+1)\mathbf{B}(i)\mathbf{u}(i), \qquad k_0 \leqq k \leqq k_0+\lambda+1. \tag{17}$$

In the sense of 10, trajectory (17) for $k=k_0+\lambda+1$ has to arrive at the origin of space W^n:

$$\mathbf{w}(k_0+\lambda+1)=\mathbf{0}=\boldsymbol{\varphi}_w(k_0+\lambda+1; k_0, \mathbf{w}(k_0), \mathbf{u}_{[k_0, k+\lambda]})=$$

$$=\mathbf{F}(k_0+\lambda+1, k_0)\left[\mathbf{w}(k_0)+\sum_{i=k_0}^{k_0+\lambda}\mathbf{F}(k_0, i+1)\mathbf{B}(i)\mathbf{u}(i)\right]. \tag{18}$$

Since $\mathbf{F}(k_0+\lambda+1, k_0)\mathbf{w}(k_0)\neq\mathbf{0}$, equality (18) can be fulfilled only if the value of the term in brackets is zero vector. As it is this very condition that is supplied by the relationship (8) written for $k=k_0$, the equality is fulfilled. Thus result (18) reveals the fact that initial state $\mathbf{w}(k_0)\in W^n$ (11) of fictive system (13) is carried into the origin of space W^n by the control sequence (5) carrying an initial state $\mathbf{x}(k_0)\in R^n$ of controlled system (1) into some final state $\mathbf{x}^F$ determined in advance.

Accordingly, the calculation of control sequence (5) to be determined, fulfilling condition (6), can be traced back also to space W^n. Obviously, control sequence $\mathbf{u}_{[k_0, k_0+\lambda]}$ (5) must be a minimum–time dead-beat sequence (MTDB) for initial state $\mathbf{w}(k_0)$ (11). Note that a possible way of producing such a sequence is one fulfilling condition (18) has been discussed in [1], in particular for the class of systems considered in the present work.

On the basis of [1], now the MTDB control law shall be defined for fictive system (13) in the following shape:

$$\mathbf{u}(k)=\mathbf{K}(k)\mathbf{w}(k), \qquad k=k_0, k_0+1, \ldots, k_0+\lambda \tag{19}$$

where matrix $\mathbf{K}(k)\in R^{r\times n}$—state feedback matrix for fictive system (13).

By substituting (19) into (13) we obtain a homogeneous state equation

$$\mathbf{w}(k+1)=[\mathbf{A}(k)+\mathbf{B}(k)\mathbf{K}(k)]\mathbf{w}(k), \qquad k_0 \leqq k \leqq k_0+\lambda \tag{20}$$

which then, by introducing matrix

$$\mathbf{W}(k)=\mathbf{A}(k)+\mathbf{B}(k)\mathbf{K}(k) \tag{21}$$

of dimension $n\times n$, can be simplified:

$$\mathbf{w}(k+1)=\mathbf{W}(k)\mathbf{w}(k), \qquad \text{initial state } \mathbf{w}(k_0). \tag{22}$$

For an initial state $\mathbf{w}(k_0) \in W^n$, the trajectory of system (22)

$$\mathbf{w}(k) \triangleq \boldsymbol{\varphi}_w(k; k_0, \mathbf{w}(k_0)) = \mathbf{F}_w(k, k_0)\mathbf{w}(k_0), \tag{23}$$

where $\mathbf{F}_w(k, k_0) \in R^{n \times n}$ is the fundamental matrix of (22) according to definition

$$\mathbf{F}_w(k, k_0) \triangleq \begin{cases} \prod_{j=k_0}^{k-1} \mathbf{W}(j) = \mathbf{W}(k-1)\mathbf{W}(k-2)\ldots\mathbf{W}(k_0), & \text{if} \quad k > k_0 \\ \mathbf{I}, & \text{if} \quad k = k_0 \\ \text{non-defined}, & \text{if} \quad k < k_0. \end{cases} \tag{24}$$

Obviously, (23) can be produced also on the basis of (17) if control law (19) is substituted into (17). State $\mathbf{w}(k)$ (23) must be zero vector for $k = k_0 + \lambda + 1$:

$$\mathbf{w}(k_0 + \lambda + 1) = \mathbf{0} = \boldsymbol{\varphi}_w(k_0 + \lambda + 1; k_0, \mathbf{w}(k_0)) = \mathbf{F}_w(k_0 + \lambda + 1, k_0)\mathbf{w}(k_0).$$

For an initial state $\mathbf{w}(k_0) \neq \mathbf{0}$ of non-zero vector value, this condition will exist only if equality

$$\mathbf{F}_w(k_0 + \lambda + 1, k_0) = [\mathbf{A}(k_0 + \lambda) + \mathbf{B}(k_0 + \lambda)\mathbf{K}(k_0 + \lambda)]\,[\mathbf{A}(k_0 + \lambda - 1) + \\ + \mathbf{B}(k_0 + \lambda - 1)\mathbf{K}(k_0 + \lambda - 1)]\ldots[\mathbf{A}(k_0) + \mathbf{B}(k_0)\mathbf{K}(k_0)] = \mathbf{0} \tag{25}$$

is fulfilled. As has been shown in [1], matrices $\mathbf{K}(k) \in R^{r \times n}$ satisfying equation (25) can be calculated on the basis of relationship

$$\mathbf{K}(k) \triangleq \begin{bmatrix} -\mathbf{T}_2(k + \lambda)\mathbf{F}(k + \lambda, k) \\ -\mathbf{T}_1(k + \lambda + 1)\mathbf{F}(k + \lambda + 1, k) \end{bmatrix}, \qquad k_0 \leqq k \leqq k_0 + \lambda \tag{26}$$

where matrices $\mathbf{T}_1(\cdot) \in R^{\beta \times n}$ and $\mathbf{T}_2(\cdot) \in R^{\gamma \times n}$, respectively, are given in (A–23), if the necessary and sufficient conditions, (A–20) and (A–21), of transformability into phase variable block canonical form (A–15) are fulfilled for controlled system (1). By substituting state $\mathbf{w}(k)$ (9) into control law (19), taking into consideration (26), we obtain for minimum-time systems state control law

$$\mathbf{u}(k) = \mathbf{K}(k)\mathbf{x}(k) + \mathbf{S}(k)\mathbf{x}^F, \qquad k = k_0, k_0 + 1, \ldots, k_0 + \lambda \tag{27}$$

where for $k = k_0, k_0 + 1, \ldots, k_0 + \lambda$, the expression for time-varying coefficient matrices $\mathbf{S}(k) \in R^{r \times n}$:

$$\mathbf{S}(k) \triangleq -\mathbf{K}(k)\mathbf{F}(k, k_0 + \lambda + 1) = \begin{bmatrix} \mathbf{T}_2(k + \lambda)\mathbf{F}(k + \lambda, k_0 + \lambda + 1) \\ \mathbf{T}_1(k + \lambda + 1)\mathbf{F}(k + \lambda + 1, k_0 + \lambda + 1) \end{bmatrix}. \tag{28}$$

The structure of control law (27) reveals the fact that minimum-time control actually takes place by state variables feedback. By substituting control law (27) into the trajectory given in (4) for controlled system (1), we obtain the trajectory of the state-variable feedback system for an initial state $\mathbf{x}(k_0)$ and a final state $\mathbf{x}^F$ determined in

advance:

$$\mathbf{x}(k) \triangleq \boldsymbol{\varphi}_x(k; k_0, \mathbf{x}(k_0)) = \mathbf{F}_w(k, k_0)\mathbf{x}(k_0) + \\ + \sum_{i=k_0}^{k-2} \mathbf{F}_w(k, i+1)\mathbf{B}(i)\mathbf{S}(i)\mathbf{x}^F + \mathbf{B}(k-1)\mathbf{S}(k-1)\mathbf{x}^F \tag{29}$$

where $k = k_0, k_0+1, \ldots, k_0+\lambda+1$. If in (29) $k = k_0+\lambda+1$, then, according to (6), relationship

$$\mathbf{x}(k_0+\lambda+1) = \mathbf{x}^F = \mathbf{B}(k_0+\lambda)\mathbf{S}(k_0+\lambda)\mathbf{x}^F + \\ + \sum_{i=k_0}^{k_0+\lambda-1} \mathbf{F}_w(k_0+\lambda+1, i+1)\mathbf{B}(i)\mathbf{S}(i)\mathbf{x}^F \tag{30}$$

must exist as according to condition (25):

$$\mathbf{F}_w(k_0+\lambda+1, k_0) = \mathbf{W}(k_0+\lambda)\ldots\mathbf{W}(k_0+1)\mathbf{W}(k_0) = \mathbf{0}. \tag{31}$$

Obviously, (30) can be fulfilled only if equality

$$\sum_{i=k_0}^{k_0+\lambda-1} \mathbf{F}_w(k_0+\lambda+1, i+1)\mathbf{B}(i)\mathbf{S}(i) + \mathbf{B}(k_0+\lambda)\mathbf{S}(k_0+\lambda) = \mathbf{I} \tag{32}$$

exists. Assume that control vectors $\mathbf{u}^F(k) \in R^r$ required to maintain a final state $\mathbf{x}^F \in R^n$ are produced according to the formula given in (27):

$$\mathbf{u}^F(k) = \mathbf{K}(k)\mathbf{x}(k) + \mathbf{S}(k)\mathbf{x}^F, \qquad k = k_0+\lambda+1, \ldots, k_0+N \tag{33}$$

where $N \geqq \lambda+1$—positive integer.

A final state $\mathbf{x}^F$ determined in advance can obviously be maintained only by control vectors (33) if relationship $\mathbf{x}(k) = \mathbf{x}^F = \mathbf{x}_E$ exists for times $k = k_0+\lambda+1, k_0+\lambda+2, \ldots, k_0+N$, where $\mathbf{x}_E \in R^n$ expresses a possible equilibrium state of controlled system (1). Therefore, if (33) is considered to be an equilibrium control sequence, then the first to be assumed is that controlled system (1) can be brought to phase–variable block canonical form (A–15) also for extended interval $k_0+\lambda+1 \leqq k \leqq k_0+N$. On this assumption, state feedback–matrices $\mathbf{K}(k) \in R^{r\times n}$ (26) will exist also over interval $k_0 \leqq k \leqq k_0+N$:

$$\mathbf{K}(k) = \begin{bmatrix} -\mathbf{T}_2(k+\lambda)\mathbf{F}(k+\lambda, k) \\ -\mathbf{T}_1(k+\lambda+1)\mathbf{F}(k+\lambda+1, k) \end{bmatrix}, \qquad k_0 \leqq k \leqq k_0+N. \tag{34}$$

Since matrices $\mathbf{S}(k) \in R^{r\times n}$ (28) are not defined for values $k \geqq k_0+\lambda+1$, also the additional hypothesis that equality (32) will exist also if values $k_0+1, k_0+2, \ldots$, are written there in place of k_0. Hence, assume that equality

$$\sum_{i=k_0}^{k-1} \mathbf{F}_w(k+1, i+1)\mathbf{B}(i)\mathbf{S}(i) + \mathbf{B}(k)\mathbf{S}(k) = \mathbf{I} \tag{35}$$

exists for all values $k = k_0 + \lambda + 1, k_0 + \lambda + 2, \ldots, k_0 + N$. Then, assuming that there exist generalized inverse $\mathbf{B}^+(k) \in R^{r \times n}$ of input matrix $\mathbf{B}(k) \in R^{n \times r}$ for all k over interval $k = k_0 + \lambda + 1, \ldots, k_0 + N$ according to relationship [5]

$$\mathbf{B}^+(k) = [\mathbf{B}^T(k)\mathbf{B}(k)]^{-1}\mathbf{B}^T(k), \qquad k_0 + \lambda + 1 \leqq k \leqq k_0 + N \tag{36}$$

then a recursive equation is supplied by (35) in the shape given below for calculation of matrices $\mathbf{S}(k) \in R^{r \times n}$ in (33)

$$\mathbf{S}(k) = \mathbf{B}^+(k)\left[\mathbf{I} - \sum_{i=k_0}^{k-1} \mathbf{F}_w(k+1, i+1)\mathbf{B}(i)\mathbf{S}(i)\right] \tag{37}$$

where $k = k_0 + \lambda + 1, k_0 + \lambda + 2, \ldots, k_0 + N$. Note that also a non–recursive relationship can be given for calculation of matrices $\mathbf{S}(k)$ in (33) in case (34) and (36) are fulfilled. If controlled system (1) is in a possible state of equilibrium $\mathbf{x}_E$, then, by the use of (33), the following equation can be derived for $k = k_0 + \lambda + 1, \ldots, k_0 + N$:

$$\mathbf{0} = [\mathbf{A}(k) - \mathbf{I} + \mathbf{B}(k)\mathbf{K}(k)]\mathbf{x}_E + \dot{\mathbf{B}}(k)\mathbf{S}_*(k)\mathbf{x}_E. \tag{38}$$

Using (36), $\mathbf{S}_*(k)$ can be then expressed from (38):

$$\mathbf{S}_*(k) \triangleq -\mathbf{B}^+(k)[\mathbf{A}(k) - \mathbf{I} + \mathbf{B}(k)\mathbf{K}(k)], \qquad k_0 + \lambda + 1 \leqq k \leqq k_0 + N. \tag{39}$$

Taking the third–order system considered in [1] as a basis, an example is given below to demonstrate the conditions.

3. Example

Let $n = 3$ and $r = 2$ to demonstrate the conditions of movement of the state-feedback controlled system. On the basis of [1], let

$$\mathbf{A}(k) = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ -e^{-k+1} & 0 & 1 \end{bmatrix} \qquad \mathbf{B}(k) = \begin{bmatrix} 0 & 1 \\ 0 & -1 \\ 1 & e^{-k} \end{bmatrix} \tag{40}$$

be the state matrices of system (1).

For controlled system (1) described by state matrices (40), $\lambda = \beta = \gamma = 1$ and thus controllability index (2) is $\alpha = \lambda + 1 = 2$. Since the necessary and sufficient conditions (A-20) and (A-21), respectively, of transformability into phase–variable block canonical form (A-15) are fulfilled, matrix $\mathbf{T}(k)$ (A-19) of linear transformation

$$\mathbf{z}(k) = \mathbf{T}(k)\mathbf{x}(k) \tag{41}$$

given in (A-14), transforming controlled system

$$\mathbf{x}(k+1) = \mathbf{A}(k)\mathbf{x}(k) + \mathbf{B}(k)\mathbf{u}(k)$$

into phase–variable block canonical form

$$\mathbf{z}(k+1)=\bar{\mathbf{A}}(k)\mathbf{z}(k)+\bar{\mathbf{B}}\mathbf{u}(k) \tag{42}$$

is given as

$$\mathbf{T}(k)=\begin{bmatrix}\mathbf{T}_1(k)\\ \mathbf{T}_2(k)\\ \mathbf{T}_1(k+1)\mathbf{A}(k)\end{bmatrix}=\begin{bmatrix}-1 & -1 & 0\\ -e^{-k+1}-m & -m & 1\\ -1 & -2 & 0\end{bmatrix}. \tag{43}$$

Now the phase-form system matrix, $\bar{\mathbf{A}}(k)$ (A-16), of phase–variable block canonical form (42) will be

$$\bar{\mathbf{A}}(k)=\mathbf{T}(k+1)\mathbf{A}(k)\mathbf{T}^{-1}(k)=\begin{bmatrix}0 & 0 & 1\\ e^{-k}-m & 1 & m\\ -1 & 0 & 2\end{bmatrix}. \tag{44}$$

In the present case, phase–form matrix $\bar{\mathbf{B}}$ (A-17) will be:

$$\bar{\mathbf{B}}=\mathbf{T}(k+1)\mathbf{B}(k)=\begin{bmatrix}0 & 0\\ 1 & 0\\ 0 & 1\end{bmatrix}.$$

Since now $\beta=\gamma=1$, matrix $\mathbf{M}_{2\lambda+1}\in R^{\gamma\times\beta}$ (A-24) constant in time, defined as a parametric matrix, degenerates into a scalar $m_{2\lambda+1}=m$. Using matrix $\mathbf{T}(k)$ (43), state–feedback matrices $\mathbf{K}(k)$ from matrix $\bar{\mathbf{A}}(k)$ (44) can be read in the following shape [1]:

$$\mathbf{K}(k)=\begin{bmatrix}-e^{-k}+m & -1 & -m\\ 1 & 0 & -2\end{bmatrix}\mathbf{T}(k). \tag{45}$$

According to (26), a simpler way offers itself for calculation of matrix $\mathbf{K}(k)$:

$$\mathbf{K}(k)=\begin{bmatrix}-\mathbf{T}_2(k+1)\mathbf{A}(k)\\ -\mathbf{T}_1(k+2)\mathbf{A}(k+1)\mathbf{A}(k)\end{bmatrix}=$$

$$=\begin{bmatrix}e^{-k}+e^{-k+1}+m & e^{-k}+2m & -1\\ 1 & 3 & 0\end{bmatrix}. \tag{46}$$

Since any initial state $\mathbf{x}(k_0)$ of the third–order system considered is controllable at any time $k_0\geqq 0$, an initial state can be given also in a form $\mathbf{x}(k)$ where $k_0=k\geqq 0$. For the sake of simplicity, assume the value of arbitrary scalar parameter $m_{2\lambda+1}=m$ to be $m=0$.

Matrices $\mathbf{S}(k)$ (28) for any arbitrary time $k_0 = k$:

$$\mathbf{S}(k) = -\mathbf{K}(k)\mathbf{F}(k, k+2) = \begin{bmatrix} 0 & 0 & 1 \\ -1 & -1 & 0 \end{bmatrix}, \tag{47}$$

$$\mathbf{S}(k+1) = -\mathbf{K}(k+1)\mathbf{F}(k+1, k+2) = \begin{bmatrix} -e^{-k-1} & 0 & 1 \\ -1 & -2 & 0 \end{bmatrix}. \tag{48}$$

Further matrices $\mathbf{S}(k+2)$, $\mathbf{S}(k+3)$, ..., $\mathbf{S}(k+N)$ can be calculated on the basis of recursive relationship

$$\mathbf{S}(k) = \mathbf{B}^+(k)\,[\mathbf{I} - \mathbf{W}(k)\mathbf{B}(k-1)\mathbf{S}(k-1)], \qquad k \in [k+\lambda+1, k+N] \tag{49}$$

given in (37), where $N \geqq \lambda + 1$. $\mathbf{W}(k)$ in (49) is the time-varying system matrix of the state-feedback controlled system, which, assuming $m = 0$, can be written as

$$\mathbf{W}(k) = \mathbf{A}(k) + \mathbf{B}(k)\mathbf{K}(k) = \begin{bmatrix} 2 & 4 & 0 \\ -1 & -2 & 0 \\ 2e^{-k} & 4e^{-k} & 0 \end{bmatrix}, \tag{50}$$

and $\mathbf{B}^+(k)$ the generalized inverse of input matrix $\mathbf{B}(k)$ according to (36):

$$\mathbf{B}^+(k) = \frac{1}{2}\begin{bmatrix} -e^{-k} & e^{-k} & 2 \\ 1 & -1 & 0 \end{bmatrix}. \tag{51}$$

With $k + \lambda + 1 = (k+2)$ written in place of k in (49), using (48), the expression for matrix $\mathbf{S}(k+2)$:

$$\mathbf{S}(k+2) = \begin{bmatrix} -e^{-k-2} & -(1/2)e^{-k-2} & 1 \\ -1 & -7/2 & 0 \end{bmatrix}. \tag{52}$$

Writing $(k+3)$ in place of k in (49) in the knowledge of (52), matrix $\mathbf{S}(k+3)$ can be calculated:

$$\mathbf{S}(k+3) = \begin{bmatrix} -e^{-k-3} & -(5/4)e^{-k-3} & 1 \\ 1 & -23/4 & 0 \end{bmatrix}. \tag{53}$$

This recursive process can then be continued in a similar way. As an alternative way to calculate matrices $\mathbf{S}(k)$, a non-recursive relationship is given in (39). According to this method on the basis of (39),

$$\mathbf{S}_*(k) = -\mathbf{B}^+(k)\,[\mathbf{W}(k) - \mathbf{I}] = \begin{bmatrix} -e^{-k} & -(1/2)e^{-k} & 1 \\ -1 & -7/2 & 0 \end{bmatrix} \tag{54}$$

where $k \in [k+\lambda+1, k+N$. It can be seen it is the very matrix $\mathbf{S}(k+2)$ (52) that occurs if $(k+2)$ is written in place of k in (54). Let then the final state determined in advance be

$$\mathbf{x}^F = [2, \quad 1, \quad 1]^T, \tag{55}$$

where T—vector transpose.

Let initial state $\mathbf{x}(k)$ taking place at any arbitrary time k be

$$\mathbf{x}(k)=[1, \quad 4, \quad -1]^T. \tag{56}$$

According to state control law (27):

$$\mathbf{u}(k)=\mathbf{K}(k)\mathbf{x}(k)+\mathbf{S}(k)\mathbf{x}^F=[e^{-k+1}+5e^{-k}+2, \quad 10]^T,$$

and as a result, state

$$\mathbf{x}(k+1)=\mathbf{A}(k)\mathbf{x}(k)+\mathbf{B}(k)\mathbf{u}(k)=$$
$$=[15, \quad -6, \quad 15e^{-k}+1]^T$$

takes place. The next control vector $\mathbf{u}(k+1)$ will be

$$\mathbf{u}(k+1)=\mathbf{K}(k+1)\mathbf{x}(k+1)+\mathbf{S}(k+1)\mathbf{x}^F=$$
$$=[7e^{-k-1}, \quad -7]^T,$$

which then carries state $\mathbf{x}(k+1)$ into final state $\mathbf{x}^F$ (55) determined in advance:

$$\mathbf{x}(k+2)=\mathbf{x}^F=\mathbf{A}(k+1)\mathbf{x}(k+1)+\mathbf{B}(k+1)\mathbf{u}(k+1)=$$
$$=[2, \quad 1, \quad 1]^T.$$

According to (33) and using matrix $\mathbf{S}(k+2)$ (52), control vector $\mathbf{u}(k+2)$ will be

$$\mathbf{u}^F(k+2)=\mathbf{K}(k+2)\mathbf{x}(k+2)+\mathbf{S}(k+2)\mathbf{x}^F=$$
$$=[2e^{-k-1}+(1/2)e^{-k-2}, \quad -1/2]^T,$$

and as a result, we arrive from state $\mathbf{x}(k+2)=\mathbf{x}^F$ at state

$$\mathbf{x}(k+3)=\mathbf{A}(k+2)\mathbf{x}(k+2)+\mathbf{B}(k+2)\mathbf{u}(k+2)=$$
$$=[5/2, \quad 3/2, \quad 1]^T.$$

As can be seen, the trajectory starting from initial state $\mathbf{x}(k)$ (56) passes through final state $\mathbf{x}^F$ (55) determined in advance. An explanation for this fact is that none of the possible equilibrium states $\mathbf{x}_E$ of controlled system (1) complies with final state $\mathbf{x}^F$ (55) as the set of possible equilibrium states $\mathbf{x}_E$ of the system is the entire co–ordinate plane (x_1, x_3) of space of states R^3.

Let now final state $\mathbf{x}^F$ determined in advance be identical with a possible equilibrium state of system (1). Assume e.g.

$$\mathbf{x}_E=\mathbf{x}^F=[2, \quad 0, \quad 1]^T. \tag{57}$$

Let initial state $\mathbf{x}(k)$ be given according to (56) also in this case. Now the sequence of control vectors and of state:

$$\mathbf{u}(k)=[e^{-k+1}+5e^{-k}+2, \quad 11]^T$$
$$\mathbf{x}(k+1)=[16, \quad -7, \quad 16e^{-k}+1]^T$$

$$\mathbf{u}(k+1)=[7e^{-k-1}, \quad -7]^T$$

$$\mathbf{x}(k+2)=[2, \quad 0, \quad 1]^T=\mathbf{x}^F.$$

According to (33), control vector $u(k+2)$ will be

$$\mathbf{u}(k+2)=\mathbf{u}^F(k+2)=\mathbf{K}(k+2)\mathbf{x}(k+2)+\mathbf{S}(k+2)\mathbf{x}^F=$$

$$=[2e^{-k-1}, \quad 0]^T, \tag{58}$$

and as a result, again final state $\mathbf{x}^F$ (57) takes place:

$$\mathbf{x}(k+3)=\mathbf{A}(k+2)\mathbf{x}(k+2)+\mathbf{B}(k+2)\mathbf{u}(k+2)=$$

$$=[2, \quad 0, \quad 1]^T=\mathbf{x}^F.$$

The next control vector according to (33), taking into consideration matrix $\mathbf{S}(k+1)$ (53)

$$\mathbf{u}(k+3)=\mathbf{u}^F(k+3)=\mathbf{K}(k+3)\mathbf{x}(k+3)+\mathbf{S}(k+3)\mathbf{x}^F=$$

$$=[2e^{-k-2}, \quad 0]^T \tag{59}$$

which brings about again final state (57) $\mathbf{x}^F$. It is therefore obvious that the state–feedback system will not leave final state $\mathbf{x}^F$ (57). Accordingly, the rule of formation of equilibrium control vectors:

$$\mathbf{u}(k+j)=\mathbf{u}^F(k+j)=\mathbf{K}(k+j)\mathbf{x}(k+j)+\mathbf{S}(k+j)\mathbf{x}^F=$$

$$=[2e^{-k-j+1}, \quad 0]^T, \tag{60}$$

where $\mathbf{x}(k+j)=\mathbf{x}^F=\mathbf{x}_E$ for all values $j \geqq N$.

It can be seen in relation with the next example that matrices $\mathbf{S}_*(k)$ given in (54) prove definable also for values $k, k+1, \ldots, k+\lambda$. Accordingly, matrices $\mathbf{S}(k)$ (47) in control law (27) can be replaced with matrices $\mathbf{S}_*(k)$ (54):

$$\mathbf{u}(k)=\mathbf{K}(k)\mathbf{x}(k)+\mathbf{S}_*(k)\mathbf{x}^F, \qquad k \in [k, k+N]. \tag{61}$$

Let the initial state and the final state be given according to (56) and (55), respectively, also in this case. Now the sequence of control vectors and state vectors will be

$$\mathbf{u}(k)=[e^{-k+1}+3e^{-k}+2, \quad 11]^T,$$

$$\mathbf{x}(k+1)=[16, \quad -7, \quad 14e^{-k}+1]^T,$$

$$\mathbf{u}(k+1)=[7e^{-k-1}+2e^{-k}, \quad -7]^T,$$

$$\mathbf{x}(k+2)=[2, \quad 0, \quad 1]^T=\mathbf{x}^F,$$

$$\mathbf{u}(k+2)=[2e^{-k-1}, \quad 0]^T=\mathbf{u}^F(k+2),$$

$$\mathbf{x}(k+3)=[2, \quad 0, \quad 1]^T=\mathbf{x}^F.$$

A result complying with what has been obtained earlier will be obtained if the process is continued in further steps.

Let now row–matrix $\mathbf{T}_2(k)$ of transformation matrix $\mathbf{T}(k)$ (43) be determined on the basis of relationship (A-28). According to the definition:

$$\hat{\mathbf{T}}_2(k)=[-(1/2)e^{-k+1}, \quad (1/2)e^{-k+1}, \quad 1]. \tag{62}$$

Now the transformation matrix:

$$\hat{\mathbf{T}}(k)=\begin{bmatrix}\mathbf{T}_1(k)\\ \hat{\mathbf{T}}_2(k)\\ \mathbf{T}_1(k+1)\mathbf{A}(k)\end{bmatrix}=\begin{bmatrix}-1 & -1 & 0\\ -(1/2)e^{-k+1} & (1/2)e^{-k+1} & 1\\ -1 & -2 & 0\end{bmatrix}. \tag{63}$$

Now phase–form system matrix $\bar{\mathbf{A}}(k)$ (A-16) will be

$$\hat{\bar{\mathbf{A}}}(k)=\hat{\mathbf{T}}(k+1)\mathbf{A}(k)\hat{\mathbf{T}}^{-1}(k)=$$

$$=\begin{bmatrix}0 & 0 & 1\\ (1/2)e^{-k+1}+e^{-k} & 1 & -(1/2)e^{-k}\\ -1 & 0 & 2\end{bmatrix}. \tag{64}$$

By the use of (62), feedback matrices $\hat{\mathbf{K}}(k)$:

$$\hat{\mathbf{K}}(k)=\begin{bmatrix}-\hat{\mathbf{T}}_2(k+1)\mathbf{A}(k)\\ -\mathbf{T}_1(k+2)\mathbf{A}(k+1)\mathbf{A}(k)\end{bmatrix}=$$

$$=\begin{bmatrix}e^{-k+1}+(1/2)e^{-k} & 0 & -1\\ 1 & 3 & 0\end{bmatrix}. \tag{65}$$

Matrices $\hat{\mathbf{S}}(k)$ can be produced also for matrices $\hat{\mathbf{K}}(k)$ (65) in a similar way. Now the state control law takes the following shape:

$$\mathbf{u}(k)=\hat{\mathbf{K}}(k)\mathbf{x}(k)+\hat{\mathbf{S}}(k)\mathbf{x}^F, \qquad k\in[k, k+N] \tag{66}$$

where $N\geqq\lambda+1$. A result similar to those obtained earlier will be obtained also by the use of (66).

4. Conclusions

In this work, a class of linear, discrete, controlled systems of state n and input $r(n\geqq r)$, varying with time, has been considered, where quotient n/r is non-integer. However, it is to emphasize that even within the class mentioned, we restricted

ourselves to the special case when the necessary and sufficient conditions of transformability into phase–variable block canonical form were fulfilled [2]. The applicability of the dimensioning method presented assumes the existence of a system state equation of phase–variable block canonical form. It was shown that the development of the minimum–time state control law was directly based on results laid down in [1].

The required state control law has been obtained in the shape of $\mathbf{u}(k)=\mathbf{K}(k)\mathbf{x}(k)+\mathbf{S}(k)\mathbf{x}^F$ for values $k=k_0, k_0+1, \ldots, k_0+\lambda$, where λ is the largest positive integer still smaller than n/r. Hence $\mathbf{K}(k)\in R^{r\times n}$ is the state–feedback matrix while matrix $\mathbf{S}(k)\in R^{r\times n}$ can be considered to be a variable amplification for input vector $\mathbf{x}^F\in R^n$ of the state–feedback system. It was seen that an initial state $\mathbf{x}(k_0)\in R^n$ of the controlled system could be carried into a final state $\mathbf{x}^F$, determined in advance, in minimum time by control steps $\mathbf{u}(k_0), \mathbf{u}(k_0+1), \ldots, \mathbf{u}(k_0+\lambda)$ the number of which complying with the value of controllability index $\alpha=\lambda+1$ according to relationship $\mathbf{x}(k_0+\lambda+1)=\mathbf{x}^F$. Also a hypothesis was necessary to derive equilibrium control vectors $\mathbf{u}^F(k)=\mathbf{K}(k)\mathbf{x}(k)+\mathbf{S}(k)\mathbf{x}^F(k=k_0+\lambda+1, k_0+\lambda+2, \ldots,)$ maintaining final state x^F in compliance with a possible equilibrium state $\mathbf{x}_E$ of the controlled system. As a result of the hypothesis, a recursive relationship was obtained for calculation of time–varying matrices $\mathbf{S}(k)\in R^{r\times n}(k=k_0+\lambda+1, k_0+\lambda+2, \ldots,)$. On the basis of an example given for third–order systems, it was seen that the hypothesis had been rightly assumed in respect of the numerical results, at least as far as the conditions arisen after the example were concerned.

Finally, note that in case there exists relationship $\mathbf{x}^F=\mathbf{0}$ for the final state determined in advance, the minimum–time dead–beat control problem outlined in [1] occurs for some initial state $\mathbf{x}(k_0)$.

Appendix

Given in Appendix are the necessary and sufficient conditions of transformability into phase–variable block canonical form for the studied class of controlled systems, taking the theorems developed in [2] as a basis. The solution of the discrete state equation describing the controlled system and the most important properties of the fundamental matrix are also discussed.

1. Solution of the discrete state equation

Let the time-varying multivariable, linear, discrete–time controlled system over set R of real numbers be given by discrete state equation

$$\mathbf{x}(k+1)=\mathbf{A}(k)\mathbf{x}(k)+\mathbf{B}(k)\mathbf{u}(k), \qquad \text{initial state} \quad \mathbf{x}(k_0) \qquad \text{(A–1)}$$

where $k \in Z$ integers,
$\mathbf{x}(k) \in R^n$ state vector,
$\mathbf{u}(k) \in R^r$ control vector,
$\mathbf{A}(k) \in R^{n \times n}$ and
$\mathbf{B}(k) \in R^{n \times r}$ properly dimensioned state matrices,
R^n and R^r Eucledian spaces, R^n being spaces of states while R^r control space.

For an initial state $\mathbf{x}(k_0) \in R^n$, assuming a control sequence

$$\mathbf{u}_{[k_0, k-1]} = \{\mathbf{u}(k_0), \mathbf{u}(k_0+1), \ldots, \mathbf{u}(k-1)\}, \tag{A–2}$$

the movement or the trajectory of the controlled system (A–1) can be given as

$$\mathbf{x}(k) \triangleq \varphi_x(k; k_0, \mathbf{x}(k_0), \mathbf{u}_{[k_0, k-1]}) = \mathbf{F}(k, k_0)\mathbf{x}(k_0) + \\ + \sum_{i=k_0}^{k-1} \mathbf{F}(k, i+1)\mathbf{B}(i)\mathbf{u}(i), \tag{A–3}$$

[3], [4], where the expression for fundamental matrix $\mathbf{F}(k, k_0) \in R^{n \times n}$:

$$\mathbf{F}(k, k_0) \triangleq \begin{cases} \prod_{j=k_0}^{k-1} \mathbf{A}(j) = \mathbf{A}(k-1)\mathbf{A}(k-2) \ldots \mathbf{A}(k_0), & \text{if} \quad k > k_0 \\ \mathbf{I}, & \text{if} \quad k = k_0 \\ \text{non-defined}, & \text{if} \quad k < k_0. \end{cases} \tag{A–4}$$

Here should the most important properties of the fundamental matrix be enhanced [3]. According to definition (A–4), the fundamental matrix complies with unit matrix $\mathbf{I} \in R^{n \times n}$ in case of $k = k_0$:

$$\mathbf{F}(k_0, k_0) = \mathbf{I}. \tag{A–5}$$

The fundamental matrix satisfies homogeneous equation ($\mathbf{u}(k) = \mathbf{0}(k)$) of (A–1):

$$\mathbf{F}(k+1, k_0) = \mathbf{A}(k)\mathbf{F}(k, k_0). \tag{A–6}$$

It follows from (A–5) and (A–6) that

$$\mathbf{F}(k_0+1, k_0) = \mathbf{A}(k_0), \qquad \mathbf{F}(k+1, k) = \mathbf{A}(k). \tag{A–7}$$

According to the group character of the fundamental matrix

$$\mathbf{F}(k_2, k_0) = \mathbf{F}(k_2, k_1)\mathbf{F}(k_1, k_0), \tag{A–8}$$

for all k_0, k_1 and k_2. It is easy to prove also the theorem concerning the inverse of the fundamental matrix, according to which

$$\mathbf{F}(k_1, k_2) = \mathbf{F}^{-1}(k_2, k_1), \qquad k_2 > k_1. \tag{A–9}$$

Namely, on the basis of (A–8),

$$\mathbf{F}(k_2, k_1)\mathbf{F}(k_1, k_2) = \mathbf{F}(k_2, k_2) = \mathbf{I}$$

will follow if $k_0 = k_2$. With this equation multiplied with inverse matrix $\mathbf{F}^{-1}(k_2, k_1)$ from the left side, actually (A–9) is obtained.

2. Phase-variable block canonical form for a class of systems

For the class of systems here considered, n/r is non–integer. Let the definition of positive integers λ, β, and γ be introduced by means of relationships

$$\beta = n - \lambda r, \qquad 0 < \beta < r, \tag{A–10a}$$

$$\gamma = r - \beta, \qquad 0 < \gamma < r \tag{A–10b}$$

[1, 2], λ being the largest positive integer smaller than n/r. Controllability index α of system (A–1) described by relationships (A–10):

$$\alpha = \lambda + 1\,. \tag{A–11}$$

Note that controllability index α is defined as the least positive integer [3] for which the rank of controllability matrix

$$\mathbf{Q}_{c,\alpha}(k_0) = [\mathbf{B}(k_0 + \alpha - 1), \mathbf{F}(k_0 + \alpha, k_0 + \alpha - 1)\mathbf{B}(k_0 + \alpha - 2), \ldots$$

$$\ldots, \mathbf{F}(k_0 + \alpha, k_0 + 1)\mathbf{B}(k_0)] \tag{A–12}$$

is n. Assume that a state $\mathbf{x} \in R^n$ of system (A–1) is controllable at initial time k_0 i.e.

$$\operatorname{rank} \mathbf{Q}_{c,\alpha}(k_0) = n\,. \tag{A–13}$$

In the present case, system (A–1) shall be transformed into phase–variable block canonical form for interval $k_0 \leqq k \leqq k_0 + \lambda$. Let a variable parameter linear transformation be introduced by means of relationship [1, 2]

$$\mathbf{z}(k) = \mathbf{T}(k)\mathbf{x}(k)\,, \tag{A–14}$$

where $\mathbf{T}(k) \in R^{n \times n}$ is a non-singular transformation matrix invariable in time. Transformation (A–14) carries discrete state equation (A–1) into phase-variable block canonical form

$$\mathbf{z}(k+1) = \bar{\mathbf{A}}(k)\mathbf{z}(k) + \bar{\mathbf{B}}\mathbf{u}(k), \qquad k_0 \leqq k \leqq k_0 + \lambda \tag{A–15}$$

where time–varying phase–form system matrix $\bar{\mathbf{A}}(k) \in R^{n \times n}$ and phase–form input

matrix $\bar{\mathbf{B}} \in R^{n \times r}$ are:

$$\bar{\mathbf{A}}(k) = \mathbf{T}(k+1)\mathbf{A}(k)\mathbf{T}^{-1}(k) =$$

$$= \begin{bmatrix} \mathbf{0}_{\beta,\beta} & \mathbf{0}_{\beta,\gamma} & \mathbf{I}_{\beta,\beta} & \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} & \dots & \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} \\ \mathbf{0}_{\gamma,\beta} & \mathbf{0}_{\gamma,\gamma} & \mathbf{0}_{\gamma,\beta} & \mathbf{I}_{\gamma,\gamma} & \mathbf{0}_{\gamma,\beta} & \dots & \mathbf{0}_{\gamma,\gamma} & \mathbf{0}_{\gamma,\beta} \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0}_{\beta,\beta} & \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} & \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} & \dots & \mathbf{0}_{\beta,\gamma} & \mathbf{I}_{\beta,\beta} \\ \mathbf{M}_1(k) & \mathbf{M}_2(k) & \mathbf{M}_3(k) & \mathbf{M}_4(k) & \mathbf{M}_5(k) & \dots & \mathbf{M}_{2\lambda}(k) & \mathbf{M}_{2\lambda+1}(k) \\ \mathbf{N}_1(k) & \mathbf{N}_2(k) & \mathbf{N}_3(k) & \mathbf{N}_4(k) & \mathbf{N}_5(k) & \dots & \mathbf{N}_{2\lambda}(k) & \mathbf{N}_{2\lambda+1}(k) \end{bmatrix}, \quad \text{(A–16)}$$

and

$$\bar{\mathbf{B}} = \mathbf{T}(k+1)\mathbf{B}(k) = \begin{bmatrix} \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} \\ \mathbf{0}_{\gamma,\gamma} & \mathbf{0}_{\gamma,\beta} \\ \vdots & \vdots \\ \mathbf{0}_{\beta,\gamma} & \mathbf{0}_{\beta,\beta} \\ \mathbf{I}_{\gamma,\gamma} & \mathbf{0}_{\beta,\beta} \\ \mathbf{0}_{\beta,\gamma} & \mathbf{I}_{\beta,\beta} \end{bmatrix}, \quad \text{(A–17)}$$

respectively.

In (A–16) and (A–17), $\mathbf{0}_{p,q}$ is a zero matrix, $\mathbf{I}_{p,p} \in R^{p \times p}$ a unit matrix, and

$$\mathbf{M}_i(k) \in R^{\gamma \times \beta}, \mathbf{N}_i(k) \in R^{\beta \times \beta}, \qquad i = 1, 3, \dots, 2\lambda+1 \quad \text{(A–18a)}$$

$$\mathbf{M}_i(k) \in R^{\gamma \times \gamma}, \mathbf{N}_i(k) \in R^{\beta \times \gamma}, \qquad i = 2, 4, \dots, 2\lambda. \quad \text{(A–18b)}$$

According to [1], transformation matrix $\mathbf{T}(k) \in R^{n \times n}$:

$$\mathbf{T}(k) = \begin{bmatrix} \mathbf{T}_1(k) \\ \mathbf{T}_2(k) \\ \vdots \\ \mathbf{T}_1(k+\lambda-1)\mathbf{F}(k+\lambda-1, k) \\ \mathbf{T}_2(k+\lambda-1)\mathbf{F}(k+\lambda-1, k) \\ \mathbf{T}_1(k+\lambda)\mathbf{F}(k+\lambda, k) \end{bmatrix} \qquad k_0 \leqq k \leqq k_0+\lambda+1 \quad \text{(A–19)}$$

where $\mathbf{T}_1(\cdot) \in R^{\beta \times n}$ and $\mathbf{T}_2(\cdot) \in R^{\gamma \times n}$. Thus system (A–1) can be transformed into

phase–variable block canonical form only if necessary and sufficient conditions

$$\text{range}\,[\mathbf{B}(k-1), \mathbf{F}(k,k-1)\mathbf{B}(k-2), \ldots, \mathbf{F}(k,k-\lambda+2)\mathbf{B}(k-\lambda+1),$$
$$\mathbf{F}(k,k-\lambda+1)\mathbf{B}(k-\lambda)] \supseteq \text{range}\,[\mathbf{F}(k,k-\lambda)\mathbf{B}_1(k-\lambda-1) \tag{A–20}$$

$$\text{rank}\,[\mathbf{B}(k-1), \mathbf{F}(k,k-1)\mathbf{B}(k-2), \ldots, \mathbf{F}(k,k-\lambda+2)\mathbf{B}(k-\lambda+1)$$
$$\mathbf{F}(k,k-\lambda+1)\mathbf{B}(k-\lambda), \mathbf{F}(k,k-\lambda)\mathbf{B}_2(k-\lambda-1)] = n \tag{A–21}$$

are fulfilled for all k in interval $k_0 \leqq k \leqq k_0 + 2\lambda + 1$ [2]. Submatrices $\mathbf{B}_1(\cdot) \in R^{n \times \gamma}$ and $\mathbf{B}_2(\cdot) \in R^{n \times \beta}$ in (A–20) and (A–21) respectively, result from partition of input matrix $\mathbf{B}(\cdot) \in R^{n \times r}$ in the following shape:

$$\mathbf{B}(\cdot) = [\mathbf{B}_1(\cdot), \quad \mathbf{B}_2(\cdot)]\,. \tag{A–22}$$

Submatrices $\mathbf{T}_1(\cdot) \in R^{\beta \times n}$ and $\mathbf{T}_2(\cdot) \in R^{\gamma \times n}$ occurring in transformation matrix $\mathbf{T}(k) \in R^{n \times n}$ (A–19) can be calculated on the basis of the following relationships[2]:

$$\mathbf{T}_1(k) = [\mathbf{0}_{\beta,\lambda r}, \mathbf{I}_{\beta,\beta}]\,[\mathbf{B}(k-1), \mathbf{F}(k,k-1)\mathbf{B}(k-2), \ldots$$
$$\ldots, \mathbf{F}(k,k-\lambda+1)\mathbf{B}(k-\lambda), \mathbf{F}(k,k-\lambda)\mathbf{B}_2(k-\lambda-1)]^{-1}$$
$$k = k_0, k_0+1, \ldots, k_0+2\lambda+1 \tag{A–23a}$$

$$\mathbf{T}_2(k) = [\mathbf{0}_{\gamma,(\lambda-1)r}, \mathbf{I}_{\gamma,\gamma}, \mathbf{0}_{\gamma,\beta}, \mathbf{M}_{2\lambda+1}]\,[\mathbf{B}(k-1), \mathbf{F}(k,k-1)\mathbf{B}(k-2),$$
$$\ldots, \mathbf{F}(k,k-\lambda+1)\mathbf{B}(k-\lambda), \mathbf{F}(k,k-\lambda)\mathbf{B}_2(k-\lambda-1)]^{-1}$$
$$k = k_0, k_0+1, \ldots, k_0+2\lambda \tag{A–23b}$$

Matrix $\mathbf{M}_{2\lambda+1} \in R^{\gamma \times \beta}$ invariable in time, appearing in (A–23b), can be defined as a parametric matrix since its elements of number $\gamma \cdot \beta$ can be selected optionally, independently of each other. Accordingly, because of the non–uniqueness of transformation matrix $\mathbf{T}(k)$ (A–19), the phase–variable block canonical form given in (A–15) is not unique either.

Note that the following additional statements complete the results arrived at in [2]. With matrix equation

$$\mathbf{M}_{2\lambda+1} = \mathbf{T}_2(k)\mathbf{F}(k,k-\lambda)\mathbf{B}_2(k-\lambda-1) = \text{const.}\,, \tag{A–24}$$

omitted from (A–23b), the number of scalar equations included in the system of matrix equations remaining over is one equation less as compared with number $\gamma \cdot n$ of unknowns occurring in matrix $\mathbf{T}_2(k) \in R^{\gamma \times n}$. Therefore, solutions $\mathbf{T}_2(k)$ of infinite number are possible to equation (A–23b) nonincluding (A–24). Equation (A–23b) takes now the following shape:

$$\mathbf{T}_2(k)\mathbf{P}(k) = [\mathbf{0}_{\gamma,(\lambda-1)r}, \mathbf{I}_{\gamma,\gamma}, \mathbf{0}_{\gamma,\beta}]\,, \qquad k_0 \leqq k \leqq k_0 + 2\lambda \tag{A–25}$$

where $\mathbf{P}(k) \in R^{n \times \lambda r}$ is a rectangular matrix:

$$\mathbf{P}(k) = [\mathbf{B}(k-1), \mathbf{F}(k, k-1)\mathbf{B}(k-2), \ldots, \mathbf{F}(k, k-\lambda+1)\mathbf{B}(k-\lambda)] . \qquad (A-26)$$

However, from among solutions $\mathbf{T}_2(k)$ of infinite number to (A–25), a well defined solution can still be selected, namely by means of the generalized inverse

$$\mathbf{P}^+(k) = [\mathbf{P}^T(k)\mathbf{P}(k)]^{-1}\mathbf{P}^T(k) \in R^{\lambda r \times n}, \qquad k_0 \leqq k \leqq k_0 + 2\lambda \qquad (A-27)$$

of matrix $\mathbf{P}(k)$ (A–26) that can be produced in accordance with [5] in the following way:

$$\hat{\mathbf{T}}_2(k) = [\mathbf{0}_{\gamma,(\lambda-1)r}, \mathbf{I}_{\gamma,\gamma}, \mathbf{0}_{\gamma,\beta}]\mathbf{P}^+(k), \qquad k_0 \leqq k \leqq k_0 + 2\lambda \qquad (A-28)$$

It can be seen that identity will be obtained if (A–28) is substituted into (A–25) because $\mathbf{P}^+(k)\mathbf{P}(k) = \mathbf{I} \in R^{\lambda r \times \lambda r}$. Hence, a condition for the existence of submatrix $\hat{\mathbf{T}}_2(k)$ **(A–28)** is the existence of generalized inverse $\mathbf{P}^+(k)$ (A–27) in interval $k_0 \leqq k \leqq k_0 + 2\lambda$ for all k. Note that in this case submatrix $\mathbf{M}_{2\lambda+1} \in R^{\gamma \times \beta}$ will in general not be constant in time that is usually it will depend on discrete variable k of time according to $\mathbf{M}_{2\lambda+1}(k)$. According to (A–28), transformation matrix $\hat{\mathbf{T}}(k)$ takes now the place of (A–19) in accordance with the following structure:

$$\hat{\mathbf{T}}(k) \hat{=} \begin{bmatrix} \mathbf{T}_1(k) \\ \hat{\mathbf{T}}_2(k) \\ \vdots \\ \mathbf{T}_1(k+\lambda-1)\mathbf{F}(k+\lambda-1, k) \\ \hat{\mathbf{T}}_2(k+\lambda-1)\mathbf{F}(k+\lambda-1, k) \\ \mathbf{T}_1(k+\lambda)\mathbf{F}(k+\lambda, k) \end{bmatrix} \qquad k_0 \leqq k \leqq k_0 + \lambda + 1 . \qquad (A-29)$$

References

1. Csapó, S.: Application of the minimum-time dead-beat control law to a class of the time-varying multivariable linear discrete-time systems. Acta Techn. Hung., submitted for publication
2. Csapó, S.: Transformation of the time-varying multivariable linear discrete-time systems to the phase-variable block canonical form. Acta Techn. Hung., submitted for publication
3. Kalman, R. E., Falb, P. L., Arbib, M. A.: Topics in Mathematical System Theory. McGraw-Hill, New York, 1969
4. Weiss, L.: Controllability realization and stability of discrete-time systems. Siam J. Contr. Vol. **10**, No. 2, pp. 230–251, May 1972
5. Ben-Israel and Greville, T. N. E.: Generalized Inverses, Theory and Applications. Wiley, New York, 1974

Acta Technica Academiae Scientiarum Hungaricae, 98 (3—4), pp. 251—268 (1985)

NUMERICAL METHOD FOR THE APPROXIMATE SOLUTION OF TECHNICAL PROBLEMS*

P. CSONKA**

[Received: 2 October 1984]

To solve certain technical problems we need a function which satisfies a given partial differential equation of the second order, and whose value assumes zero along the boundary of the domain concerned. This paper presents an approximate solution for these problems by replacing the governing differential equation by the corresponding difference equation and by approximately solving this difference equation. In the case of the *homogeneous* problem, the paper only presents the approximate value of the first eigenvalue and determines an approximate function for the first eigenfunction. The first eigenvalue of the difference equation is determined by using upper and lower bounds. In the case of the *inhomogeneous* problem, the results obtained for the homogeneous problem can directly be used provided that the inhomogeneous part of the difference equation is of constant sign at the internal points of the domain concerned.

1. Introduction

Many technical problems require the solution of the partial homogeneous differential equation

$$\frac{\partial^2 w}{\partial x^2}+\frac{\partial^2 w}{\partial y^2}+\lambda w=0\,, \tag{1}$$

or of the partial inhomogeneous differential equation

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}+V=0 \tag{2}$$

in the simply connected domain T, with the boundary conditions $w=0$ and $u=0$. Closed formulae can only be presented for exceptionally simple cases and consequently practice has developed different approximate methods [1, 2, 3, 4].

The aim of this paper is to present an approximate method for the above problems. Domain T is replaced by the rectangular network R of spacing $\Delta x=\Delta y=h$

* This paper is published as a historic example to show what possibilities the method of differences developed in the twenties and thirties offered for the approximate numerical solution of differential equations.

** P. Csonka, H-1114 Budapest, Bartók B. út 31, Hungary

and, instead of differential equations (1) and (2), the homogeneous difference equation

$$\frac{\Delta^2 w}{\Delta x^2}+\frac{\Delta^2 w}{\Delta y^2}+\lambda w=0 \tag{3}$$

and the inhomogeneous difference equation

$$\frac{\Delta^2 u}{\Delta x^2}+\frac{\Delta^2 u}{\Delta y^2}+V=0 \tag{4}$$

defined at the nodal points of network R are introduced. The method to be presented gives approximate solutions for these difference equations.

2. Notation

To simplify the treatment, it is expedient to use the following notations:

$$\mathscr{D}f=-\frac{h^2}{8}\left(\frac{\Delta^2 f}{\Delta x^2}+\frac{\Delta^2 f}{\Delta y^2}\right)\equiv-\frac{h^2}{8}\Delta f, \tag{5}$$

$$\mathscr{M}f=f+\frac{h^2}{4}\left(\frac{\Delta^2 f}{\Delta x^2}+\frac{\Delta^2 f}{\Delta y^2}\right)\equiv f+\frac{h^2}{4}\Delta f. \tag{6}$$

With these notations we have

$$\mathscr{D}f=\frac{1}{8}\left[4f(x,y)-f(x+h,y)-f(x,y+h)-f(x-h,y)-f(x,y-h)\right], \tag{7}$$

$$\mathscr{M}f=\frac{1}{4}\left[\qquad f(x+h,y)+f(x,y+h)+f(x-h,y)+f(x,y-h)\right]. \tag{8}$$

Further simplifying notations are

$$\mathscr{M}_0 f=f$$

$$\mathscr{M}_1 f=\mathscr{M}f$$

$$\mathscr{M}_2 f=\mathscr{M}\mathscr{M}_1 f$$

$$\mathscr{M}_3 f=\mathscr{M}\mathscr{M}_2 f$$

$$\cdots\cdots\cdots\cdots$$

and

$$v=\frac{h^2}{8}V, \qquad l'=\frac{h^2}{8}\lambda. \tag{9a, b}$$

Difference equations (3), (4) now assume the form

$$\mathscr{D}w-\lambda' w=0, \tag{10}$$

$$\mathscr{D}u-v=0. \tag{11}$$

3. Basic theorems

Through our derivations we will rely on the following four theorems based on existing analogies.

Basic Theorem 1.: Difference equation (10) with the boundary condition $w=0$ *has as many eigenfunctions* $w_1, w_2, \ldots w_n$ *as the number of the internal nodal points of network R.* Eigenvalues corresponding to these eigenfunctions are denoted by $l_1, l_2, \ldots, l_n$ where l_1 is the smallest eigenvalue in absolute value, l_2 is the next one, then l_3 etc, i.e.:

$$|l_1| \leqq |l_2| \leqq \ldots \leqq |l_n| .$$

Basic Theorem 2.: Eigenfunctions $w_1, w_2, \ldots, w_n$ *are orthogonal functions,* i.e. the expression

$$\sum_R w_i w_k = 0 , \qquad i \neq k \tag{12}$$

holds where summation must cover every internal nodal point of network R. Every other function proportional to an eigenfunction is also an eigenfunction. The one of these which satisfies the condition

$$h^2 \sum_R w_i^2 = 1 \tag{13}$$

is called the *normal eigenfunction.*

Basic Theorem 3.: Every function v *which vanishes at the external nodal points, can be expanded to a series like*

$$v = c_1 w_1 + c_2 w_2 + \ldots + c_n w_n \tag{14}$$

where, assuming normal functions w_i, the coefficients take on the form

$$c_i = h^2 \sum_R v w_i . \tag{14a}$$

Basic Theorem 4.: Eigenfunction w_1 *is of constant sign at every internal nodal point of network R.*

4. Auxiliary theorems

In the following, we shall derive some auxiliary theorems from the basic theorems introduced in Section 3.

Auxiliary Theorem 1. The homogeneous difference equations

$$\mathscr{S} w - l_i' w = 0 \tag{15}$$

and

$$\mathscr{M} w - l_i'' w = 0 \tag{16}$$

have the same eigenfunctions and the relation

$$l_i'' = 1 - 2l_i' \tag{17}$$

holds between their eigenvalues.

Combining Eqs (5) and (6) we obtain

$$\mathcal{M}f = f - 2\mathcal{D}f$$

which, applying to eigenfunction $f = w_i$, yields

$$\mathcal{M}w_i = w_i - 2\mathcal{D}w_i .$$

Substituting Eq. (10) to this equation, we arrive at

$$\mathcal{M}w_i = w_i - 2l_i'w_i = (1 - 2l_i')w_i .$$

It follows from this equation that, apart from Eq. (15), function w_i also satisfies Eq. (16), i.e. it is an eigenfunction of both difference equations (15), (16) and that the relation (17) between the eigenvalues l_i' and l_i'' holds.

Auxiliary Theorem 2. The sign of the linearly independent eigenfunctions w_i and w_k cannot be the same at every internal nodal point but they cannot be the opposite either at every internal nodal point.

This auxiliary theorem directly follows from Basic Theorem 2. If this auxiliary theorem is applied to the case $i = 1$, $k \neq 1$, it can be easily seen that, apart from the first eigenfunction, all the other eigenfunctions are of alternant sign at the internal nodal points.

Auxiliary Theorem 3. Only one eigenfunction belongs to eigenvalue l_1''.

If, apart from w_1, another eigenfunction $\bar{w}_1$ independent of w_1 belonged to l_1'', then, according to Basic Theorem 4 at every internal nodal point both eigenfunctions would have the same sign (e.g. positive) or their sign would be opposite (e.g. negative). According to Auxiliary Theorem 2, however, this is impossible.

Auxiliary Theorem 4. Unequalities

$$0 < l_i' < 1 \tag{18}$$

and

$$-1 < l_i'' < 1 \tag{19}$$

hold for every $i = 1, 2, \ldots, n$.

Of these statements, it is enough to prove unequality (19) since then unequality (18) automatically follows.

Through the demonstration, we have to take into consideration that $w_i = 0$ at the external nodal points and therefore there must be at least one internal nodal point where the absolute value of w_i is maximum. Let us choose such a nodal point and determine the value of l_i'' from Eq. (16):

$$l_i'' = \frac{\mathcal{M}w_i}{w_i} . \tag{20}$$

Since $|w_i|$ is maximum at this point, there are two possibilities here, namely, $|\mathcal{M}w_i|$ is smaller or equal to $|w_i|$.

In the first case unequality (19) automatically holds.

In the second case we have $l_i''=1$ or $l_i''=-1$. In this case the value of w_1 at the neighbouring points is the same as or the opposite of the value at the point in question. Passing to one of the four neighbouring points and applying Eq. (20), we find the same situation as at the point in question. Advancing from point to point, the situation is the same, till we arrive at a point next to one of the external points. Because of the condition $w_i=0$, the neighbouring w_i values cannot be equal to the original w_i. It follows that this second case is not a real possibility and therefore the unequality

$$|l_i''|<1,$$

i.e. Auxiliary Theorem 4. holds.

Auxiliary Theorem 5. Eigenvalues l_1'', l_2'', ..., l_n'' *and* $l_1', l_2', \ldots, l_n'$ *are symmetrical to 1/2 and 0, respectively.*

To prove this theorem, let us choose one of the nodal points as the origin of the co-ordinate system x, y and assign the function

$$w_k(x, y)=(-1)^{(x+y)/h}w_j(x, y) \tag{21}$$

as a *conjugate function* to eigenfunction $w_j(x, y)$ where $k\neq j$.

If, at the point chosen $w_j(x, y)=w_k(x, y)$ holds—*even nodal point*—, then we have

$$\mathcal{M}w_k(x, y)=-\mathcal{M}w_j(x, y),$$

i.e.

$$\mathcal{M}w_k(x, y)=-l_j''w_j(x, y)=-l_j'w_k(x, y).$$

The situation is similar when at the point chosen $w_k(x, y)=-w_j(x, y)$ holds—*odd nodal point*. In this case we have

$$\mathcal{M}w_k(x, y)=\mathcal{M}w_j(x, y)$$

and correspondingly

$$\mathcal{M}w_k(x, y)=l_j''w_j(x, y)=-l_j''w_k(x, y).$$

It follows that function w_k is an eigenfunction of difference equation (16), but the corresponding eigenvalue is the opposite of the eigenvalue of function w_j. Eigenvalues $l_1'', l_2'', \ldots, l_n''$ are therefore symmetrical to 0 and consequently eigenvalues $l_1', l_2', \ldots, l_n'$ are symmetrical to 1/2.

When n is an odd number, the middle eigenvalue of difference equation (16) is 0 and the middle eigenvalue of difference equation (15) is 1/2.

Auxiliary Theorem 6. If function v is of the same sign at every internal nodal point, then the sign of the expression

$$w_{1n}=c_1w_1-c_nw_n$$

constructed from eigenfunctions w_1 *and* w_n *is the same at every internal nodal point.*

It the case of normal eigenfunctions, formula (14a) yields

$$c_1 = h^2 \sum_R v w_1 , \qquad c_n = h^2 \sum_R v w_n$$

where, according to Basic Theorem 4, the terms in the summation in the formula of coefficient c_1 are all of the same sign and, according to Auxiliary Theorem 5., the same terms in the formula of c_n are of positive and negative sign. Thus

$$|c_n| < |c_1|, \tag{21a}$$

and, since Auxiliary Theorem 5 yields

$$|w_1| = |w_n|,$$

in the expression

$$w_{1n} = c_1 w_1 - c_n w_n$$

it is always the first term which dominates, i.e. w_{1n} is of the same sign at every internal nodal point.

Auxiliary Theorem 7. If we have $c_1 \neq 0$ *or* $c_n \neq 0$ *in series* (14), *then, for m great enough, the asymptotic equality*

$$\mathcal{M}_m v \cong c_1 l_1'' w_1 + (-1)^m c_n l_1'' w_n$$

holds.

To prove this statement, let us start from series (14) and apply the $\mathcal{M}$ operation m times:

$$\mathcal{M}_m v = c_1 l_1''^m w_1 + c_2 l_2''^m w_2 + \ldots + c_{n-1} l_{n-1}''^m w_{n-1} + c_n l_n''^m w_n. \tag{22}$$

According to Auxiliary Theorem 5, we have

$$c_n l_n''^m w_n = (-1)^m c_n l_n''^m w_n,$$

with which Eq. (22) takes on the form

$$\mathcal{M}_m v = c_1 l_1''^m w_1 + c_2 l_2''^m w_2 + \ldots + c_{n-1} l_{n-1}''^m w_{n-1} + (-1)^m c_n l_1''^m w_n. \tag{22a}$$

According to Auxiliary Theorem 3, l_1'' is a single eigenvalue and from Auxiliary Theorem 5 we obtain

$$|l_1''| > |l_i''| < |l_n''|, \qquad |w_1| = |w_n|$$

and unequality (21a) yields

$$|c_1| > |c_n|.$$

It follows that, for m great enough, the expression

$$c_2 l_2''^m w_2 + \ldots + c_{n-1} l_{n-1}''^m w_{n-1}$$

in formula (22a) is negligible in comparison with the sum of the first and last terms, i.e. the theorem is proved.

In the case $m<4$, instead of $\cong$, we can use the sign of equality since for $m=1$ and $m=2$ there is no internal term in formula (22) at all, and for $m=3$, according to Auxiliary Theorem 5, the medium term vanishes.

Auxiliary Theorem 8. *If the network contains at least two positive and at least two negative internal nodal points, then, in the case of* $j>1$, $c_j^2+c_k^2\neq 0$, *the expression*

$$w_{jk}=c_j w_j+c_k w_k$$

constructed of conjugate eigenfunctions w_j *and* w_k *is alternant or its value varies between zero and an extreme value at the internal nodal points.*

For $c_j=0$ and $c_k=0$ the theorem goes without saying since in these cases we have an expression of w_{jk} of one term which is proportional to either w_j or w_k and as such it must change sign at the internal nodal points.

For $c_j\neq 0$, $c_k\neq 0$, formula (21) related to conjugate eigenfunctions yields

$$w_{jk}=(c_j-c_k)w_j$$

at the even nodal points and

$$w_{jk}=(c_j+c_k)w_j$$

at the odd nodal points, so eigenfunction w_j has to be multiplied by a constant in both cases. If neither of these constants takes on zero value, then, according to Auxiliary Theorem 2, the expression of w_{jk} is an alternant one. If, however, we have $c_j-c_k=0$, the expression of w_{jk} can alternate between two values of opposite sign but one of these values may be zero. The situation is similar when $c_j+c_k=0$.

In the above Auxiliary theorem two conjugate eigenfunctions were assumed, i.e. the theorem does not hold for the case $k=n-j+1$ when w_j has no conjugate function.

5. Solution to the homogeneous difference equation

Let us first solve the homogeneous difference equation

$$\mathcal{M}w-l''w=0 \tag{24}$$

with the boundary condition $w=0$. We start from Auxiliary Theorem 7. which yields

$$\mathcal{M}_m v\cong c_1 l_1''^m w_1+(-1)^m c_n l_1''^m w_n,$$

$$\mathcal{M}_{m+1} v\cong c_1 l_1''^{m+1} w_1-(-1)^m c_n l_1''^{m+1} w_n,$$

$$\mathcal{M}_{m+2} v\cong c_1 l_1''^{m+2} w_1+(-1)^m c_n l_1''^{m+2} w_n.$$

From these equations we obtain

$$\mathcal{M}_{m+2} v\cong l_1''^2\,\mathcal{M}_m v$$

and

$$\mathcal{M}_{m+1} v+l_1''\mathcal{M}_m v\cong 2c_1 l_1''^{m+1} w_1$$

which lead us to the square of the eigenvalue

$$l_1''^2 \cong \frac{\mathscr{M}_{m+2}v}{\mathscr{M}_m v} \tag{25}$$

and to the first eigenfunction

$$c_1 w_1 \cong \frac{\mathscr{M}_{m+1}v + l_1''\mathscr{M}_m v}{2l_1''^{m+1}} \tag{26}$$

of the difference equation (24).

Knowing the eigenvalue and the eigenfunction of (24), now we can go back to our original problem, the computation of the first eigenvalue λ_1 and the first eigenfunction w_1 of the homogeneous difference equation.

$$\mathscr{D}w - l_1' w = 0, \tag{27}$$

identical to difference equation (3). The relationship defined by (17) exists between the eigenvalues of Eqs (27) and (24), i.e. we have

$$l_1' = \frac{1}{2}(1 - l_1'') \cong \frac{1}{2}\left(1 - \sqrt{\frac{\mathscr{M}_{m+2}v}{\mathscr{M}_m v}}\right)$$

which, taking into consideration formula (9b), yields the *first eigenvalue of difference equation* (3) as

$$\lambda_1 = \frac{8l_1'}{h^2} = \frac{4}{h^2}\left(1 - \sqrt{\frac{\mathscr{M}_{m+2}v}{\mathscr{M}_m v}}\right). \tag{28}$$

Since, according to Auxiliary Theorem 1, the eigenfunctions of difference equations (24) and (27) are identical, the first eigenfunction of difference equation (27) can be computed from formula (26) as

$$cw_1 \cong \frac{\mathscr{M}_{m+1}v + (1 - 2l_1')\mathscr{M}_m v}{2\left(1 - \frac{h^2 l_1'}{4}\right)^{m+1}}. \tag{29}$$

The function

$$w_1 \cong \mathscr{M}_{m+1}v + (1 - 2l_1')\mathscr{M}_m v \tag{29a}$$

proportional to function (29) is also a first eigenfunction and the *first normal eigenfunction of difference equation* (3) assumes the form

$$w_1 \cong \frac{\mathscr{M}_{m+1}v + (1 - 2l_1')\mathscr{M}_m v}{h\sqrt{\sum [\mathscr{M}_{m+1}v + (1 - 2l_1')\mathscr{M}_m v]^2}}. \tag{30}$$

In the case $m < 4$, the approximately equal sign $\cong$ in the above approximate formulas can be replaced by the sign equality since the formulas presented yield exact results in this case.

Accuracy analysis. The accuracy of the first eigenvalue λ_1 of the homogeneous difference equation (3) depends on the accuracy of l_1'', therefore it is expedient to establish a lower and an upper bound for the value of l_1''. We start from expansion (14) and neglect the terms at the beginning and at the end of the series which vanish together with their conjugates. Applying operation $\mathcal{M}$ by m and $(m+2)$ times, respectively, we arrive at the two series:

$$\begin{aligned} \mathcal{M}_m v &= l_1''^m c_1 w_1 + l_j''^m c_j w_j + \ldots + l_k''^m c_k w_k + l_n''^m c_n w_n, \\ \mathcal{M}_{m+2} w &= l_1''^{m+2} c_1 w_1 + l_j''^{m+2} c_j w_j + \ldots + l_k''^{m+2} c_k w_k + l_n''^{m+2} c_n w_n. \end{aligned} \tag{31}$$

According to Auxiliary Theorem 5, we have

$$l_k'' = -l_j'', \quad l_{k-1}'' = -l_{j+1}'', \quad \ldots, \quad l_n'' = -l_1''$$

in these series which leads us to

$$\begin{aligned} \mathcal{M}_m v &= l_1''^m (c_1 w_1 - c_n w_n) + l_j''^m (c_j w_j - c_k w_k) + \ldots, \\ \mathcal{M}_{m+2} v &= l_1''^{m+2} (c_1 w_1 - c_n w_n) + l_j''^{m+2} (c_j w_j - c_k w_k) + \ldots. \end{aligned}$$

Making use of the notation

$$\frac{l_j''}{l_1''} = s_j < 1, \tag{32}$$

we arrive at the formula

$$\frac{\mathcal{M}_{m+2} v}{\mathcal{M}_m v} = l_1''^2 \frac{1 + s_j^{m+2} \dfrac{c_j w_j - c_k w_k}{c_1 w_1 - c_n w_n} + \delta}{1 + s_j^m \dfrac{c_j w_j - c_k w_k}{c_1 w_1 - c_n w_n} + \varepsilon},$$

where, in the case of m great enough, quantities δ and ε are negligible compared to the other terms. Rearranging the above formula yields

$$\frac{\mathcal{M}_{m+2} v}{\mathcal{M}_m v} = l_1''^2 \left[1 - \frac{s_j^m (1 - s_j^2) \dfrac{c_j w_j - c_k w_k}{c_1 w_1 - c_n w_n} - \delta + \varepsilon}{1 + s_j^m \dfrac{c_j w_j - c_k w_k}{c_1 w_1 - c_n w_n} + \varepsilon} \right]. \tag{33}$$

In formula (33) the expression

$$s_j^m (1 - s_j^2)$$

is always positive and, according to Auxiliary Theorem 6, the expression

$$w_{1n} = c_1 w_1 - c_n w_n$$

is of the same sign at all the internal nodal points while according to Auxiliary

Theorem 8, the expression

$$w_{jk} = c_j w_j - c_k w_k$$

is either an alternant one or it varies between zero and an extreme value, provided the network contains at least two even and at least two odd internal nodal points. Consequently, the numerator of the big fraction is either an alternant quantity or (approximately) assumes zero value and its denominator is always positive. We can conclude that, depending on the place of the external nodal point, the value of the expression in brackets can be smaller or greater than unity and it may even be (approximately) equal to unity. It follows that the value of $l_1''^2$ has to fall between the greatest and smallest values of quotient $\mathcal{M}_{m+2}v/\mathcal{M}_m v$ or on its extreme value, i.e.

$$\left[\frac{\mathcal{M}_{m+2}v}{\mathcal{M}_m v}\right]_{\min} \leqq l_1''^2 \leqq \left[\frac{\mathcal{M}_{m+2}v}{\mathcal{M}_m v}\right]_{\max}.$$

When, after a sufficient number of steps, we find that the value of the expression

$$\left[\frac{\mathcal{M}_{m+2}v}{\mathcal{M}_m v}\right]_{\max} - \left[\frac{\mathcal{M}_{m+2}v}{\mathcal{M}_m v}\right]_{\min}$$

is small enough, we can accept the value of $l_1''^2$ as a good approximation.

There is no need for any accuracy analysis for the cases $m<4$ and $j=(1+n)/2$, since formulas (25) and (26) yield the exact eigenvalue and eigenfunction.

The above accuracy analysis can also be applied to cases where l_j'' or l_k'' are multiple eigenvalues.

Remark. When computing the first eigenvalue, it is sufficient to accept rough estimates at the beginning and the accuracy can be increased later on. It is also sufficient to rely on a coarse network at the beginning and later on, introducing new nodal points, the density of the network can be increased. In this way the method is relatively simple and efficient.

6. Solution to the inhomogeneous difference equation

Let us try to find a solution to the inhomogeneous difference equation

$$\mathcal{D}u - v = 0$$

identical to difference equation (4) with the boundary condition $u=0$. Let us start from Eq. (14) and apply operation M repeatedly:

$$\mathcal{M}_0 v = c_1 w_1 + c_2 w_2 + \ldots + c_n w_n,$$

$$\mathcal{M}_1 v = c_1 l_1'' w_1 + c_2 l_2'' w_2 + \ldots + c_n l_n'' w_n,$$

$$\mathcal{M}_2 v = c_1 l_1''^2 w_1 + c_2 l_2''^2 w_2 + \ldots + c_n l_n''^2 w_n,$$

. . . . . .

By producing the sum of the above equations we obtain

$$\sum_{i=0}^{\infty} \mathcal{M}_i v = c_1 w_1 \sum_{k=0}^{\infty} l_1''^k + c_2 w_2 \sum_{k=0}^{\infty} l_1''^k + \ldots + c_n w_n \sum_{k=0}^{\infty} l_n''^k \ldots \tag{35}$$

Since, by virtue of inequality (19), we have $|l''| < 1$ and according to the formula for the sum to infinity of geometrical series we have

$$\sum_{k=0}^{\infty} l_i''^k = \frac{1}{1 - l_1''},$$

Eq. (35) can be written as

$$\sum_{i=0}^{\infty} \mathcal{M}_i v = c_1 w_1 \frac{1}{1 - l_1''} + c_2 w_2 \frac{1}{1 - l_2''} + \ldots + c_n w_n \frac{1}{1 - l_n''},$$

or

$$\sum_{i=0}^{\infty} \mathcal{M}_i v = \sum_{k=1}^{n} c_k w_k \frac{1}{1 - l_k''}. \tag{36}$$

Instead of l_k'', we can introduce l_k' by making use of formula (17). By so doing, we arrive at

$$2 \sum_{i=0}^{\infty} \mathcal{M}_i v = \sum_{k=1}^{n} \frac{c_k}{l_k'} w_k \tag{37}$$

instead of (36). By carrying out operation $\mathcal{D}$ on both sides of the equation and taking into consideration Eq. (15) we obtain

$$2\mathcal{D} \sum_{i=0}^{\infty} \mathcal{M}_i v = \sum_{k=1}^{n} \frac{c_k}{l_k'} \mathcal{D} w_k = \sum_{k=1}^{n} c_k w_k.$$

According to Eq. (14) the right hand side of this equation represents function v itself, i.e.

$$2\mathcal{D} \sum_{i=0}^{\infty} \mathcal{M}_i v = v.$$

Eq. (11) shows that function v equals $\mathcal{D}u$, so that we have

$$2\mathcal{D} \sum_{i=0}^{\infty} \mathcal{M}_i v = \mathcal{D}u,$$

from which we obtain the unknown function u:

$$u = 2 \sum_{i=0}^{\infty} \mathcal{M}_i v. \tag{38}$$

This formula can be rearranged as

$$u=2\sum_{i=0}^{\infty}\mathcal{M}_i v=2\sum_{i=0}^{m}\mathcal{M}_i v+2(\mathcal{M}_{m+1}v+\mathcal{M}_{m+3}v+\mathcal{M}_{m+5}v+\ldots)+$$

$$+2(\mathcal{M}_{m+2}v+\mathcal{M}_{m+4}v+\mathcal{M}_{m+6}v+\ldots). \tag{39}$$

By making use of Eq. (24), we arrive at

$$u=2\sum_{i=0}^{\infty}\mathcal{M}_i v=2\sum_{i=0}^{m}\mathcal{M}_i v+2\mathcal{M}_{m+1}v(1+l_1''^2+l_1''^4+\ldots)+$$

$$+2\mathcal{M}_{m+2}v(1+l_1''^2+l_1''^4+\ldots),$$

which, applying the formula for the sum to infinity of geometrical series, takes on the form

$$u=2\sum_{0}^{m}\mathcal{M}_i v+2\mathcal{M}_{m+1}v\frac{1}{1-l_1''^2}+2\mathcal{M}_{m+2}v\frac{1}{1-l_1''^2}.$$

Finally, the solution to the *inhomogeneous difference equation* (34) reads

$$u=\sum_{i=0}^{m}\mathcal{M}_i v+\frac{2}{1-l_1''^2}(\mathcal{M}_{m+1}v+\mathcal{M}_{m+2}v). \tag{40}$$

Remark. Contrary to the homogeneous problem, we cannot start from an arbitrary function v when establishing the solution to the inhomogeneous problem, since function v and function V are defined by relationship (9). Neither can we accept rough estimates at the beginning of the process, since the terms used at the beginning also have a great influence on the accuracy of the method.

7. Numerical examples

In the following we shall present two numerical examples to illustrate the above numerical methods. Both examples refer to the rectangular network shown in Fig. 1 where the spacing is unity, i.e. $h=1$. The network contains more than two positive and more than two negative internal nodal points.

First, we shall present the approximate solution to the *homogeneous difference equation*

$$\Delta w+\lambda w=0 \tag{41}$$

with the boundary condition $w=0$.

Second, we shall discuss the *inhomogeneous difference equation*

$$\Delta u+8=0. \tag{42}$$

with the boundary condition $u=0$. With this problem we have $V=8$, i.e. Eq. (9a) yields

$$v=\frac{h^2V}{8}=1.$$

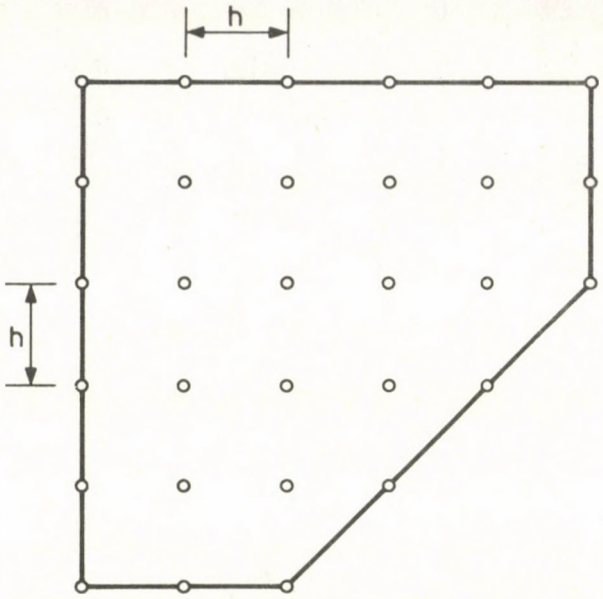


Fig. 1. Rectangular network

7.1 Approximate solution to the homogeneous difference equation

It is expedient to choose function v for the solution to difference equation (41) as $v = 1$ as it appears in the second numerical example, since in this case the results of this numerical example can be used for the approximate solution of difference equation (42). Since we intend to use the results for the second numerical example, we have to compute as exactly as possible from the beginning.

The first eigenvalue

To start with, we need the table for values $\mathcal{M}_0 v$ (Fig. 2):

The next step is to produce the table for values $\mathcal{M}_1 v$. Function $\mathcal{M}_1 v$ vanishes at every external nodal point and the values at the internal nodal points are obtained by applying operation $\mathcal{M}$ (production of the mean value). E.g. the value 0.7500 framed in the table of $\mathcal{M}_1 v$ is obtained by producing the arithmetic mean of the four values framed in the table of $\mathcal{M}_0 v$. Proceeding in a similar way at the other nodal points, we obtain the table of values $\mathcal{M}_1 v$ (Fig. 3):

By repeating the above process and omitting the zero-values at the external nodal points, we obtain the table of values $\mathcal{M}_2 v$ (Fig. 4):

0	0	[0]	0	0	0
0	[1]	1	[1]	1	0
0	1	[1]	1	1	0
0	1	1	1	0	
0	1	1	0		
0	0	0			

Fig. 2. Table of values $\mathcal{M}_0 v$

0	0	0	0	0	0
0	0.5000	[0.7500]	0.7500	0.5000	0
0	0.7500	1.0000	1.0000	0.5000	0
0	0.7500	1.0000	0.5000	0	
0	0.5000	0.5000	0		
0	0	0			

Fig. 3. Table of values $\mathcal{M}_1 v$

0.3750	0.5625	0.5625	0.3125
0.5625	0.8750	0.6875	0.3750
0.5625	0.6875	0.5000	
0.3125	0.3750		

Fig. 4. Table of values $\mathcal{M}_2 v$

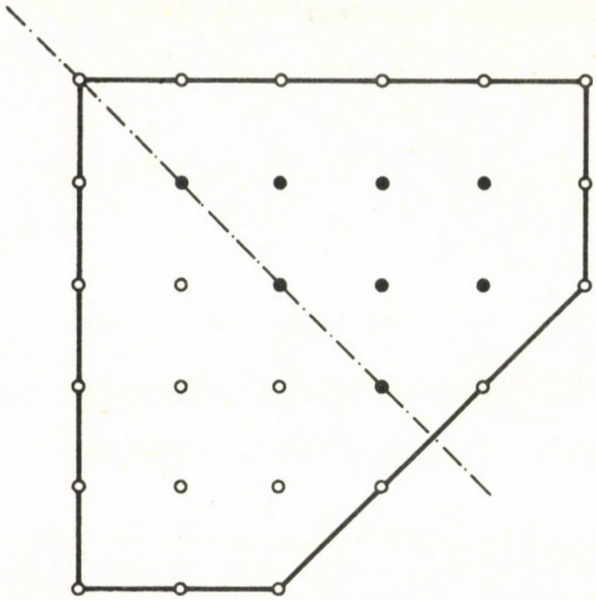


Fig. 5. Places of the numerical values are marked by full circles

Proceeding in this way, we obtain the tables for $\mathcal{M}_3 v, \mathcal{M}_4 v, \ldots, \mathcal{M}_7 v$. Leaving out the symmetrical values (Fig. 5), the table for values $\mathcal{M}_8 v$ is obtained as

0.076 447	0.107 468	0.104 204	0.051 957
	0.171 449	0.133 225	0.066 040
		0.095 002	

the table for values $\mathcal{M}_9 v$ as

0.053 734	0.088 040	0.073 163	0.042 576
	0.120 347	0.109 189	0.046 296
		0.066 613	

and finally the table for values $\mathcal{M}_{10} v$ as

0.044 020	0.061 811	0.059 951	0.029 865
	0.098 615	0.076 605	0.037 941
		0.054 595	

Knowing values $\mathcal{M}_{10}v$ and $\mathcal{M}_8v$, we can now produce the table for values $\mathcal{M}_{10}v/\mathcal{M}_8v$:

0.575 824	0.575 157	0.574 992	0.574 802
	0.575 185	0.575 005	0.574 515
		0.574 672	

It can be seen that the values for quotient $\mathcal{M}_{10}v/\mathcal{M}_8v$ fall between 0.574 515 and 0.575 824 and their mean value is

$$l_1'^2 = 0.574\,971\,.$$

By making use of this value, we obtain

$$l_1' = \frac{1}{2}(1 - \sqrt{0.574\,971}) = 0.120\,866 \tag{43}$$

from formula (28). We can now use formula (9b) which, with good approximation yields *the first eigenvalue of the homogeneous difference equation:*

$$\lambda_1 \cong \frac{8l'}{h^2} = 0.966\,926\,. \tag{44}$$

If, to obtain greater accuracy, we continue the computation till the determination of $\mathcal{M}_{20}v$, the first eigenvalue emerges as

$$\lambda_1 = 0.966\,750\,. \tag{45}$$

As can be seen, there is only a little difference between (44) and the more accurate (45).

The first eigenfunction

Eigenfunction w_1 can be approximately obtained from formula (29a). By making use of the values of the tables for $\mathcal{M}_9v$ and $\mathcal{M}_{10}v$, we obtain the values of w_1:

0.084 765	0.128 569	0.115 428	0.062 149
	0.189 870	0.159 400	0.073 046
		0.105 106	

To obtain the normal form of function w_1, we have to produce the squares of the above values, then the sum of the squares. In our case this sum assumes the value

$$\sum_R w_1^2 = 0.183\,027\,.$$

With $h=1$, the denominator of formula (30) for the normal function w_1 assumes the value

$$1 \cdot \sqrt{\sum_R w_1^2} = 0.427\,817\,.$$

We obtain the normal form of the first eigenfunction by dividing the values of the last table by 0.427 817 computed above. We compiled the approximate values of the first normal eigenfunction of the homogeneous difference equation at the nodal points in the following table:

0.198 134	0.300 524	0.269 807	0.145 270
	0.443 812	0.372 590	0.170 741
		0.245 680	

By computing the approximate values of the first eigenvalue and the first eigenfunction, we have solved the problem of the homogeneous difference equation.

If, to obtain greater accuracy in the computation of the first eigenfunction, we continue the process till the determination of $\mathcal{M}_{20}v$, we obtain the more accurate values of the normal function w_1 as

0.198 123	0.300 479	0.269 670	0.145 151
	0.443 634	0.372 348	0.170 609
		0.245 511	

The difference between these values and those less accurate ones in the previous table is negligible.

7.2 Approximate solution to the inhomogeneous difference equation

To obtain an approximate solution to difference equation (42), we use formula (40) which, for $m=8$, yields

$$u = 2\sum_{i=0}^{8} \mathcal{M}_i v + \frac{2}{1-l_1'^2}(\mathcal{M}_9 v + \mathcal{M}_{10} v)\,.$$

When computing the term $\sum_{i=0}^{8} \mathcal{M}_i v$ in the above formula, we can make use of the tables of $\mathcal{M}_0 v, \mathcal{M}_1 v, \ldots, \mathcal{M}_8 v$ compiled for the solution of the homogeneous problem. The sums of the corresponding values of these tables, i.e. the table of values $\sum_{i=0}^{8} \mathcal{M}_i v$ is given

as

2.847 079	3.801 623	3.654 131	2.553 605
	5.057 437	4.553 942	2.730 592
		3.210 360	

Similarly, values $\mathcal{M}_9 v$ and $\mathcal{M}_{10} v$ also come from the tables complied for the solution of the homogeneous problem. With these values, the table of values $(\mathcal{M}_9 v + \mathcal{M}_{10} v)/(1 - l_1'^2)$ is as follows:

0.229 994	0.352 567	0.313 188	0.170 438
	0.515 170	0.437 133	0.198 191
		0.285 176	

Finally, formula (40) yields the approximate values of the unknown function u as the double sum of the values of the above two tables:

6.154 146	8.308 380	7.934 638	5.448 086
	11.145 214	9.982 150	5.857 566
		6.991 072	

This is the approximate solution to the inhomogeneous problem.

Acknowledgements

The Author is grateful to the late György Alexits, mathematician, later academician, with whom he had useful discussions about the subject of this paper 55 years ago.

References

1. Runge, C.: Über eine Methode die partielle Differentialgleichung $\Delta u = \text{Constans}$ numerisch zu integrieren. *Zeitschrift für Mathematik und Physik* **56** (1909), 226–232
2. Liebmann, H.: *Sitzungsberichte der mathematisch-physikalischen Klasse der Bayerischen Akademie der Wissenschaften zu München.* Jahrgang 1918, 385–416
3. Courant, R.–Hilbert, D.: Methoden der mathematischen Physik. Bd. I. Verlag von Julius Springer, Berlin 1924
4. Wolf, F.: Über die angenäherte numerische Berechnung harmonischer und biharmonischer Funktionen. *Zeitschrift für angewandte Mathematik und Physik* **6** (1928), 118–150

Acta Technica Academiae Scientiarum Hungaricae, 98 (3—4), pp. 269—274 (1985)

TORSION OF BARS WITH A TRIANGULAR HOLLOW CROSS SECTION

P. CSONKA*

[Received: 10 January 1984]

The subject of this paper is to investigate the pure torsion of an elastic prismatic bar whose cross section is an equilateral triangle with a circular hole in its middle point. The treatment of this task requires the determination of the stress function of the problem. For this purpose a triple symmetric expression with two free parameters is chosen as approximation. The values of the parameters are determined in such a way that the curve defined by the equation of the stress function passes the corner points of the cross section and at the same time these points are double points of the curve.

1. Introduction

The purpose of this paper is to present an approximate method for the determination of the stress function of pure torsion for an elastic prismatic bar whose cross section is an equilateral triangle with a circular hole in its middle point.

A polar coordinate system $0(r, \varphi)$ is introduced whose origin 0 is placed in the centre of the circular hole and the straight $\varphi = 0$ halves one side of the ground plan triangle (Fig. 1). The radius of the inscribed and circumscribed circle is denoted by a and R respectively, while the radius of the circular hole is marked r_0.

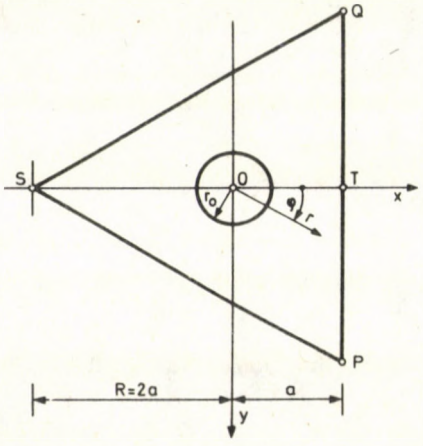


Fig. 1. Equilateral triangle cross section with a circular hole in its centre

* P. Csonka, H-1114 Budapest, Bartók B. út 31., Hungary

The solution of the problem requires the construction of a stress function $F = F(r, \varphi)$ satisfying differential equation

$$F = \frac{\partial^2 F}{\partial r^2} + \frac{1}{r}\frac{\partial F}{\partial r} + \frac{1}{r^2}\frac{\partial^2 F}{\partial \varphi^2} = \text{const} \tag{1}$$

as well boundary conditions

$$F = 0 \tag{2}$$

and

$$F = \text{const} \neq 0 \tag{3}$$

along the outer and inner boundary line of the cross section.

2. Solution of the problem

To produce the stress function of torsion—at least approximately—let us start from the triple symmetric function

$$F(r, \varphi) = r^2 - r_0^2 + A\frac{r^3}{r_0}\left(1 - \frac{r_0^6}{r^6}\right)\cos 3\varphi + Br_0^2 = 0 \tag{4}$$

where A and B are constants. This function fulfils a priori conditions (1) and (3), and when attributing appropriate values to A and B, it also fulfils condition (2). Namely, if parameters A and B are suitable chosen, the curve corresponding to the function in question—the *function curve*—consists of three intersecting branches which are nearly straight in sections between the corner points (Fig. 2).

In order to ensure that the function curve consisting of three intersecting branches encloses a configuration having nearly the same shape as the given triangle, the points of intersection of the three branches have to coincide with the three corner points of the triangle. Since the intersections of the three branches are double points, conditions

$$F = 0; \qquad \frac{\partial F}{\partial r} = 0; \qquad \frac{\partial F}{\partial \varphi} = 0 \tag{5}$$

—valid for double points—have to be fulfilled at the corner points of the cross section.

As at the three corner points the relations

$$r = R; \qquad \cos 3\varphi = -1; \qquad \sin 3\varphi = 0$$

are valid, among conditions (5) the third one is a priori fulfilled and the first two are also fulfilled, if

$$R^2 - r_0^2 - A\frac{R^3}{r_0}\left(1 - \frac{r_0^6}{R^6}\right) + Br_0^2 = 0 \tag{6}$$

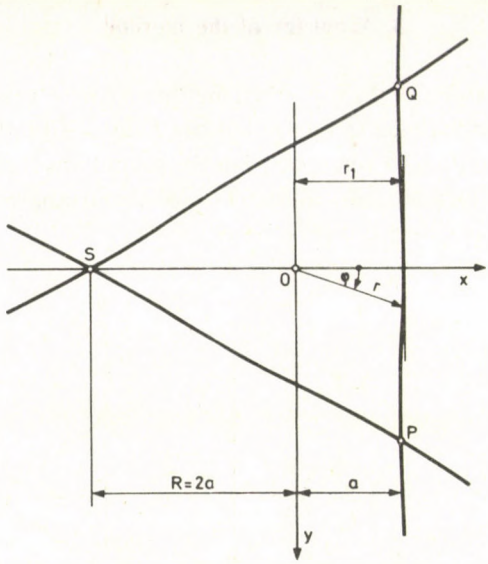


Fig. 2. The three intersecting branches of the function curve

and

$$2R-3A\left(\frac{R^2}{r_0}+\frac{r_0^5}{R^4}\right)=0 \tag{7}$$

hold.

Introducing the simplifying notation

$$\rho=\frac{r_0}{R}=\frac{r_0}{2a},$$

equations (6) and (7) may be transformed as

$$A(\rho^6-1)-\rho^5-(B-1)\rho^3=0,$$

$$2\rho^5-3A(1+\rho^6)=0$$

yielding, in turn, A and B

$$A=\frac{2\rho^5}{3(1+\rho^6)},$$

$$B=-\frac{\rho^8+5\rho^2}{3(1+\rho^6)}+1, \tag{8}$$

making stress function (4),—so far indefinite—perfectly known.

3. Accuracy of the method

To check the accuracy of the presented method we need to know the distance Δx between the points of the function curve and the points of the boundary line of the triangle. As the greatest Δx is to be expected at the point lying farthest from the corner points, that is at the middle point of the side of the triangle, we are particularly interested in its value at this point (Fig. 3).

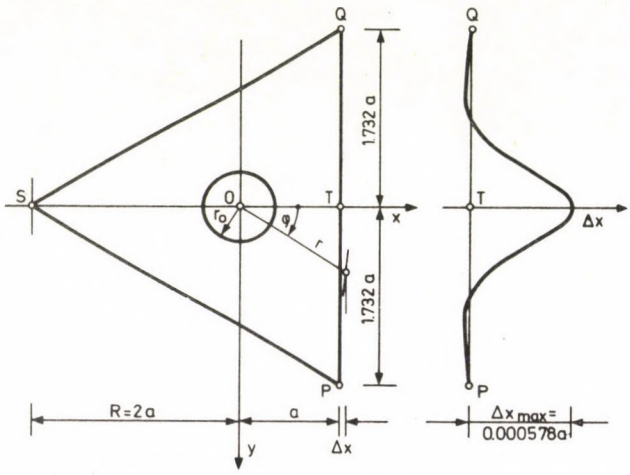


Fig. 3. Distance between the function curve and the side PQ of the triangle, enlarged 5000 times

For this purpose values of parameters A and B to be attributed to different values of ρ have been determined from (8) and compiled in Table 1.

Radius vector r_1 of the function curve at $\varphi = 0$ is obtained from formula

$$r_1^2 - r_0^2 + A\frac{r_1^3}{r_0}\left(1 - \frac{r_0^6}{r_1^6}\right) Br_0^2 = 0,$$

Table 1

Parameters A and B

ρ	a/r_0	A	B
4	2.0	0.166 625 986	− 4.338 540 396
5	2.5	0.133 324 801	− 7.335 466 530
6	3.0	0.111 108 730	−11.001 028 785
7	3.5	0.095 237 286	−15.333 888 653
8	4.0	0.083 333 015	−20.333 658 852
9	4.5	0.074 073 935	−26.000 203 221
10	5.0	0.066 666 600	−32.333 466 667

analogous to (4), which can be brought with notation

$$q=\frac{r_1}{r_0},$$

to the form

$$q^2+Aq^3\left(1-\frac{1}{q^6}\right)+B-1=0,$$

yielding values q compiled in Table 2. It is to be seen that the value $q-a/r_0$ is negligibly small in cases $\rho\geqq 4$, that is practically in all cases occurring in practice.

Table 2

Quotient q

ρ	a/r_0	q
4	2.0	2.004 365
5	2.5	2.501 437
6	3.0	3.000 578
7	3.5	3.500 263
8	4.0	4.000 147
9	4.5	4.500 076
10	5.0	5.000 045

As a definite example let us look at the cross section to be seen in Figure 1, where $\rho=6$. For this case we have determined and compiled in Table 3 the values Δx for different points of the boundary line of the triangle. The greatest value of Δx is 0.000 578a,—a very small value—which is about 40 times (!) smaller, than the thickness of the boundary line of the triangle in Figure 1.

Table 3

Data of the function curve $\rho=6$

r/r_0	x/r_0	$\lvert y\rvert/r_0$	x/r_0
3.000 578	3.000 578	0.000 000	0.000 578
3.10	3.000 429	0.799 374	0.000 429
3.25	3.000 269	1.249 354	0.000 269
3.5	3.000 112	1.793 257	0.000 112
4.0	2.999 995	2.646 754	−0.000 005
4.5	2.999 973	3.354 235	−0.000 027
5.0	2.999 978	4.000 017	−0.000 022
5.5	2.999 989	4.771 013	−0.000 011
6.0	3.000 000	5.196 152	0.000 000

References

1. Grammel, R.: Mechanik elastischer Körper (Handbuch der Physik, Bd. VI) Julius Springer, Berlin 1928, 1. Aufl. 143–181.
2. Timoshenko, S.–Goodier, J. N.: Theory of Elasticity, McGraw-Hill Book Company, Inc. New York–Toronto–London, 2. Ed. (1951), 258–315.
3. L'Hermite, R.: Résistance des matériaux. Théorique et expérimentale, Tome 1, Dunod, Paris 1954, 196–244.

Acta Technica Academiae Scientiarum Hungaricae, 98 (3—4), pp. 275—293 (1985)

A SPECIAL CASE OF THE PROBLEM OF TORSION OF HOLLOW-CORE SOLIDS OF REVOLUTION

I. ECSEDI*

[Received 20 September 1984]

The torsion problem of hollow-core solids of revolution made of an elastic material will be handled by adopting the usual assumptions of the theory of elasticity. Meridian section of the examined solid of revolution is bounded by coordinate lines of a plane orthogonal curvilinear coordinate system.

Symbols

α, φ, β	orthogonal curvilinear coordinates;
r, z	orthogonal coordinates in the meridian plane;
$\partial T = \partial T_1 \cup \partial T_2 \cup \partial T_3 \cup \partial T_4$	boundary curve of the meridian section;
$\partial V_i (i = 1, 2, 3, 4)$	surfaces of revolution;
$\boldsymbol{\rho} = \boldsymbol{\rho}(r, z) = r\mathbf{e}_r + z\mathbf{e}_Z$	place vector in the meridian plane;
$\mathbf{e}_r, \mathbf{e}_\varphi, \mathbf{e}_\chi, \mathbf{e}_\alpha, \mathbf{e}_\beta$	unit vectors;
ds	arc element;
$H_\alpha, H_\beta, H_\varphi$	Lamé coefficients;
V	Hamiltonian differential operator;
$\mathbf{u}$	displacement vector;
"$\cdot$"	symbol of scalar multiplication;
$\mathbf{u}V$ and $V\mathbf{u}$	diadic products of vectors $\mathbf{u}$ and V;
$\boldsymbol{\varepsilon}$	strain tensor;
$\varepsilon_\alpha, \varepsilon_\varphi, \varepsilon_\beta$	specific strains;
$\gamma_{\alpha\beta} = \gamma_{\beta\alpha}, \gamma_{\alpha\varphi} = \gamma_{\varphi\alpha}, \gamma_{\beta\varphi} = \gamma_{\varphi\beta}$	specific angular rotations;
$\sigma_\alpha, \sigma_\varphi, \sigma_\beta$	normal stresses;
$\tau_{\alpha\beta} = \tau_{\beta\alpha}, \tau_{\alpha\varphi} = \tau_{\varphi\alpha}, \tau_{\beta\varphi} = \tau_{\varphi\beta}$	shear stresses;
$\mathbf{T}$	stress tensor;
$\mathbf{q}$	volume load;
$\mathbf{p}$	surface load;
$U = U(\beta)$	stress function;
$V = V(\alpha)$	auxiliary function;
$a = a(\alpha), b = b(\beta)$	auxiliary functions;
G	modulus of elasticity in shear;
M	torque;
S	torsional rigidity;
$\sqrt{-1} = i$	imaginary unit;
$\xi = \alpha + i\beta$	complex variables;
C	constant.

Other magnitudes, variables are interpreted in the text.

* I. Ecsedi, H-3524 Miskolc, Klapka Gy. u. 36, IX/2, Hungary

1. Introduction

Assumptions usual in the theory of elasticity will be applied, namely, that:
— deformations and displacements are small;
— the material is homogeneous, isotropic and linear elastic;
— thermal effects are negligible;
— initial stresses and displacements are zero;
— the problem is a quasi-static one.

Meridian section of the tested solid of revolution symmetry is seen in Fig. 1. Boundary surfaces of the solid of revolution are surfaces of revolution ∂V_1, ∂V_2, ∂V_3,

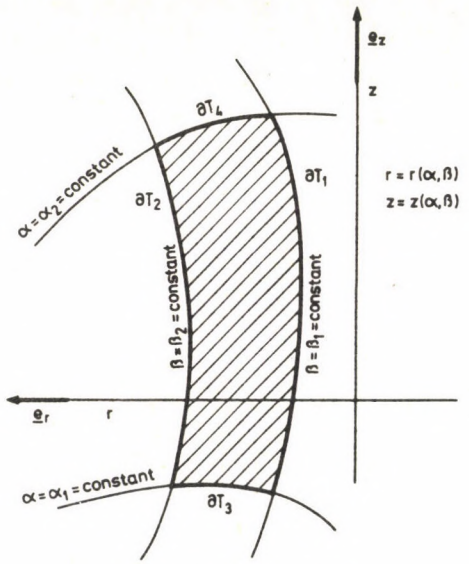


Fig. 1. Meridian section

∂V_4 obtained by rotating curves ∂T_1, ∂T_2, ∂T_3, ∂T_4. Curves $\partial T_i (i = 1, 2, 3, 4)$ are bounded by coordinate lines of orthogonal curvilinear coordinate system $\alpha\beta$ in plane r, z (Fig. 2).

Computations are made in the spatial orthogonal curvilinear coordinate system (α, φ, β). Spatial point P is located in terms of polar angle φ of the meridian plane comprising point P, and of orthogonal curvilinear coordinates α, β interpreted in the meridian plane.

In the meridian plane set out by polar angle φ, place vector $\boldsymbol{\rho}$ of point P is

$$\boldsymbol{\rho} = r\mathbf{e}_r + z\mathbf{e}_z \tag{1.1}$$

where

$$r = r(\alpha, \beta), \tag{1.2}$$

$$z = z(\alpha, \beta). \tag{1.3}$$

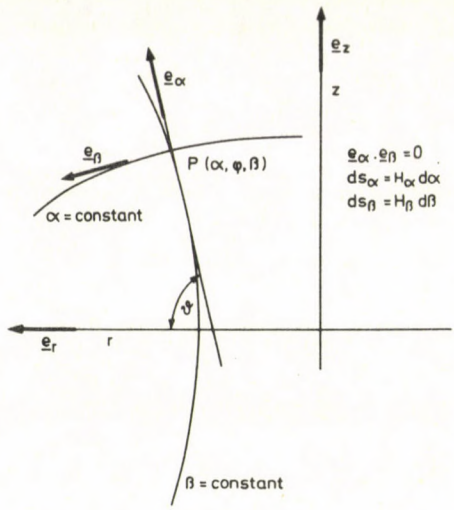


Fig. 2. Coordinate lines

Along lines α, and β of orthogonal curvilinear coordinate system $\alpha\beta$, β, and $\alpha = \text{constant}$, respectively (Fig. 2).

Tangential unit vector of curve $\beta = \text{constant}$ ([1], [3]):

$$\mathbf{e}_\alpha = \frac{1}{H_\alpha}\frac{\partial \boldsymbol{\rho}}{\partial \alpha} = \frac{1}{H_\alpha}\frac{\partial r}{\partial \alpha}\mathbf{e}_r + \frac{1}{H_\alpha}\frac{\partial z}{\partial \alpha}\mathbf{e}_z \tag{1.4}$$

and tangential unit vector of curve $\alpha = \text{constant}$:

$$\mathbf{e}_\beta = \frac{1}{H_\beta}\frac{\partial \boldsymbol{\rho}}{\partial \beta} = \frac{1}{H_\beta}\frac{\partial r}{\partial \beta}\mathbf{e}_r + \frac{1}{H_\beta}\frac{\partial z}{\partial \beta}\mathbf{e}_z . \tag{1.5}$$

In formulae above [1]

$$H_\alpha = \sqrt{\left(\frac{\partial r}{\partial \alpha}\right)^2 + \left(\frac{\partial z}{\partial \alpha}\right)^2}, \tag{1.6}$$

$$H_\beta = \sqrt{\left(\frac{\partial r}{\partial \beta}\right)^2 + \left(\frac{\partial z}{\partial \beta}\right)^2} . \tag{1.7}$$

In consequence of orthogonality

$$\frac{\partial \boldsymbol{\rho}}{\partial \alpha}\cdot\frac{\partial \boldsymbol{\rho}}{\partial \beta} = \frac{\partial r}{\partial \alpha}\frac{\partial r}{\partial \beta} + \frac{\partial z}{\partial \alpha}\frac{\partial z}{\partial \beta} = 0 . \tag{1.8}$$

In the meridian plane:

$$(\mathrm{d}s)^2 = H_\alpha^2(\mathrm{d}\alpha)^2 + H_\beta^2(\mathrm{d}\beta)^2 \tag{1.9}$$

expression for the square of the arc element.

Hamiltonian differential operator V is expressed as ([1], [2])

$$V = \frac{1}{H_\alpha}\frac{\partial}{\partial\alpha}\mathbf{e}_\alpha + \frac{1}{H_\varphi}\frac{\partial}{\partial\varphi}\mathbf{e}_\varphi + \frac{1}{H_\beta}\frac{\partial}{\partial\beta}\mathbf{e}_\beta \tag{1.10}$$

where:

$$H_\varphi = r(\alpha, \beta)\,. \tag{1.11}$$

Correctness of

$$\mathbf{e}_\alpha = \mathbf{e}_r \cos\vartheta + \mathbf{e}_z \sin\vartheta\,, \tag{1.12}$$

$$\mathbf{e}_\beta = \mathbf{e}_r \sin\vartheta - \mathbf{e}_z \cos\vartheta \tag{1.13}$$

is understood from Fig. 2.

It should be noted that

$$\cos\vartheta = \frac{1}{H_\alpha}\frac{\partial r}{\partial\alpha} = -\frac{1}{H_\beta}\frac{\partial z}{\partial\beta}\,, \tag{1.14}$$

$$\sin\vartheta = \frac{1}{H_\alpha}\frac{\partial z}{\partial\alpha} = \frac{1}{H_\beta}\frac{\partial r}{\partial\beta}\,. \tag{1.15}$$

Elementary calculation may verify the following rules of differentiation:

$$\frac{\partial\mathbf{e}_\alpha}{\partial\varphi} = \cos\vartheta\,\mathbf{e}_\varphi\,, \tag{1.16}$$

$$\frac{\partial\mathbf{e}_\beta}{\partial\varphi} = \sin\vartheta\,\mathbf{e}_\varphi\,, \tag{1.17}$$

$$\frac{\partial\mathbf{e}_\varphi}{\partial\varphi} = -\mathbf{e}_r = -\cos\vartheta\,\mathbf{e}_\alpha - \sin\vartheta\,\mathbf{e}_\beta\,. \tag{1.18}$$

2. A torsion problem

Torsion problem of the solid of revolution in Fig. 3 will concern a solid with a displacement vector $\mathbf{u} = \mathbf{u}(\alpha, \varphi, \beta)$ to be indicated as

$$\mathbf{u} = v(\alpha, \beta)\mathbf{e}_\varphi \tag{2.1}$$

Furthermore, "outer and inner mantle surfaces" of the solid, revolution surfaces ∂V_1 and ∂V_2 are assumed to bear no load, and volume load density $\mathbf{q}$ to be zero vector at any point of the solid.

Let the solid of revolution be clamped over surface section ∂V_3, that is

$$\mathbf{u}(P) = \mathbf{0} \qquad P \in \partial V_3\,. \tag{2.2}$$

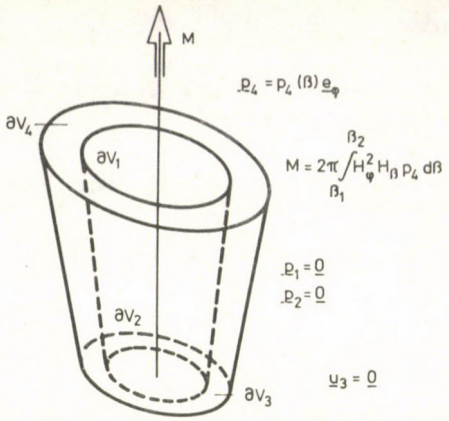


Fig. 3. Solid of revolution bounded by coordinate surfaces

Furthermore, let surface load $\mathbf{p}_4$ acting on surface section ∂V_4 be of the form

$$\mathbf{p}_4 = p_4(\beta)\mathbf{e}_\varphi \,. \tag{2.3}$$

The pair of resultant vectors $\mathbf{F}$, $\mathbf{M}_A$ of surface load of density $\mathbf{p}_4$ is obtained as

$$\mathbf{F} = \int_{\beta_1}^{\beta_2} p_4(\beta) H_\varphi(\alpha_2, \beta) H_\beta(\alpha_2, \beta) \mathrm{d}\varphi \int_0^{2\pi} \mathbf{e}_\varphi \, \mathrm{d}\varphi = \mathbf{0} \,, \tag{2.4}$$

$$\mathbf{M}_A = \{2\pi \int_{\beta_1}^{\beta_2} p_4(\beta) H_\varphi^2(\alpha_2, \beta) H_\beta(\alpha_2, \beta) \, \mathrm{d}\beta\} \mathbf{e}_z \,. \tag{2.5}$$

A in Eq. (2.5) stands for an arbitrary point along axis z of revolution.

From Eqs (2.4) and (2.5) it appears that, however function $p_4 = p_4(\beta)$ is specified, resultant vector $\mathbf{F}$ of the load acting on surface section ∂V_4 is in any case zero vector.

In writing Eqs (2.3), (2.4), (2.5) it has been made use of the fact that

along curve ∂T_3 $\quad \alpha = \alpha_1 = \text{constant}$,

along curve ∂T_4 $\quad \alpha = \alpha_2 = \text{constant}$,

$$\beta_1 \leq \beta \leq \beta_2 \,,$$

namely along curve ∂T_1 $\quad \beta = \beta_1 = \text{constant}$,

along curve ∂T_2 $\quad \beta = \beta_2 = \text{constant}$.

Meridian curves $\partial T_i (i = 1, 2, 3, 4)$ are illustrated by Fig. 1.

Strain tensor

$$\boldsymbol{\varepsilon} = \frac{1}{2}(\mathbf{u}\nabla + \nabla\mathbf{u}) \tag{2.6}$$

can be produced in form

$$\boldsymbol{\varepsilon} = \frac{1}{2H_\alpha}\frac{\partial v}{\partial \alpha}(\mathbf{e}_\varphi \mathbf{e}_\alpha + \mathbf{e}_\alpha \mathbf{e}_\varphi) +$$

$$+ \frac{1}{2H_\beta}\frac{\partial v}{\partial \beta}(\mathbf{e}_\varphi \mathbf{e}_\beta + \mathbf{e}_\beta \mathbf{e}_\varphi) + \frac{v}{H_\varphi}(\mathbf{e}_r \mathbf{e}_\varphi + \mathbf{e}_\varphi \mathbf{e}_r). \tag{2.7}$$

Correctness of Eq. (2.7) may be verified by means of relationships

$$\mathbf{u}\nabla = \frac{1}{H_\alpha}\frac{\partial v}{\partial \alpha}\mathbf{e}_\varphi \mathbf{e}_\alpha - \frac{v}{H_\varphi}\mathbf{e}_r \mathbf{e}_\varphi - \frac{1}{H_\beta}\frac{\partial v}{\partial \beta}\mathbf{e}_\varphi \mathbf{e}_\beta, \tag{2.8}$$

$$\nabla \mathbf{u} = \frac{1}{H_\alpha}\frac{\partial v}{\partial \alpha}\mathbf{e}_\alpha \mathbf{e}_\varphi - \frac{v}{H_\varphi}\mathbf{e}_\varphi \mathbf{e}_r + \frac{1}{H_\beta}\frac{\partial v}{\partial \beta}\mathbf{e}_\beta \mathbf{e}_\varphi \tag{2.9}$$

written according to Eq. (2.1). After some calculation

$$\varepsilon_\alpha = \mathbf{e}_\alpha \cdot \boldsymbol{\varepsilon} \cdot \mathbf{e}_\alpha = 0, \tag{2.10}$$

$$\varepsilon_\beta = \mathbf{e}_\beta \cdot \boldsymbol{\varepsilon} \cdot \mathbf{e}_\beta = 0, \tag{2.11}$$

$$\varepsilon_\varphi = \mathbf{e}_\varphi \cdot \boldsymbol{\varepsilon} \cdot \mathbf{e}_\varphi = 0, \tag{2.12}$$

$$\gamma_{\alpha\beta} = \gamma_{\beta\alpha} = 2\mathbf{e}_\beta \cdot \boldsymbol{\varepsilon} \cdot \mathbf{e}_\alpha = 0, \tag{2.13}$$

$$\gamma_{\alpha\varphi} = \gamma_{\varphi\alpha} = 2\mathbf{e}_\alpha \cdot \boldsymbol{\varepsilon} \cdot \mathbf{e}_\varphi = \frac{1}{H_\alpha}\frac{\partial v}{\partial \alpha} - \frac{v}{H_\varphi}\frac{1}{H_\alpha}\frac{\partial H_\varphi}{\partial \alpha} =$$

$$= \frac{H_\varphi}{H_\alpha}\frac{\partial}{\partial \alpha}\left(\frac{v}{H_\varphi}\right), \tag{2.14}$$

$$\gamma_{\beta\varphi} = \gamma_{\varphi\beta} = 2\mathbf{e}_\beta \cdot \boldsymbol{\varepsilon} \cdot \mathbf{e}_\varphi = \frac{1}{H_\beta}\frac{\partial v}{\partial \beta} -$$

$$- \frac{v}{H_\varphi}\frac{1}{H_\beta}\frac{\partial H_\varphi}{\partial \beta} = \frac{H_\varphi}{H_\beta}\frac{\partial}{\partial \beta}\left(\frac{v}{H_\varphi}\right). \tag{2.15}$$

Thereafter the analysis will be restricted to a function $v = v(\alpha, \beta)$ of the form

$$v(\alpha, \beta) = V(\alpha) H_\varphi(\alpha, \beta). \tag{2.16}$$

In this case:

$$\gamma_{\alpha\varphi} = \frac{H_\varphi}{H_\alpha}\frac{\mathrm{d}V}{\mathrm{d}\alpha}, \tag{2.17}$$

$$\gamma_{\beta\varphi} = 0. \tag{2.18}$$

Utilizing the general Hooke's law

$$\sigma_\alpha = \sigma_\varphi = \sigma_\beta = \tau_{\alpha\beta} = \tau_{\beta\varphi} = 0 \tag{2.19}$$

at every point of the solid, furthermore

$$\tau_{\alpha\varphi} = G \frac{H_\varphi}{H_\beta} \frac{\mathrm{d}V}{\mathrm{d}\alpha}. \tag{2.20}$$

Accordingly, stress tensor $\mathbf{T}$ can be given in form

$$\mathbf{T} = \tau(\alpha, \beta)\,[\mathbf{e}_\alpha \mathbf{e}_\varphi + \mathbf{e}_\varphi \mathbf{e}_\alpha]\,, \tag{2.21}$$

introducing notation

$$\tau_{\alpha\varphi} = \tau_{\varphi\alpha} = \tau\,. \tag{2.22}$$

It is easy to demonstrate that the surface load value belonging to a stress tensor of this form is zero vector on surface sections ∂V_1 and ∂V_2, that is

$$\mathbf{T} \cdot \mathbf{n}_1 = \mathbf{p}_1 = \mathbf{0}\,; \qquad \mathbf{T} \cdot \mathbf{n}_2 = \mathbf{p}_2 = \mathbf{0}\,,$$

namely

$$\mathbf{n}_1 = -\mathbf{e}_\beta\,; \qquad \mathbf{n}_2 = \mathbf{e}_\beta\,.$$

While vector of surface load $\mathbf{p}_4$ acting on surface section ∂V_4 is

$$\mathbf{p}_4 = \mathbf{T} \cdot \mathbf{n}_4 = \mathbf{T} \cdot \mathbf{e}_\alpha = \tau(\alpha_2\,, \beta)\mathbf{e}_\varphi\,. \tag{2.23}$$

This latter agrees with Eq. (2.3).

Expanding

$$\mathbf{T} \cdot \nabla = \mathbf{0} \tag{2.24}$$

expressing the necessary condition of mechanical equilibrium yields

$$\frac{1}{H_\alpha} \frac{\partial \tau}{\partial \alpha} \mathbf{e}_\varphi + \frac{1}{H_\varphi} \tau \frac{\partial \mathbf{e}_\alpha}{\partial \varphi} + \frac{1}{H_\varphi} \tau \mathbf{e}_\varphi \left(\frac{\partial \mathbf{e}_\alpha}{\partial \varphi} \cdot \mathbf{e}_\varphi \right) +$$

$$+ \frac{1}{H_\beta} \tau \mathbf{e}_\varphi \left(\frac{\partial \mathbf{e}_\alpha}{\partial \beta} \cdot \mathbf{e}_\beta \right) = \mathbf{0}. \tag{2.25}$$

Utilizing

$$\frac{\partial \mathbf{e}_\alpha}{\partial \beta} \cdot \mathbf{e}_\beta = \frac{1}{H_\alpha} \frac{\partial H_\beta}{\partial \alpha}. \tag{2.26}$$

deduced from Eqs (1.16), (1.17), (1.18) and

$$\frac{\partial \mathbf{e}_\alpha}{\partial \beta} = \frac{\partial}{\partial \beta} \left(\frac{1}{H_\alpha} \frac{\partial \boldsymbol{\rho}}{\partial \alpha} \right) = \frac{\partial}{\partial \beta} \left(\frac{1}{H_\alpha} \right) H_\alpha \mathbf{e}_\alpha +$$

$$+ \frac{1}{H_\alpha} \frac{\partial^2 \boldsymbol{\rho}}{\partial \beta \partial \alpha} = \frac{\partial}{\partial \beta} \left(\frac{1}{H_\beta} \right) H_\alpha \mathbf{e}_\alpha +$$

$$+ \frac{1}{H_\alpha} \frac{\partial}{\partial \alpha} (H_\beta \mathbf{e}_\beta) = \frac{\partial}{\partial \beta} \left(\frac{1}{H_\alpha} \right) H_\alpha \mathbf{e}_\alpha +$$

$$+\frac{1}{H_\alpha}\frac{\partial H_\beta}{\partial \alpha}\mathbf{e}_\beta+\frac{H_\beta}{H_\alpha}\frac{\partial \mathbf{e}_\beta}{\partial \alpha} \tag{2.27}$$

yields from Eq. (2.25) the differential equation

$$\frac{1}{H_\alpha}\frac{\partial \tau}{\partial \alpha}+2\frac{\tau}{H_\alpha H_\varphi}\frac{\partial H_\varphi}{\partial \alpha}+\frac{\tau}{H_\alpha H_\varphi}\frac{\partial H_\beta}{\partial \alpha}=0 \tag{2.28}$$

for function $\tau=\tau(\alpha, \beta)$.

In writing Eq. (2.27),

$$\mathbf{e}_\beta\cdot\frac{\partial \mathbf{e}_\beta}{\partial \beta}=0 \tag{2.29}$$

has been made use of.

Combining identity

$$\frac{\partial}{\partial \alpha}(\tau H_\varphi^2 H_\beta)=\frac{\partial \tau}{\partial \alpha}H_\varphi^2 H_\beta+2\tau H_\varphi\frac{\partial H_\varphi}{\partial \alpha}H_\beta+$$

$$+\tau H_\varphi^2\frac{\partial H_\beta}{\partial \alpha}=0 \tag{2.30}$$

and differential equation (2.28) yields

$$\frac{\partial}{\partial \alpha}(\tau H_\varphi^2 H_\beta)=0. \tag{2.31}$$

General solution of (2.31) is taken in the form

$$\tau=\frac{1}{H_\varphi^2 H_\beta}\frac{\mathrm{d}U}{\mathrm{d}\beta} \tag{2.32}$$

where $U=U(\beta)$ is an arbitrary, at least once continuously differentiable, single-variable function.

Involvement of the derivative in (2.32) permits simple deduction of the expression for moment M.

Surface load $\mathbf{p}$ acting on a surface with arbitrary coordinates $\alpha=\alpha_i=$constant is of the form

$$\mathbf{p}=p(\alpha_i, \beta)\mathbf{e}_\varphi=\tau(\alpha_i, \beta)\mathbf{e}_\varphi \tag{2.33}$$

with moment $M=M(\alpha_i)$ about the z-axis

$$M=2\pi\int_{\beta_1}^{\beta_2}H_\varphi^2 H_\beta p(\alpha_i, \beta)\,\mathrm{d}\beta=2\pi\int_{\beta_1}^{\beta_2}\frac{\mathrm{d}U}{\mathrm{d}\beta}\,\mathrm{d}\beta=2\pi[U(\beta_2)-U(\beta_1)] \tag{2.34}$$

demonstrating—in agreement with conditions of mechanical equilibrium—that

$$M=M(\alpha)=\text{constant}. \tag{2.35}$$

In conformity with Eqs (2.20) and (2.32):

$$\frac{1}{H_\varphi^2 H_\beta}\frac{\mathrm{d}U}{\mathrm{d}\beta} = G\frac{H_\varphi}{H_\alpha}\frac{\mathrm{d}V}{\mathrm{d}\alpha}. \tag{2.36}$$

Existence of this latter condition guarantees that the field of deformation produced from the stress tensor field of the form (2.21) meeting equilibrium equations meets also conditions of compatibility. That is, the problem has only a solution if Eq. (2.36) exists. Let us have an orthogonal curvilinear coordinate system $\alpha\beta$ such that in possession of suitable functions

$$a = a(\alpha) \qquad \text{and} \qquad b = b(\beta)$$

it is:

$$\frac{H_\alpha}{H_\varphi^3 H_\beta} = \frac{a(\alpha)}{b(\beta)}. \tag{2.37}$$

In this case—elementary calculation may show that only in this case—Eq. (2.36) will have solution for functions $V = V(\alpha)$ and $U = U(\beta)$, of the form

$$V(\alpha) = \frac{C}{G}\int_{\alpha_1}^{\alpha} a(\xi)\,\mathrm{d}\xi = \frac{C}{G}A(\alpha), \tag{2.38}$$

$$U(\beta) = C\int_{\beta_1}^{\beta_2} b(\xi)\,\mathrm{d}\xi = CB(\beta). \tag{2.39}$$

Combining Eqs (2.34) and (2.39)

$$C = \frac{M}{2\pi B(\beta_2)}. \tag{2.40}$$

In conformity with boundary conditions (2.2), function $V = V(\alpha)$ obtained form (2.38) meets boundary condition

$$V(\alpha_1) = 0. \tag{2.41}$$

Utilizing (2.38) and (2.40) it is easy to show that

$$V(\alpha) = \frac{M}{2\pi G}\frac{A(\alpha)}{B(\beta_2)}. \tag{2.42}$$

Deformation energy W of hollow core solid of revolution is obtained from:

$$W = \frac{1}{2}\int_{\partial V_4} \mathbf{p}_4 \cdot \mathbf{u}\,\mathrm{d}\partial V_4 = \frac{1}{2}2\pi\int_{\beta_1}^{\beta_2} V(\alpha_2)\tau(\alpha_2,\beta)H_\varphi^2 H_\beta\,\mathrm{d}\beta =$$

$$= \frac{1}{2}V(\alpha_2)2\pi\int_{\beta_1}^{\beta_2}\tau(\alpha_2,\beta)H_\varphi^2 H_\beta\,\mathrm{d}\beta = \frac{1}{2}MV(\alpha_2) \tag{2.43}$$

written by making use of

$$M = 2\pi \int_{\beta_1}^{\beta_2} \tau(\alpha_2, \beta) H_\varphi^2 H_\beta \, \mathrm{d}\beta. \tag{2.44}$$

Combining (2.42) and (2.43) yields

$$W = \frac{1}{2} M^2 \frac{A(\alpha_2)}{2\pi G B(\beta_2)}. \tag{2.45}$$

Specified magnitude

$$S = \frac{M}{V(\alpha_2)} \tag{2.46}$$

is termed torsion stiffness of the hollow-core solid of revolution.

Correctness of

$$S = 2\pi G \frac{B(\beta_2)}{A(\alpha_2)} \tag{2.47}$$

and

$$W = \frac{1}{2} \frac{M^2}{S} \tag{2.48}$$

is obvious.

Utilizing (2.20) and (2.44) it may be written

$$M = \left(2\pi G \int_{\beta_1}^{\beta_2} \frac{H_\varphi^3 H_\beta}{H_\alpha} \, \mathrm{d}\beta \right) \frac{\mathrm{d}V}{\mathrm{d}\alpha}. \tag{2.49}$$

Integrating (2.49), since $M = \text{constant}$,

$$V(\alpha_1) = 0$$

we obtain

$$V(\alpha) = \frac{M}{2\pi G} \int_{\alpha_1}^{\alpha} \frac{\mathrm{d}\alpha}{\int_{\beta_1}^{\beta_2} \frac{H_\varphi^3 H_\beta}{H_\alpha} \, \mathrm{d}\beta}. \tag{2.50}$$

This formula directly yields

$$S = \frac{2\pi G}{\int_{\alpha_1}^{\alpha_2} \frac{\mathrm{d}\alpha}{\int_{\beta_1}^{\beta_2} \frac{H_\varphi^3 H_\beta}{H_\alpha} \, \mathrm{d}\beta}} \tag{2.51}$$

It should be stressed that these formulae refer only to the case where Eq. (2.37) is valid.

Direct consequence of inequalities

$$H_\alpha \geq 0, \quad H_\varphi \geq 0, \quad H_\beta \geq 0$$

is non-negativity of S.

3. Isometric orthogonal curvilinear coordinate system

Let us consider analytic complex variable function

$$r + iz = f(\alpha + i\beta). \tag{3.1}$$

Its curves $\alpha = \text{constant}$, $\beta = \text{constant}$ define an orthogonal curvilinear coordinate system on plane rz.

Utilizing fundamental results of the complex theory of functions

$$f'(\xi) = \frac{\partial r}{\partial \alpha} + i\frac{\partial z}{\partial \alpha} = -i\frac{\partial r}{\partial \beta} + \frac{\partial z}{\partial \beta} \tag{3.2}$$

where

$$\xi = \alpha + i\beta. \tag{3.3}$$

In conformity with the Cauchy–Riemann equations,

$$\frac{\partial r}{\partial \alpha} = \frac{\partial z}{\partial \beta}, \tag{3.4}$$

$$\frac{\partial z}{\partial \alpha} = -\frac{\partial r}{\partial \beta}. \tag{3.5}$$

It is easy to demonstrate therefrom that

$$\frac{\partial r}{\partial \alpha}\frac{\partial r}{\partial \beta} + \frac{\partial z}{\partial \alpha}\frac{\partial z}{\partial \beta} = 0 \tag{3.6}$$

there is, in fact, an orthogonal, curvilinear coordinate system.

Expressions for the Lamé coefficients:

$$H_\alpha = \sqrt{\left(\frac{\partial r}{\partial \alpha}\right)^2 + \left(\frac{\partial z}{\partial \alpha}\right)^2} = |f'(\xi)|, \tag{3.7}$$

$$H_\beta = \sqrt{\left(\frac{\partial r}{\partial \beta}\right)^2 + \left(\frac{\partial z}{\partial \beta}\right)^2} = |f'(\xi)|. \tag{3.8}$$

Since

$$H_\alpha = H_\beta = H$$

it is a so called isometric curvilinear coordinate system. In this case, Eq. (2.37) yields

$$\frac{a(\alpha)}{b(\beta)} = \frac{1}{[r(\alpha, \beta)]^3}. \tag{3.9}$$

That is, for every case where $r(\alpha, \beta)$ can be written as product of two functions α and β a solution be constructed.

Formula of torsional rigidity will be rather simple in case of an isometric orthogonal curvilinear coordinate system

$$S = \frac{2\pi G}{\int_{\alpha_1}^{\alpha_2} \frac{\mathrm{d}\alpha}{\int_{\beta_1}^{\beta_2} [r(\alpha, \beta)]^3 \, \mathrm{d}\beta}}. \tag{3.10}$$

4. Examples

4.1. Let us consider the orthogonal curvilinear coordinate system defined by

$$\alpha = \arctan \frac{z}{r}, \tag{4.1}$$

$$\beta = \sqrt{r^2 + z^2}. \tag{4.2}$$

Meridian section for these orthogonal curvilinear coordinates is seen in Fig. 4. Making use of arc element expressions

$$\mathrm{d}s_\alpha = H_\alpha \, \mathrm{d}\alpha = \beta \, \mathrm{d}\alpha, \tag{4.3}$$

$$\mathrm{d}s_\beta = H_\beta \, \mathrm{d}\beta \tag{4.4}$$

it may be written:

$$H_\alpha = \beta, \tag{4.5}$$

$$H_\beta = 1. \tag{4.6}$$

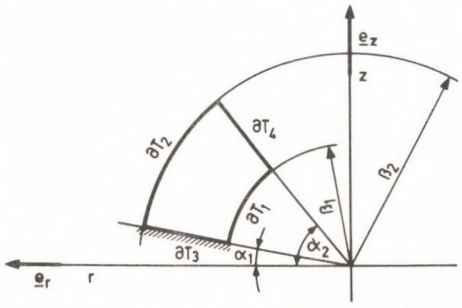


Fig. 4. Meridian section bounded by circular arcs and straight lines

Obviously

$$r = H_\varphi = \beta \cos \alpha. \tag{4.7}$$

Combining Eqs (2.37), (2.38), (2.39), (2.40), (4.5), (4.6) and (4.7)

$$\frac{H_\alpha}{H_\beta H_\varphi^3} = \frac{a(\alpha)}{b(\beta)} = \frac{1}{\beta^2} \frac{1}{\cos^3 \alpha}, \tag{4.8}$$

$$a(\alpha) = \frac{1}{\cos^3 \alpha}, \qquad b(\beta) = \beta^2, \tag{4.9) (4.10}$$

$$A(\alpha) = \int_{\alpha_1}^{\alpha_2} \frac{1}{\cos^3 \alpha} \, \mathrm{d}\alpha = \frac{\sin \alpha}{2 \cos^2 \alpha} - \frac{\sin \alpha_1}{2 \cos^2 \alpha_1} + \frac{1}{2} \ln \left| \frac{\tan\left(\frac{\pi}{4} + \frac{\alpha}{2}\right)}{\tan\left(\frac{\pi}{4} + \frac{\alpha_1}{2}\right)} \right|, \tag{4.11}$$

$$B(\beta) = \int_{\beta_1}^{\beta} \beta^2 \, \mathrm{d}\beta = \frac{\beta^3 - \beta_1^3}{3}, \tag{4.12}$$

$$V(\alpha) = \frac{M}{4\pi G} \frac{3}{\beta_2^3 - \beta_1^3} \left(\frac{\sin \alpha}{\cos^2 \alpha} - \frac{\sin \alpha_1}{\cos^2 \alpha_1} + \frac{1}{2} \ln \left| \frac{\tan\left(\frac{\pi}{4} + \frac{\alpha}{2}\right)}{\tan\left(\frac{\pi}{4} + \frac{\alpha}{2} 1\right)} \right| \right), \tag{4.13}$$

$$U = \frac{3M}{2\pi} \frac{\beta^3 - \beta_1^3}{\beta_2^3 - \beta_1^3}, \tag{4.14}$$

$$S = \frac{4\pi G(\beta_2^3 - \beta_1^3)}{3\left(\frac{\sin \alpha_2}{\cos^2 \alpha_2} - \frac{\sin \alpha_1}{\cos^2 \alpha_1} + \frac{1}{2} \ln \left| \frac{\tan\left(\frac{\pi}{4} + \frac{\alpha_2}{2}\right)}{\tan\left(\frac{\pi}{4} + \frac{\alpha_1}{2}\right)} \right| \right)}. \tag{4.15}$$

While Eqs (2.32), (4.14) yield

$$\tau = \frac{3M}{2\pi(\beta_2^3 - \beta_1^3)} \frac{1}{\cos^2 \alpha} = \frac{3M}{2\pi(\beta_2^3 - \beta_1^3)} \frac{r^2 + z^2}{z^2}. \tag{4.16}$$

4.2. Let

$$\alpha = \sqrt{r^2 + z^2}, \tag{4.17}$$

$$\beta = \mathrm{Ar} \tan \frac{z}{r}. \tag{4.18}$$

In this case, surfaces of revolution ∂V_1 and ∂V_2 will be cones or revolution with common vertex (Fig. 5).

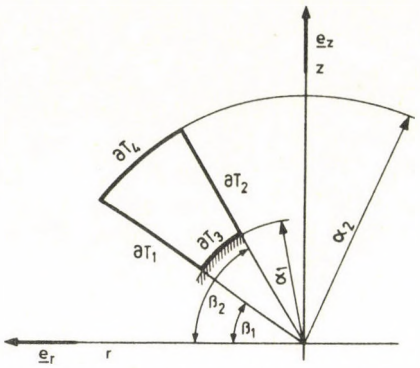


Fig. 5. Meridian section of a solid bounded by conic surfaces

Obviously, in the actual case

$$H_\alpha = 1\,, \tag{4.19}$$

$$H_\varphi = \alpha \cos\beta\,, \tag{4.20}$$

$$H_\beta = \alpha\,, \tag{4.21}$$

$$a(\alpha) = \frac{1}{\alpha^4}\,, \tag{4.22}$$

$$b(\beta) = \cos^3\beta\,, \tag{4.23}$$

$$A(\alpha) = -\frac{1}{3}\left(\frac{1}{\alpha^3} - \frac{1}{\alpha_1^3}\right), \tag{4.24}$$

$$B(\beta) = \sin\beta - \frac{1}{3}\sin^3\beta - \sin\beta_1 + \frac{1}{3}\sin^3\beta_1\,, \tag{4.25}$$

$$C = \frac{M}{2\pi\left[(\sin\beta_2 - \sin\beta_1) + \frac{1}{3}(\sin^3\beta_1 - \sin^3\beta_2)\right]}, \tag{4.26}$$

$$S = 2\pi G \frac{3(\sin\beta_2 - \sin\beta_1) - (\sin^3\beta_2 - \sin^3\beta_1)}{\frac{1}{\alpha_1^3} - \frac{1}{\alpha_2^3}}, \tag{4.27}$$

$$\tau_{\alpha\varphi} = \frac{M}{2\pi\alpha^3} \cdot \frac{\cos\beta}{\sin\beta_2 - \frac{1}{3}\sin^3\beta_2 - \sin\beta_1 + \frac{1}{3}\sin^3\beta_1} =$$

$$= \frac{M}{2\pi\left(\sin\beta_2 - \frac{1}{3}\sin^3\beta_2 - \sin\beta_1 + \frac{1}{3}\sin^3\beta_1\right)} \frac{r}{(r^2+z^2)^2}. \tag{4.28}$$

Substituting $\beta_2 = \pi/2$, the deduced formulae lend themselves for solids of revolution with conic surfaces.

4.3. Meridian curve of a hollow-core solid of revolution bounded by ellipse and hyperbola arcs is seen in Fig. 6.

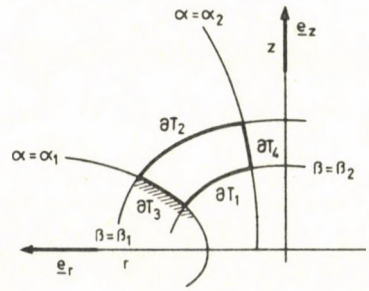


Fig. 6. Meridian section bounded by ellipse and hyperbola arcs

Actually:

$$r = D \sin\alpha \cosh\beta, \tag{4.29}$$

$$z = D \cos\alpha \sinh\beta. \tag{4.30}$$

Orthogonal curvilinear coordinate system defined by (4.29) and (4.30) may also be given by complex function

$$r + iz = D \sin(\alpha + i\beta). \tag{4.31}$$

Elementary calculation yields

$$H_\alpha = H_\beta = D\sqrt{\sinh^2\beta + \cos^2\alpha}, \tag{4.32}$$

$$H_\varphi = D \sin\alpha \cosh\beta, \tag{4.33}$$

$$a(\alpha) = \frac{1}{\sin^3\alpha}, \tag{4.34}$$

$$b(\beta) = D^3 \cosh^3\beta, \tag{4.35}$$

$$A(\alpha) = -\frac{\cos\alpha}{2\sin^2\alpha} + \frac{\cos\alpha_1}{2\sin^2\alpha_1} + \frac{1}{2}\ln\left|\frac{\tan\frac{\alpha}{2}}{\tan\frac{\alpha_1}{2}}\right|, \tag{4.36}$$

$$B(\beta)=D^3\left(\sinh\beta+\frac{1}{3}\sinh^3\beta-\sinh\beta_1-\frac{1}{3}\sinh^3\beta_1\right), \tag{4.37}$$

$$C=\frac{M}{2\pi D^3\left(\sinh\beta_2+\frac{1}{3}\sinh^3\beta_2-\sinh\beta_1-\frac{1}{3}\sinh^3\beta_1\right)}, \tag{4.38}$$

$$V(\alpha)=\frac{M}{4\pi GD^3}\frac{-\frac{\cos\alpha}{\sin^2\alpha}+\frac{\cos\alpha_1}{\sin^2\alpha_1}+\ln\left|\frac{\tan\frac{\alpha}{2}}{\tan\frac{\alpha_1}{2}}\right|}{\left(\sinh\beta_2+\frac{1}{3}\sinh^3\beta_2-\sinh\beta_1-\frac{1}{3}\sinh^3\beta_1\right)}, \tag{4.39}$$

$$S=4\pi GD^3\frac{\sinh\beta_2+\frac{1}{3}\sinh^3\beta_2-\sinh\beta_1-\frac{1}{3}\sinh^3\beta_1}{-\frac{\cos\alpha_2}{\sin^2\alpha_2}+\frac{\cos^2\alpha_1}{\sin^2\alpha_1}+\ln\left|\frac{\tan\frac{\alpha_1}{2}}{\tan\frac{\alpha_2}{2}}\right|}, \tag{4.40}$$

$$\tau=\frac{M}{2\pi D^3}\frac{1}{\sinh\beta_2+\frac{1}{3}\sinh^3\beta_2-\sinh\beta_1-\frac{1}{3}\sinh^3\beta_1}\cdot\frac{\cosh\beta}{\sin^2\alpha\sqrt{\sinh^2\alpha+\cos^2\alpha}}. \tag{4.41}$$

4.4. Let us consider the orthogonal curvilinear coordinate system in plane rz, defined by

$$f(\zeta)=-iD\zeta^2=-iD(\alpha^2-\beta^2)+2D\alpha\beta\,. \tag{4.42}$$

Obviously,

$$r=2D\alpha\beta\,, \tag{4.43}$$

$$z=D(\beta^2-\alpha^2)\,. \tag{4.44}$$

Eqs (4.43) and (4.44) show in the actual case coordinate lines to be parabolae normal to each other (Fig. 7).

Elementary calculation yields

$$H_\alpha=H_\beta=H=|f'(\zeta)|=2D\sqrt{\alpha^2+\beta^2}\,, \tag{4.45}$$

$$H_\varphi=2D\alpha\beta\,, \tag{4.46}$$

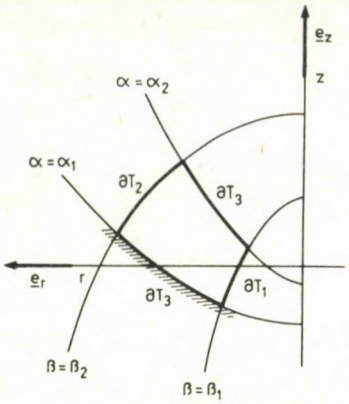


Fig. 7. Meridian section bounded by parabola arcs

$$a(\alpha)=\frac{1}{\alpha^3}, \tag{4.47}$$

$$b(\beta)=8D^3\beta^3, \tag{4.48}$$

$$A(\alpha)=\frac{1}{2}\left(\frac{1}{\alpha_1^2}-\frac{1}{\alpha^2}\right), \tag{4.49}$$

$$B(\beta)=2D^3(\beta^4-\beta_1^4), \tag{4.50}$$

$$C=\frac{M}{4\pi D^3(\beta_2^4-\beta_1^4)}, \tag{4.51}$$

$$V(\alpha)=\frac{M}{8\pi D^3 G(\beta_2^4-\beta_1^4)}\left(\frac{1}{\alpha_1^2}-\frac{1}{\alpha^2}\right), \tag{4.52}$$

$$S=8\pi D^3 G\frac{\beta_2^4-\beta_1^4}{\dfrac{1}{\alpha_1^2}-\dfrac{1}{\alpha_2^2}}, \tag{4.53}$$

$$\tau=\frac{M}{4\pi D^3(\beta_2^4-\beta_1^4)}\frac{\beta}{\alpha^2\sqrt{\alpha^2+\beta^2}}. \tag{4.54}$$

5. Comments

5.1. Figure 8 represents meridian section of a solid of revolution bounded by concentrical spherical surfaces and a conic surface. In conformity with Fig. 8

$$\tau_{z\varphi}=\tau\sin\beta, \tag{5.1}$$

$$\tau_{r\varphi}=\tau\cos\beta, \tag{5.2}$$

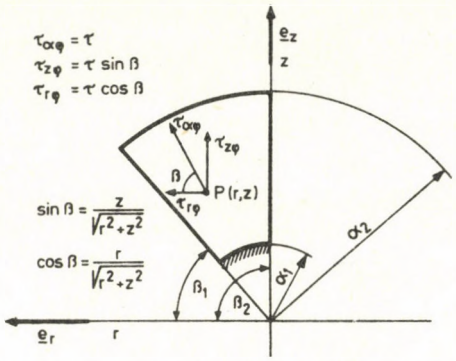


Fig. 8. Meridian section of a solid bar with conic surface

$$\cos\beta = \frac{r}{\sqrt{r^2+z^2}}, \tag{5.3}$$

$$\sin\beta = \frac{z}{\sqrt{r^2+z^2}}. \tag{5.4}$$

Utilizing Eq. (4.28) substituting $\beta_2 = \pi/2$, as well as Eqs (5.1), (5.2), (5.3), (5.4) yields

$$\tau_{z\varphi} = \frac{M}{2\pi\left(\frac{2}{3} - \sin\beta_1 + \frac{1}{3}\sin^3\beta_1\right)} \frac{rz}{(r^2+z^2)^{5/2}}, \tag{5.5}$$

$$\tau_{r\varphi} = \frac{M}{2\pi\left(\frac{2}{3} - \sin\beta_1 + \frac{1}{3}\sin^3\beta_1\right)} \frac{r^2}{(r^2+z^2)^{5/2}}. \tag{5.6}$$

The same result as that obtained from Eq. (5.5) is obtained by combining relationships (*p*) and (*q*) on p. 309 of Ref. [3].

5.2. Expression for the displacement vector

$$\mathbf{u} = H_\varphi V(\alpha)\mathbf{e}_\varphi = rV(\alpha)\mathbf{e}_\varphi \tag{5.7}$$

shows coordinate surfaces $\alpha =$ constant to rotate rigid bodylike at an angle $V(\alpha)$ about the z-axis as axis of revolution.

5.3. In conformity with Eq. (5.7), the considered problem of torsion of a solid of revolution bounded by coordinate surfaces of an orthogonal curvilinear coordinate system (α, φ, β) is defined by the following static and kinematic requirements:

$$\mathbf{u}(P) = \mathbf{0} \qquad P \in \partial V_3, \tag{5.8}$$

$$\mathbf{p}(P) = \mathbf{0} \qquad P \in \partial V_1 \cup \partial V_2, \tag{5.9}$$

$$\mathbf{u}(P)=r\vartheta\mathbf{e}_{\varphi} \qquad P\in\partial V_4\,, \tag{5.10}$$

$$\mathbf{q}(P)=\mathbf{0} \qquad P\in V. \tag{5.11}$$

It should be noted that:

$$\vartheta=V(\alpha_2)\,. \tag{5.12}$$

References

1. Lurje A. I.: Teorija uprugosti. Izd. Nauka Glav. red. fiz-mat. lit. Moskva 1971 850–891 cmp
2. Frank-Mises: Die Differetial- und Integralgleichungen der Mechanik und Physik. Dover Publications, Inc. New York 1961, S. 82–86
3. Timoshenko S. and Goodier I. N.: Theory of Elasticity. Mc. Graw Hill. Book. Comp. Inc. 1951. p. 309
4. Lamé G.: Leçons sur les Coordonnées Curvilignes, Gauthier- Villars. Paris 1859

Acta Technica Academiae Scientiarum Hungaricae, 98 (3—4), pp. 295—307 (1985)

THE EFFECT OF SOME IMPERFECTIONS ON THE STRESS OF ONE-BAY THIN-WALLED CHANNEL PURLINS WORKING TOGETHER WITH CORRUGATED PLATE-COVER

B. GOSOWSKI*, E. KUBICA*, K. RYKALUK*

[Received: 6 March 1984]

A solution of the construction built up of a thin-walled channel purlin initially twisted or stiffened discreetly against torsion and corrugated sheets has been presented. The purlin is elastically restrained at a covering made of corrugated sheets and has an upper flange connected with a covering. Algorithm of the strength calculation of the construction has been derived based on the second order theory. The effect of the initial torsion and discreet stiffeners on the strain of the cold-formed purlin has been estimated. Theoretical results have been verified on the part of the full-scale roof model.

1. Introduction

Screw joints between corrugated plates and U-purlins commonly used because of their many advantages, are characterized by plastic strains which occur following the first loading thus producing, in effect, the torsional deflection of the purlins [1]. The value of the twist depends on the level of the first loading. The torsional deflection that remains after the first unloading may be treated as initial deflection (geometric imperfection) for the next loading cycles. Therefore, the problem arises of estimating the effect of this imperfection on the stress of purlins observed in the next loading cycles. The solutions of this problem that can be found in the literature (cf. e.g. [2]) cannot be directly applied in this case since they consider bisymmetric I-purlins only.

In [1] it was shown that the torsional deflection of U-purlin affected the state of displacements to a greater degree than the state of stresses. Hence, it is very often the case that the horizontal displacements of the bottom flange resulting from purlin twist has to be reduced. One of the recent constructional solutions which stiffens U-purlins with sheet cover was suggested in [3]. The solution was based on experimental observations. And again, the problem arises of how point stiffening (constructional imperfections) affects the stress of the purlins.

In this work, a solution is presented for a thin-walled U-purlin working with corrugated plate-cover. The purlin was pre-deflected torsionally or point-stiffened. Theoretical results are verified on the full-scale model of a fragment of a roof.

* Institute of Building Engineering of Wrocław Technical University. Wybrzeże Wyspiańskiego 27. 50-370 Wrocław. Poland

2. Static-and-strength analysis of a pre-deflected purlin

2.1. Differential equations of equilibrium

For an elementary section of the purlin with defined axis of revolution (point N in Fig. 1) and torsional pre-deflection $\Phi_0(x)$, the following set of differential equations of equilibrium has been deriveed:

$$EI_y w'''' + EI_z z_N[(\Phi + \Phi_0)\Phi'']'' = q_z, \tag{1a}$$

$$(EI_\omega + EI_z z_N^2)\Phi'''' - GI_d\Phi'' + k_\Phi\Phi + 2b_y z_N EI_z[(\Phi + \Phi_0)' \cdot \Phi'']' -$$

$$- 2z_N EI_y[(\Phi + \Phi_0)''w'' + (\Phi + \Phi_0)'w'''] + EI_z z_N(w''\Phi'' - w\Phi'') = q_z(y_N - y_S). \tag{1b}$$

where: E, G are Young's modulus of elasticity and shear modulus of elasticity respectively: I_y, I_z, I_ω, I_d are the moments of inertia about axes y, z the sector moment and the pure torsional moment, respectively; w, Φ are linear and angular displacements

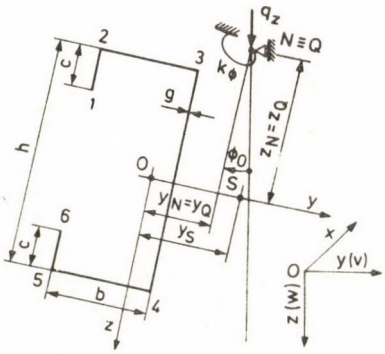


Fig. 1. Cross section of U-purlin with pre-twist

measured with respect to axis y in deformed state and around axis x, respectively; Φ_0 is purlin's torsional pre-deflection; k_Φ is the coefficient of elastic fixing of the purlin in cover:

$$b_y = \frac{1}{2I_z}\int_F y(y^2 + z^2)\,dF - y_S;$$

F represents the purlin's cross sectional area. All the remaining symbols are presented in Fig. 1.

2.2. *Solution of the set of equations for a purlin with free-ends*

For a purlin of length ℓ which, due to linear and angular displacements, has free supports at both ends, the boundary conditions of the set (1) are as follows:

$$w(0)=w(l)=0, \qquad w''(0)=w''(l)=0, \tag{2a}$$

$$\Phi(0)=\Phi(l)=0, \qquad \Phi''(0)=\Phi''(l)=0. \tag{2b}$$

Here, the initial torsion can be described by the function:

$$\Phi_0=a_0\sin\frac{\pi x}{l}. \tag{3}$$

The set (1) with boundary conditions (2) for Φ_0 and acc. to (3) was solved with Bubnov–Galerkin orthogonalization method assuming the functions of displacements in the form of the series:

$$\Phi=\sum_{n=1}^{\infty}a_n\sin\frac{n\pi x}{l}. \qquad w=\sum_{n=1}^{\infty}b_n\sin\frac{n\pi x}{l}. \tag{4}$$

For the load distributed along the purlin $q_z=$const. and for the coefficient of elastic fixing $k_\Phi=$const., the shape of the purlin's deformation stays symmetric in relation to half-span ($x=l/2$). Then, in (4), only the terms with odd n should be taken into consideration. From the ortogonality conditions, the following set of non-linear algebraic equations with regard to a_n and b_n coefficients of series (4) was obtained:

$$p^5b_p-\frac{8}{\pi}z_N\frac{II_z}{I_y}\left[\sum_{n=1}^{\infty}\sum_{r=1}^{\infty}\frac{n^3p^4r}{(n^2-p^2+r^2)^2-4n^2r^2}a_na_r+\right.$$

$$\left.+a_0\sum_{n=1}^{\infty}\frac{n^3p^4}{(n^2-p^2+1)^2-4n^2}a_n\right]=\frac{4l^4q_z}{\pi^5EI_y}. \tag{5a}$$

$$\left[\left(\frac{I_\omega}{I_y}+z_N^2\frac{I_z}{I_y}\right)p^5+\frac{GI_dl^2}{\pi^2EI_y}p^3+\frac{k_\Phi l^4}{\pi^4EI_y}p\right]a_p-$$

$$-\frac{8}{\pi}z_Nb_y\frac{I_z}{I_y}\left[\sum_{n=1}^{\infty}\sum_{r=1}^{\infty}\frac{n^3p^2r(p^2+r^2-n^2)}{(n^2-p^2+r^2)^2-4n^2r^2}a_na_r+\right.$$

$$\left.+a_0\sum_{n=1}^{\infty}\frac{n^3p^2(p^2-n^2+1)}{(n^2-p^2+1)^2-4n^2}a_n\right]+$$

$$+\frac{8}{\pi}z_N\sum_{n=1}^{\infty}\sum_{r=1}^{\infty}\frac{np^2r[n^2p^2-(n^2-r^2)(n^2+r^2I_z/I_y)]}{(n^2-p^2+r^2)^2-4n^2r^2}a_rb_n+$$

$$+\frac{8}{\pi}z_Na_0\sum_{n=1}^{\infty}\frac{n^3p^2(p^2-n^2+1)}{(n^2-p^2+1)^2-4n^2}b_n=\frac{4l^4q_z}{\pi^5EI_y}(y_N-y_S). \tag{5b}$$

where n, p, r are odd.

2.3. Determination of the displacements and stresses

After finding a_n and b_n from (5), the displacements Φ and w of the purlin can be determined on the basis of (4) and the displacement v on the basis of Φ as follows:

$$v = z_N \Phi = z_N \sum_{n=1}^{\infty} a_n \sin \frac{n\pi x}{l}. \tag{6}$$

Bending moments M_y and M_z and bimoment B are derived from the known differential equations:

$$M_y = -EI_y w'', \qquad M_z = EI_z v'' = EI_z z_N \Phi'', \qquad B = -EI_\omega \Phi''. \tag{7}$$

Normal stresses in fibre j defined by the co-ordinates y_j, z_j, ω_j of the purlin's cross section are calculated as follows:

$$\sigma_j = \frac{M_y}{I_y} z_j - \frac{M_z}{I_z} y_j + \frac{B}{I_\omega} \omega_j. \tag{8}$$

2.4. Numerical example

Numerical calculations were performed for a $\mathbb{C}$ $180 \times 65 \times 25/2.5$ U-purlin. The cross sectional dimensions of the U-bar (cf. Fig. 1) were $b = 6.25$ [cm], $c = 2.375$ [cm], $h = 17.75$ [cm], $g = 0.25$ [cm]. The span of the purlin was assumed $l = 5.88$ [m] and the co-ordinates of the predetermined axis of revolution were $y_N = 1.465$ [cm], $z_N = -9.0$ [cm]. Geometric characteristics of the cross section and the material constant are to be found in [1].

By assuming $n = 1$ in (4) one arrives at the set (5) in the form:

$$\frac{8}{3\pi} z_N \frac{I_z}{I_y} (a_1^2 + a_0 a_1) + b_1 = \frac{4l^4 q_z}{\pi^5 EI_y}, \tag{9a}$$

$$\left(\frac{I_\omega}{I_y} + z_N^2 \frac{I_z}{I_y} + \frac{GI_d l^2}{\pi^2 EI_y} + \frac{k_\Phi l^4}{\pi^4 EI_y} \right) a_1 + \frac{8}{3\pi} b_y z_N \frac{I_z}{I_y} (a_1^2 + a_0 a_1) -$$

$$- \frac{8}{3\pi} z_N (a_1 b_1 + a_0 b_1) = \frac{4l^4 q_z}{\pi^5 EI_y} (y_N - y_S). \tag{9b}$$

This set was solved for selected values of q_z and k_Φ while assuming different values of the purlin's pre-twist $a_0 = \Phi_{0\max}$. The increment of the angle of twist in the purlin's mid-span $\Phi(l/2) = \Phi_{\max}$ depending on q_z, k_Φ, and a_0, is presented in Fig. 2. It is clear that the purlin's pre-twist may have a varying effect on angle Φ depending on coefficient k_Φ. For a definite k_Φ, this effect is controlled by the sign of angle Φ_0 (positive Φ_0 increases angle Φ while negative decreases it).

For $q_z = 2.608$ [kN/m] and $k_\Phi = 1.0$ [kNm/m] and $\Phi_0 = -0.015$ [rad], the values for $a_1 = -0.0810$ [rad] and $b_1 = 4.59$ [cm] were obtained from (9). The displacement

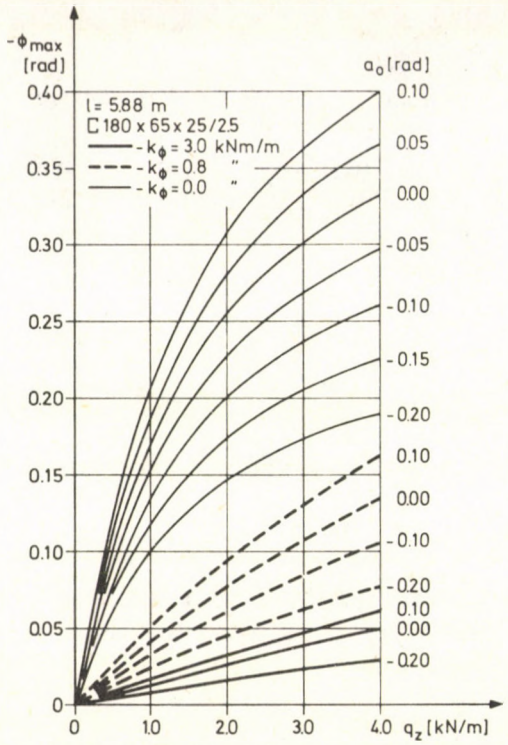


Fig. 2. The increment of the angle of torsion Φ of the purlin's cross section at its half-span depending on q_z, k_Φ and a_0

components in the purlin's mid-span thus are: $\Phi_{max} = -0.0810$ [rad], $w_{max} = 4.59$ [cm] $v_{max} = 0.73$ [cm]. The bending moment and the bimoment in the purlin's mid-span calculated from (7) are: $M_{y\,max} = 11.645$ [kNm]. $M_{z\,max} = -0.228$ [kNm], $B_{max} = -0.018$ [kNm^2]. This makes it possible to calculate the stresses from (8) at particular points of the purlin's cross section (points from *1* to *6* in Fig. 1). Results are presented in Table 1.

Table 1

Normal stresses in the central section of pre-twisted U-purlin C180 × 65 × 25/2.5 ($l = 5.88$ [m], $q_z = 2.608$ [kN/m], $k_\Phi = 1.0$ [kNm/m], $a_0 = -0.015$ rad)

Point j of the cross section (cf. *Fig. 1*)		1	2	3	4	5	6
$\sigma_j^y = M_y z_j / I_y$	[N/mm^2]	−174.6	−238.5	−238.5	238.5	238.5	174.6
$\sigma_j^z = -M_z y_j / I_z$	[N/mm^2]	−18.3	−18.3	8.4	8.4	−18.3	−18.3
$\sigma_j^\omega = B\omega_j / I_\omega$	[N/mm^2]	23.8	13.3	−13.0	13.0	−13.3	−23.8
$\sigma_j = \sigma_j^y + \sigma_j^z + \sigma_j^\omega$	[N/mm^2]	−169.1	−243.5	−243.1	259.9	206.9	132.5

3. Static and strength analysis of anti-twist point-stiffened purlin

3.10. Differential equation of equilibrium

For a purlin with a fixed axis of revolution of co-ordinates y_N, z_N which works with cover and is fitted with additional elastic anti-twist stiffenings spaced along the span at m points, the set of differential equations of equilibrium [4] is written as:

$$EI_y w'''' + z_N EI_z(\Phi''\Phi)'' = q_z, \tag{10a}$$

$$(EI_\omega + z_N^2 EI_z)\Phi'''' - GI_d\Phi'' + k_\Phi\Phi + \sum_{i=1}^{m} K_{\Phi i}\Phi(x_i)\delta(x-x_i) +$$

$$+ 2b_y z_N EI_z(\Phi''\Phi')' - 2z_N EI_y(w''\Phi'' + w'''\Phi') +$$

$$+ z_N EI_z(w''\Phi'' - w\Phi'''') = q_z(y_N - y_S), \tag{10b}$$

where $K_{\Phi i}$ is the elastic constant of the i-th point-stiffening of the purlin located at distance x_i from the support, $\delta(x-x_i)$ is Dirac distribution.

3.2. Solution of the set of equations for free-supported purlin

For a purlin with length l and with boundary conditions (2), the set of equations (10) was solved by Bubnov–Galerkin orthogonality method assuming the functions of displacements to take the form of (4) series.

In the case when the load is uniformly distributed along the purlin $q_z = \text{const.}$, the coefficient of elastic fixing $k_\Phi = \text{const.}$, and identical point-stiffenings are spaced symmetrically with regard to the purlin's half span ($K_{\Phi i} = K_\Phi$), then, in series (4) only those terms with odd n should be made allowances for due to a symmetric form of the purlin's deformation. From orthogonality conditions, the following set of nonlinear (with regard to coefficients a_n and b_n) algebraic equations is obtained:

$$p^5 b_p - \frac{8}{\pi} z_N \frac{I_z}{I_y} \sum_{n=1}^{\infty}\sum_{r=1}^{\infty} \frac{n^3 p^4 r}{(n^2 - p^2 + r^2)^2 - 4n^2 r^2} a_n a_r = \frac{4l^4 q_z}{\pi^5 EI_y}, \tag{11a}$$

$$\left[\left(\frac{I_\omega}{I_y} + z_N^2 \frac{I_z}{I_y}\right)p^5 + \frac{GI_d l^2}{\pi^2 EI_y} p^3 + \frac{k_\Phi l^4}{\pi^4 EI_y} p\right] a_p +$$

$$+ \frac{2K_\Phi l^3}{\pi^4 EI_y} p \sum_{i=1}^{\infty}\sum_{n=1}^{\infty} a_n \sin\frac{n\pi x_i}{l} \sin\frac{p\pi x_i}{l} -$$

$$-\frac{8}{\pi} z_N b_y \frac{I_z}{I_y} \sum_{n=1}^{\infty} \sum_{r=1}^{\infty} \frac{n^3 p^2 r (p^2 + r^2 - n^2)}{(n^2 - p^2 + r^2)^2 - 4n^2 r^2} a_n a_r +$$

$$+\frac{8}{\pi} z_N \sum_{n=1}^{\infty} \sum_{r=1}^{\infty} \frac{np^2 r[n^2 p^2 - (n^2 - r^2)(n^2 + r^2 I_z/I_y)]}{(n^2 - p^2 + r^2)^2 - 4n^2 r^2} a_r b_n =$$

$$= \frac{4l^4 q_z}{\pi^5 E I_y} (y_N - y_S). \tag{11b}$$

where n, p, r are odd.

After solving (11), the displacements and stresses are found similarly as in *2.3*.

3.3. Numerical example

Numerical calculations were performed for the purlin identical to the one described in *2.4*. The purlin was additionally stiffened at one ($x_1 = l/2$) or two ($x_1 = l/3$, $x_2 = 2l/3$) points along its length. Taking only one term from each series (4), the set of equations (11) is reduced to the form:

$$\frac{8}{3\pi} z_N \frac{I_z}{I_y} a_1^2 + b_1 = \frac{4l^4 q_z}{\pi^5 E I_y}, \tag{12a}$$

$$\left(\frac{I_\omega}{I_y} + z_N^2 \frac{I_z}{I_y} + \frac{G I_d l^2}{\pi^2 E I_y} + \frac{k_\Phi l^4}{\pi^4 E I_y} + \frac{2 K_\Phi l^3}{\pi^4 E I_y} \sum_{i=1}^{m} \sin^2 \frac{\pi x_i}{l} \right) a_1 +$$

$$+\frac{8}{3\pi} b_y z_N \frac{I_z}{I_y} a_1^2 - \frac{8}{3\pi} z_N a_1 b_1 = \frac{4l^4 q_z}{\pi^5 E I_y} (y_N - y_S). \tag{12b}$$

where for the purlin stiffened only at one point $m = 1$ and $x_1 = l/2$ and for that stiffened at two points $m = 2$ and $x_1 = l/3$ and $x_2 = 2l/3$.

Figure 3 provides a graphical representation of the solution of (12) set for purlin-corrugated plate system ($k_\Phi = 1.0$ [kNm/m]) with one or two point-stiffenings of different elastic constants K_Φ. Practically, point-stiffenings have no effect on the linear displacements w but, on the other hand, they considerably reduce angular displacements Φ and, thereby, the linear displacements v.

For the purlin with one point-stiffening at its half-span ($K_\Phi = 3.0$ [kNm]), elastically fixed to the cover ($k_\Phi = 1.0$ [kNm/m]) and uniformly loaded ($q_z = 2.608$ [kN/m]), the displacement components at the purlin's half-span are: $\Phi_{max} = -0.0487$ [rad], $w_{max} = 4.59$ [cm], $v_{max} = 0.44$ [cm], the bending moments and bimoment are $M_{y\,max} = 11.638$ [kNm], $M_{z\,max} = -0.137$ [kNm], $B_{max} = -0.011$ [kNm^2]. Normal stresses at particular points of the purlin's cross section are given in Table 2.

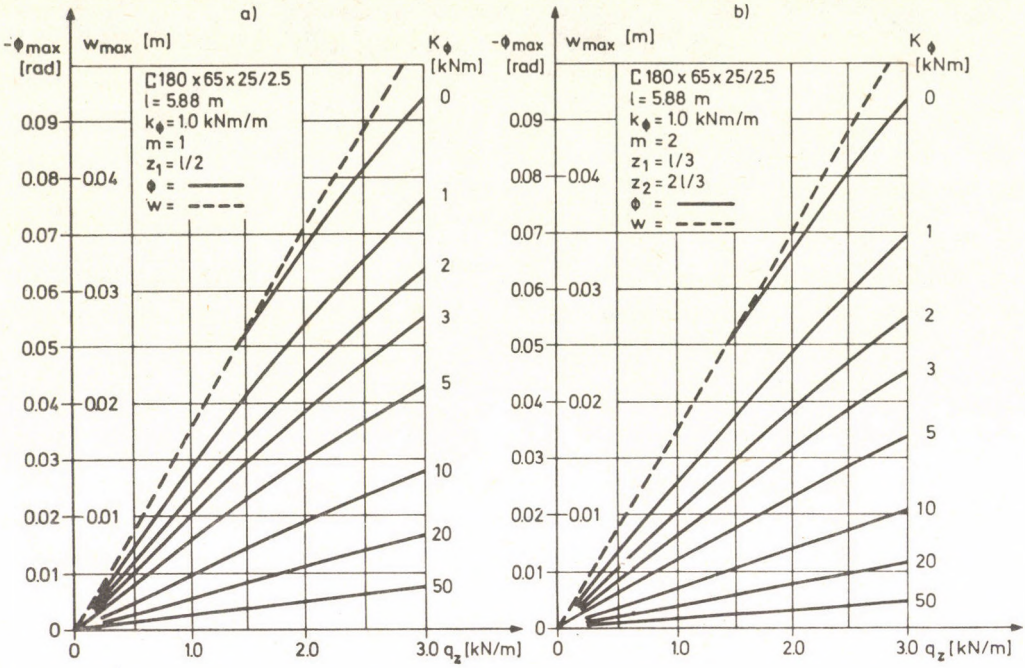


Fig. 3. The dependence of the angle of torsion of the purlin's cross section at its half-span on q_z for different values of K_Φ: a—purlin with one additional point-stiffening, *b*—purlin with two additional point-stiffenings

Table 2

Normal stresses in the central section of U-purlin C180 × 65 × 25/2.5 with additional point-stiffening at half-span ($l = 5.88$ [m], $q_z = 2.608$ [kN/m], $k_\Phi = 1.0$ [kN/m], $K_\Phi = 3.0$ [kNm])

Point j of the cross section (cf. *Fig. 1*)		1	2	3	4	5	6
$\sigma_j^y = M_y z_j / I_y$	[N/mm²]	−174.5	−238.3	−238.3	238.3	238.3	174.5
$\sigma_j^z = -M_z y_j / I_z$	[N/mm²]	−11.0	−11.0	5.0	5.0	−11.0	−11.0
$\sigma_j^\omega = B\omega_j / I_\omega$	[N/mm²]	14.3	8.0	−7.8	7.8	−8.0	−14.3
$\sigma_j = \sigma_j^y + \sigma_j^z + \sigma_j^\omega$	[N/mm²]	−171.2	−241.3	−241.1	251.1	219.3	149.2

4. Model verification

4.1. Model studies

The purlin was studied on a full-scale model of a fragment of the roof. The model was constructed in such a way that precise representation of the purlin-roof working conditions was possible. The purlin studied was a cold-formed U-bar C 180 × 65

$\times 25/2.5$, while the covering was made of corrugated plate T $55 \times 188 - 750$, with thickness 0.75 [mm]. The joint between the plate and the upper flange of the purlin was pre-stressed with screws B–6.3×80 at 37.5 [cm] distance using special profiles to reinforce the plates. The dimensions of the fragment of the roof were 6.0×3.6 [m].

The model was examined at the same test stand which was used in [1].

The program of tests included:

1) testing the purlin with pre-twist Φ_0 (left as a permanent set after the first cycle of loading) during two successive cycles,
2) testing the purlin with one additional point-stiffening during one cycle of loading.

The point-stiffening had the form of bracket shown in Fig. 4. It was mounted in the purlin's half-span with special connection clips as described in [3].

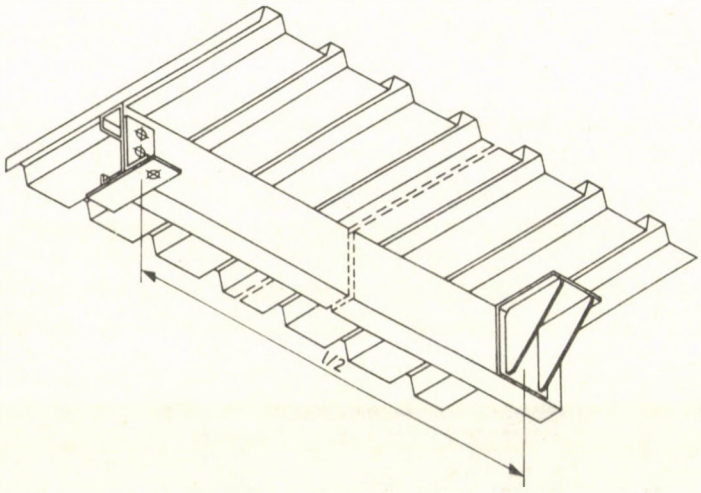


Fig. 4. Fragment of the model—bracket visible

The following parameters were measured during the tests:

— strain at selected points of the purlin's central section,
— torsion and displacements (horizontal and vertical) of the central section,
— toorsion in the purlin's support sections,
— displacements of plates in relation to the purlin's upper flange,
— vertical deflections of extreme purlins of the model.

The results are presented in Figs 5 and 6. The increment of the angle of torsion (Φ) and the vertical displacement (w) for the pre-twisted purlin $\Phi_{0\,max} = a_0 = -0.015$ [rad] are shown in Fig. 5. The same parameters but for bracket-stiffened purlin are shown in Fig. 6.

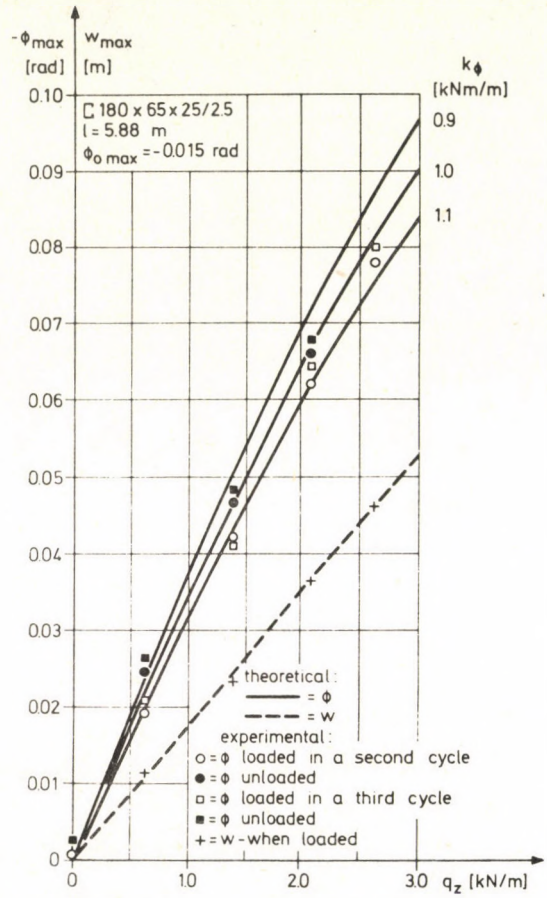


Fig. 5. Theoretical and experimental dependences of the angle of torsion and vertical displacement on loading q_z for purlin with pre-twist

4.2. Discussion of the results

The results of measurements for the pre-twisted purlin shown in Fig. 5 include two cycles of loading. Vertical displacements were practically the same in both cycles and agreed well with the results obtained theoretically from the algorithm presented in section *2.4*. Angles of torsion observed at the purlin's half-span (understood as the increments in relation to preliminary angle Φ_0) are, in both cycles, i.e., loading and unloading, similar in value and their dependence on q_z runs around convex curve. This dependence is similar to that obtained in [1] for the purlin's unloading after the first cycle. Fig 5 also presents theoretical curves found by the algorithm described in Section *2.4* in this paper, for three values of coefficient k_Φ and $\Phi_0 = -0.015$ [rad]. These results confirm the thesis put forward in [1] that the coefficient of elastic joint between the

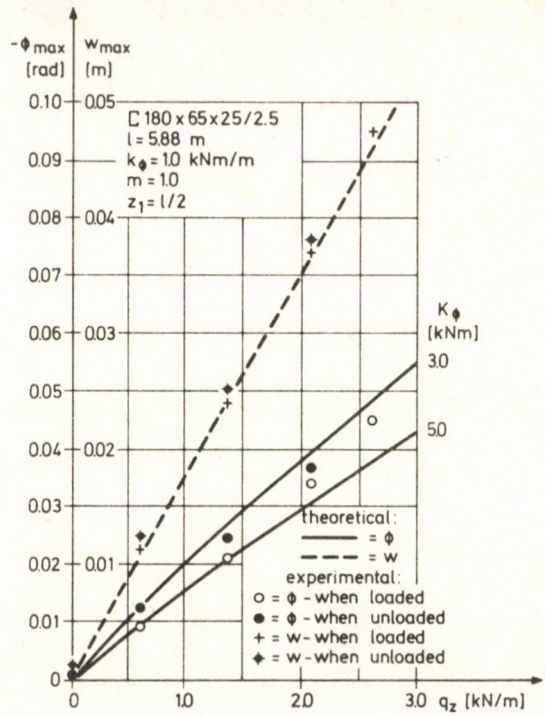


Fig. 6. Theoretical and experimental dependences of the angle of torsion and vertical displacement on loading q_z for bracket-stiffened purlin

purlin and the cover k_Φ is determined as for unloading in the first cycle (in that case $k_\Phi \approx 1$ [kNm/m]).

The results obtained for bracket-stiffened purlin are shown in Fig. 6. Vertical displacements agree well with theoretical results obtained with the algorithm presented in section *3.3* of this paper. The results of measurements of angle of torsion for loading and unloading run between theoretical curves determined from *3.3* algorithm where $k_\Phi = 1.0$ [kNm/m] and $K_\Phi = 3.0$ and 5.0 [kNm]. In the above algorithm it was assumed that the function of angle of torsion had the form of sinusoid half-wave. This assumption was confirmed by the tests.

Normal stresses σ_x in selected fibres of the central section of the pre-twisted purlin and the bracket-stiffened purlin are presented in Fig. 7 with points of stress. These normal stresses were determined from measured strain for $q_z = 2.608$ [kN/m]. For the sake of comparison, theoretical distributions of stresses σ_x calculated from the II order theory were also plotted in the Figure (cf. Tables 1 and 2) as well as the theoretical distribution of stresses caused by bending in the plane normal to the covering (acc. to elementary theory).

Stresses σ_x calculated from the II order theory for the pre-twisted purlin are almost identical across the width of the upper flange. At the lower flange, however, clear

differences occur. Stresses measured at the lower flange confirm this distribution. The maximum stresses observed in the purlin's cross section occur at point *4*.

Similar distributions of stresses were obtained for the bracket-stiffened purlin, only the values for the upper flange were smaller and less differentiated for the lower flange. This was the result of the stiffening effect of the bracket.

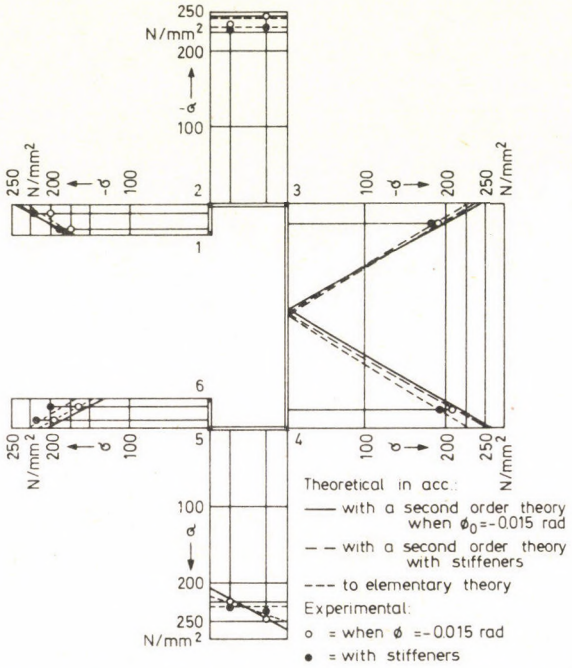


Fig. 7. The distribution of normal stresses σ_x in the central section of U-purlin [$180 \times 65 \times 25/2.5$

5. Conclusions

Thin-walled U-purlins develop both bending and torsion. The imperfections discussed in this paper (pre-twist and point-stiffening) affect, in practice, only the torsional deflexion.

The algorithms presented enable the estimation (acc. to the II order theory) of the stress of single-span U-purlins with the above mentioned imperfections.

Practically, the pre-twist of the purlin does not affect the deflections but exerts a considerable influence on the angle of torsion. The smaller is the coefficient of the purlin-cover elastic joint k_Φ the more pronounced is this effect. For a pre-determined k_Φ, the pre-twist Φ_0 may increase or decrease the purlin's angle of torsion Φ depending on whether Φ and Φ_0 have opposite or same signs. Indeed, when the signs of Φ and Φ_0

are the same, the torsional stiffness of the purlin increases but, at the same time, the forces acting at the purlin-cover joint increases, too.

Point-stiffenings of the purlin (e.g., with brackets) while having practically no effect on the deflections decrease significantly the effect of torsion, i.e., the angle of torsion (to a greater degree) and the stresses (to a smaller degree). During the tests on the sigle-bracket purlin the angle of torsion was reduced by 50% and the maximum stresses by 5%.

The coefficient of elasticity of the point-stiffening K_Φ as well as the coefficient of the purlin-cover elastic joint k_Φ should be determined experimentally for each constructional solution. In the constructional solution presented in this paper $K_\Phi \approx 4$ [kNm] was obtained.

References

1. Gosowski, B., Kubica, E., Rykaluk, K.: Beanspruchung einfeldriger dünnwandiger C-Pfetten, die mit Trapezblechen zusammenwirken. Der Stahlbau **52** (1983). 335–338
2. Oxfort, J.: Zur Kippstabilisierung stählerner I-Dachpfetten mit Imperfektionen in geneigten Dächern bis zum Erreichen der plastischen Grenzlast durch die Biege- und Schubsteifigkeit der Dacheindeckung. Der Stahlbau **45** (1976). 307–311, 365–371
3. Gosowski, B., Kubica, E., Rykaluk, K.: Stiffening of cold-rolled C-iron beams matching roofing sheets with beam holders (in Polish). Przegląd Budowlany **55** (1983). 106–108
4. Gosowski, B.: Torsion of thin-walled bars by the distribution calculus (in Polish). Archiwum Inżynierii Lądowej **25** (1979), 231–243

Acta Technica Academiae Scientiarum Hungaricae, 98 (3—4), pp. 309—316 (1985)

THE STRESS FUNCTION OF PLANE GRIDS OF A GENERAL TRIANGULAR NETWORK

I. HEGEDŰS*

[Received: 24 May 1985]

The paper shows that a stress function, analogous to Airy's stress function of elastic discs (i.e. plates subjected to in-plane loads,) can be used for analyzing stress states of plane grids of a general triangular network loaded by in-plane boundary loads. The equation of the stress function of a grid describes an open polyhedron having the same projection network as the network of the grid. The bar forces are proportional to the changes in slope of the broken surface lines of intersection of the polyhedron and of planes perpendicular to the bars in question. Having expressed the compatibility conditions of elastic deformations of the grid in terms of the stress function, a system of linear equations can be developed for solving the grid problem. The paper also shows the similarities between the grid problem and the finite difference method for solving analogous disc problems.

Introduction

The author has shown in a previous paper [2] that the stress state of a pin-jointed single-layer space grid of triangular network can be considered to be the degenerated stress state of an open polyhedron-shaped membrane shell. This stress state can be described, analogously to the stress states of the common membrane shells, by using a stress function which is the analogue of Pucher's stress function.

Taking the equation $F(x, y)$ of an arbitrary open polyhedron having the same projection network on the horizontal plane (x, y) as the grid in question, the horizontal components of bar forces determined by $F(x, y)$ as the stress function of the grid automatically meet the equilibrium conditions at each joint. Hence, if the joints of the grid are subjected only to forces perpendicular to the plane (x, y), the equilibrium condition of the vertical components uniquely determines a vertical external force at each joint. Inversely: if the grid is loaded by a given set of forces perpendicular to the plane (x, y), and the boundary conditions of the grid ensure a statically determinate stresses state, the equilibrium equation determine the stress function of the stress state, except for three free parameters. These three parameters do not affect the values of the bar forces, because they only represent the equation of an arbitrary plane.

* I. Hegedűs, H-2083 Solymár, Váci Mihály u. 10, Hungary

The stress function of plane grids

Though the problem of pin-jointed single-layer plane grids differs in many respects from that of space grids, the analogy between the nodal equilibrium conditions of bar forces of a plane grid and of projected bar forces of a space grid permits us to use a stress function in solving plane grid problems, too.

As in the previous case, the stress function of the grid has to be the equation of an open polyhedron which has the same projection network as the network of the plane grid. The bar forces are interpreted as the changes in slope of the broken surface lines which have projections perpendicular to the bar axes containing the projections of the kneeing points (Fig. 1). It is shown in [2] that this interpretation is in accordance with

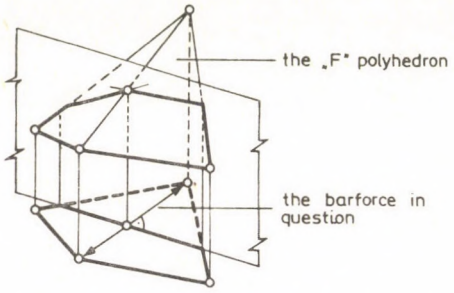


Fig. 1. The interpretation of a bar force

that of stress states determined by Pucher's or Airy's stress functions. The bar forces determined by the polyhedron automatically satisfy the equilibrium condition at each unloaded internal point of the network, so that each stress function represents a statically possible stress state of the grid.

The compatibility of the grid deformations

Assuming that the self–stress state of the grid is zero, the stress state and also the elastic deformations of the bars depend only on the edge loads. The bar forces have to meet the equilibrium conditions at each node, and the elastic deformations have to be compatible.

If a plane grid of triangular network has n internal and m boundary joints, the number of the bars (i.e. of the bar forces) is

$$b = 2m + 3n - 3 ,$$

and the number of static redundants is

$$r = b + 3 - 2(n + m) = n .$$

It follows from the latter equation that the deformations have to meet n independent compatibility conditions. These conditions can be formulated as follows.

The closest neighbourhood of any internal joint of the grid forms an elementary grid fragment as shown in Fig. 2a. As can be seen, this grid is statically indeterminate,

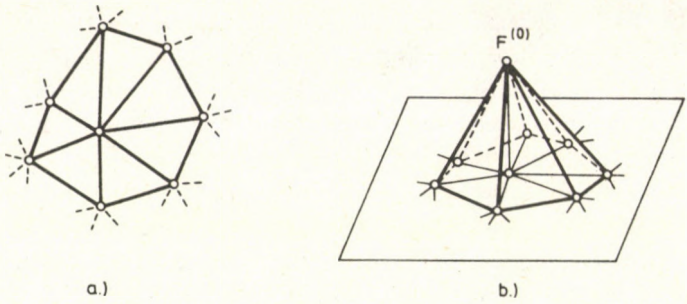


Fig. 2. An elementary grid fragment (a), and the relief of its self-stress function (b)

but the number of redundants is always one, so that, the elastic deformations have to comply with one condition of compatibility. Denoting the bar lengths of the k–gon shaped grid fragment by $l_1, \ldots, l_{2k}$, and the corresponding tensile rigidities by $EA_1, \ldots, EA_{2k}$, the compatibility equation can be written as

$$\sum_{i=1}^{2k} \frac{N_i N_i^0}{EA_i} l_i = 0 . \tag{1}$$

where $N_1, \ldots, N_{2k}$ denote the bar forces caused by the loads of the whole grid and $N_1^0, \ldots, N_{2k}^0$ denote the bar forces of the grid fragment belonging to its self-stress state.

Since Eq. (1) can be independently written for n elementary grid fragments, the set of these equations form the total system of compatibility conditions of the grid.

The boundary conditions

For the actual calculation of bar forces we have to know the nodal values of F at $(n+m)$ points as well as the changes in slope of the faces of the polyhedron of F along m boundary sections, so that the total number of independent data of the stress function has to be at least $(n+2m)$.

If we want to construct the stress function of the grid, subjected to a given boundary load, we have to set up $(n+2m)$ independent equations for determining the required data, or we have to determine $2m$ of them before using the compatibility equations.

In case of externally statically determinate grids of triangular network we can follow the latter way.

The formal degree of freedom of the external loads is $2m$, but the actual one is less by three, because the loads of the grid have to form a system of equilibrium. Consequently, we may assume three arbitrary boundary values as the data of the stress function without losing the possibility of taking into account any system of boundary loads. We may set, for example, the first and the m–th boundary value equal to zero, and let the external continuation of the stress function surface between the first and m-th nodes of the boundary be a horizontal plane coinciding with the plane (x, y). Thus, the other boundary values and slopes of F can be successively determined by considering the external loads as if they were bar forces of fictitious bars. The relief of the external continuation of F has to break over each fictitious bar in such a way that the changes in slope of the surface lines over the lines of the plane (x, y) perpendicular to the fictitious bars have to be equal to the fictitious bar forces in question. The changes in slope are positive or negative values, their signs depending on the signs of stresses in the fictitious bars. If all signs are correctly assumed, then the relief of the extended values of F has to be a connected open polyhedron (see e.g. in Fig. 5), and each edge of the polyhedron has to lie over the line of action of a boundary force. It is not too difficult to realize that the diagram of boundary values of F is similar to the moment diagram of the boundary loads acting on a bar which takes the shape of a broken line, congruent with the boundary of the grid. The free end of this bar geometrically coincides with the fixed one and lies at an arbitrary point of the boundary section between the first and the m–th boundary nodes. Also the slopes of the external faces of the polyhedron can be interpreted as shear forces of the same bar.

The presented method of determining the boundary values of F tacitly involves the assumption that the grid of triangular network covers a simple connected plane domain. If the domain covered by the grid is multiply connected, in other words, if the network contains four or more sided polygons without diagonals, the procedure cannot be used in the simple way as is presented, because the internal boundaries also need boundary conditions, even if there are no internal boundary loads.

The method also fails if some points of application of the external loads are internal points of the network.

Both difficulties are analogous to those of using a stress function in the analysis of elastic discs [1]. The latter can be overcome by using various singular solutions of the differential equation of the problem [1, 4]. The analogies between using stress functions in either problem suggest the idea of constructing "singular solutions" of the grid problem, too. To the author's knowledge the general method of constructing such "singular solutions" has not yet been worked out.

Illustrative example

Let us use the stress function for calculating the bar forces of the grid shown in Fig. 3. Let the length of each bar be equal to l, the tensile rigidity of each bar be equal to EA.

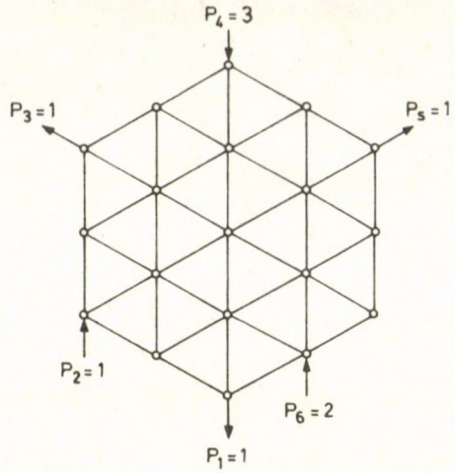


Fig. 3. The network and the loads of the analysed grid.

Since the network of the grid is regular, it is expedient to use an operator symbolism similar to that of the finite difference method. Accordingly, heavy lines or circles refer to the places of reference, and circled numbers show the multipliers of the nodal values of the stress function at their relative place in star–shaped operator diagrams.

The operators of bar forces and of bar elongations (positive if tension) are shown in Figs 4a and 4b.

The elementary grid fragments of the grid are regular hexagons with three pairs of diagonal bars. The relief of their self–stress states is a hexagon based pyramid with zero external continuation.

The operator of the compatibility equation can be constructed by using the operators of bar forces and of bar elongations for calculating the bar forces of the sum in Eq. (1). Its final form is shown in Fig. 4c. A common multiplier containing the actual

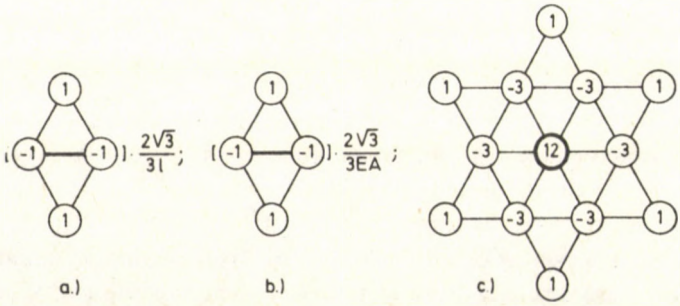


Fig. 4. The operators of bar forces (a), of bar elongations (b), and of the compatibility equations (c)

data of bar lengths and of tensile rigidities also belongs to the operator, but this has been dropped out, because it has no further role in the calculation.

The arbitrarily assumed boundary values are the zeros at the points of application of the forces P_1 and P_6 and the zero slope of the plane between the lines of action of these forces. The relief of the extended values of F is shown in Fig. 5.

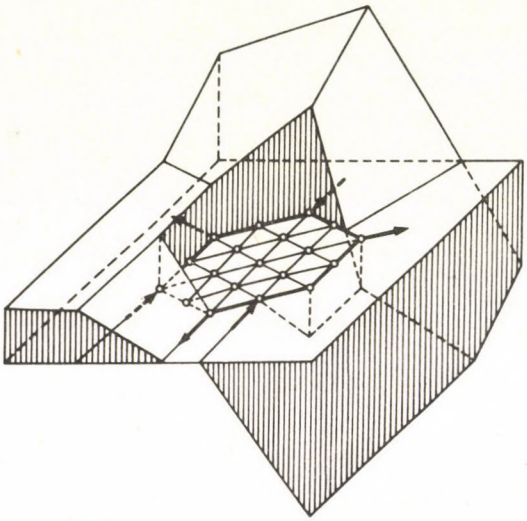


Fig. 5. The relief of the extended stress function

In order to make the algorithmic use of the operator of the compatibility equation possible, the network of the grid has to be completed with fictitious nodes. The values of the stress function at these nodes can easily be calculated from the boundary values and the slopes of the external continuation.

The nodal values of the stress function calculated from the boundary loads and by solving the inhomogeneous system of compatibility equations of seven unknowns, are shown in Fig. 6a. Fig. 6b shows the values of the bar forces as the final result.

Closing remarks

Though our example refers to a grid of regular network and of uniform tensile rigidities, the presented method can be used in cases of irregular network and varying rigidities, too. In these cases the operators vary from node to node, but the procedure of their construction remains the same.

It is worth mentioning that the operator of the compatibility equation shown in the preceding section is exactly the same as the difference operator which can be used in solving Airy's differential equation, if a regular triangular network of finite differences is

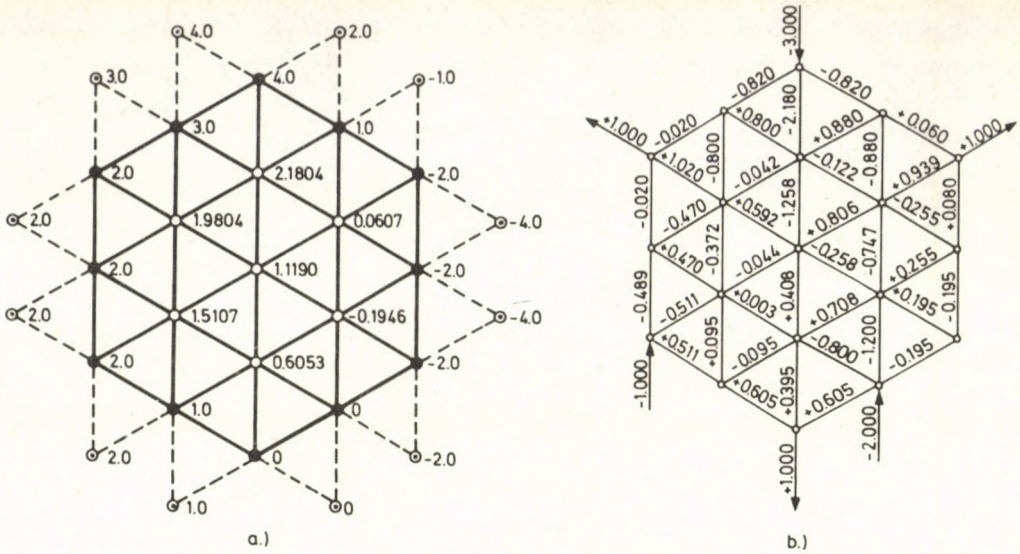


Fig. 6. The nodal values of the stress function (a), and the bar forces of the grid (b)

used. Since the calculation of the boundary values is also the same in both problems, the solution of the grid can be considered as an approximate solution of the analogously shaped and loaded continuous disc.

The widely used analogy between the stress states of isotropic discs and of their replacement by Hrennikoff–type grids (square network with two diagonals, the ratio of cross–sectional areas of the bars forming the squares to those of the diagonals is $\sqrt{2}$; [3]) can also be checked by comparing the operators of the compatibility equations of the grid shown in Figs 7a. and 7b. with the "biharmonic" difference operator of the

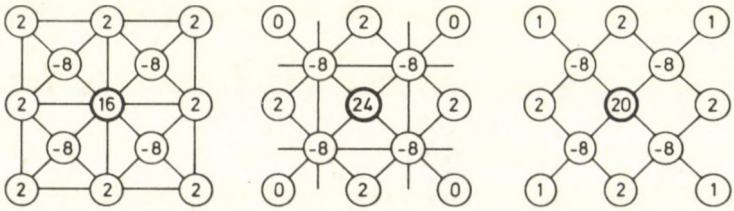


Fig. 7. The operators of compatibility equations of a Hrennikoff-type grid (a and b), and the "biharmonic" difference operator (c)

square network used in solving disc problems shown in Fig. 7c. The comparison shows that the correlation between the computed stress states of the grid and of the disc may be much worse in the case of Hrennikoff–type grids than in the case of regular

triangular grids, if the networks are loose, because the operators of the Hrennikoff grids obviously differ from the "biharmonic" operator of square network, while the operator of the grids of regular triangular network is exactly the same as the "biharmonic" operator of a regular triangular network.

References

1. Girkmann, K.: Flachentragwerke. Dritte Auflage, Springer Verlag, Wien, 1954.
2. Hegedűs, I.: The stress function of single-layer reticulated shells and its relation to that of continuous membrane shells. Acta Technica Acad. Sci. Hung., 97, 103–110, 1984.
3. Hrennikoff, A.: Solution of problems of elasticity by the framework method. Journ. Appl. Mech., 8, 169–175, 1941.
4. Lur'e, A. I.: Teoria uprugosti. Izdatel'stvo "Nauka", Moscow, 1970.

Acta Technica Academiae Scientiarum Hungaricae, 98 (3—4), pp. 317—327 (1985)

THE IDENTIFICATION METHOD OF A DYNAMIC SYSTEM WITH A KNOWN STRUCTURE

V. A. KAMINSKY,* V. I. MAKAROV

[Received: 22 January 1985]

This paper introduces the identification method of linear dynamic systems with a known structure. Usually the identification of the dynamic system is understood as an "inverse" problem in reference to a "primal" problem which can be presented as a solution of a set of differential equations. Thus the "inverse" problem deals with determining the system's unknown parameters when its solution and characteristics are given. The case of vibrating systems is considered in this paper. The method is illustrated by examples given at the end of the paper.

1. Formulation of the problem and its notation

Consider the linear dynamic system

$$M\ddot{x}+R\dot{x}+Cx=F\,, \tag{1}$$

where

$$M=[m_{ij}]\,,\quad R=[r_{ij}]\,,\quad C=[c_{ij}]\quad (i,j=1,\ldots,n)$$

are square matrices with constant coefficients of the order n and matrix M is a diagonal one;

$$x=X(t)=(X_1(t),\ldots,X_i(t),\ldots,X_n(t))$$

and

$$F=F(t)=(F_1(t),\ldots,F_i(t),\ldots,F_n(t))$$

are vector-functions. In future, for reasons of simplicity, the diagonal elements of matrix M are written in the form $m_{ii}=m_i$. In the "primal" problem, the elements of matrices M, R, C, and the right part of system (1) $F(t)$ are known. The task is to find the solution of system $X(t)$ for fixed initial conditions $X(t_0)=X_0$. Another possible approach to the analysis of the "primal" problem deals with the study of the spectral properties of the dynamic system. In this case, again, matrices M, R, C are known and vector-function $F(t)$ is a broadband random process with given characteristics. The problem is to determine the spectral properties of solution $X(t)$. Actually, the last task is to find the set of transfer functions for system (1)—$G(s)=(G_1(s),\ldots,G_i(s),\ldots,G_n(s))$

* V. A. Kaminsky, 113114 Moskow, Shluzovaia Naberezknaia 8, USSR 4

where $G_i(s)$—is a transfer function corresponding to the generalized coordinate $X_i(t)$ of the vector-function $X(t)$. Then the corresponding amplitudo-frequence response

$$A_i(\ldots, m_i, \ldots, r_{ij}, \ldots, c_{ij} \ldots, \omega) \quad (i=1, \ldots, n)$$

and the phase-frequence response

$$\Phi_i(\ldots, m_i \ldots, r_{ij}, \ldots, c_{ij} \ldots, \omega) \quad (i=1, \ldots, n)$$

are to be plotted. These characteristics manifest the properties of the system.

The identification problem of the dynamic system is "inverse" in relation to the second modification of the "primal" problem. The vector-functions of the amplitudo-frequence response

$$a(\omega)=(a_1(\omega), \ldots, a_i(\omega), \ldots, a_n(\omega))$$

and of the phase-frequence response

$$\varphi(\omega)=(\varphi_1(\omega), \ldots, \varphi_i(\omega), \ldots \varphi_n(\omega))$$

are given. However, the elements of matrices M, R, C are unknown and should be determined. If the problem is "inverse", then instead of the amplitudo-f=requence response $a(\omega)$ and the phase-frequence response $\varphi(\omega)$ a class of solutions of system (1) may be given, which were obtained at the broadband spectrum of excitation $F(t)$. Therefore the "inverse" problem, as a rule, leads to the solution of the overdetermined set of equations

$$\begin{gathered} A_i(\xi, \omega_k)=a_i(\omega_k), \qquad \Phi_i(\xi, \omega_k)=\varphi_i(\omega_k) \\ i=1, \ldots, n; \qquad k=1, \ldots, k \end{gathered} \tag{2}$$

where: 1) $A_i(\xi, \omega_k)$, $\Phi_i(\xi, \omega_k)$ $(i=1, \ldots, n)$ are expressions obtained from the given structure of the system (1) and contained unknown parameters m_i, r_{ij}, c_{ij} $(i, j=1, \ldots, n)$ written as N-dimensional $(N=2n^2+n)$ vector

$$\xi=(\xi_1, \ldots, \xi_l, \ldots, \xi_N) \qquad \text{with} \quad \xi_l=m_l (l=1, \ldots, n); \quad \xi_l=r_{ij}$$

$$(l=n+1, \ldots, n^2+n; \quad i, j=1, \ldots, n); \quad \xi_l=c_{ij}$$

$$(l=n+n^2+1, \ldots, n+2n^2; \quad i, j=1, \ldots, n).$$

2) $a_i(\omega)$, $\varphi_i(\omega)$ $(i=1, \ldots, n)$ are measured values of amplitudo-frequence response and phase-frequence response of system (1) with fixed argument ω ($a_i(\omega_k)$ and $\varphi_i(\omega_k)$ $(k=1, \ldots, K)$ given for problem (2)).

System (2) contains N unknown values and consists of $2Kn$ equations where K is usually taken to be $2Kn \geq n^2+n$. The set of equations for system (2) always has a solution.

Remark 1

Due to the form of amplitudo-frequence response and phase-frequence response, it is obvious that functions

$$A_i(\xi, \omega_k), \quad \Phi_i(\xi, \omega_k) \quad (i=1, \ldots, n;\ k=1, \ldots, K)$$

are twice differentiable in almost all points of the parameter space except at points $\{y\}$ where these functions are discontinuous and unlimited, i.e.

$$\lim_{y_p \to y} A_i(y_p, \omega_k) = \pm\infty$$

$$\left(\text{or } \lim_{y_p \to y} \Phi_i(y_p, \omega_k) = \pm\infty\right).$$

Instead of problem (2), an analogous but stochastic one may be considered

$$A_i(\xi, \omega_k) = \tilde{a}_i(\omega_k), \qquad \Phi_i(\xi, \omega_k) = \tilde{\varphi}_i(\omega_k)$$
$$i=1, \ldots, n; \quad k=1, \ldots, K \tag{3}$$

where $\tilde{a}_i(\omega_k)$ and $\tilde{\varphi}_i(\omega_k)$ $(i=1, \ldots, n;\ k=1, \ldots, K)$, are random values for which

$$M[\tilde{a}_i(\omega_k)] = a_i(\omega_k), \qquad M[\tilde{\varphi}_i(\omega_k)] = \varphi_i(\omega_k) \tag{3'}$$
$$D[\tilde{a}_i(\omega_k)] = d'_{ik}, \quad D[\tilde{\varphi}_i(\omega_k)] = d''_{ik} \quad (i=1, \ldots, n;\ k=1, \ldots, K)$$

but d'_{ik} and d''_{ik} are non-negative numbers. By $M[Z]$ and correspondingly by $D[Z]$ we mean the mathematic expectation and variance of random value Z. Such formulation of problem (3) corresponds to the amplitudo-frequence response and the phase–frequence response received with noises having different characteristics depending on argument ω. In this case the overdetermined set of equations (3) cannot always be solved in the classical sense. Therefore, in spite of the solution of problem (2)

$$\xi^* = (\ldots, m_i^*, \ldots, r_{ij}^*, \ldots, c_{ij}^*, \ldots)$$

where the identities

$$A_i(\xi^*, \omega_k) \equiv a_i(\omega_k), \quad \Phi_i(\xi^*, \omega_k) \equiv \varphi_i(\omega_k) \quad (i=1, \ldots, n);\quad k=1, \ldots, K)$$

are valid, in case (3), only a quasi-solution

$$\tilde{\xi}^* = (\ldots, \tilde{m}_i^*, \ldots, \tilde{r}_{ij}^*, \ldots, \tilde{c}_{ij}^*, \ldots)$$

should be sought for which would, to a certain extent, minimize errors

$$\Delta'_{ik}(\xi) = |A_i(\xi, \omega_k) - a_i(\omega_k)|, \qquad \Delta''_{ik}(\xi) = |\Phi_i(\xi, \omega_k) - \varphi_i(\omega_k)|$$
$$(i=1, \ldots, n;\quad k=1, \ldots, K).$$

The geometrical meaning of a quasi solution is the following: in the space of unknown parameters R_N, we are looking for the point $\tilde{\xi}^*$ which has the "minimum distance"

from the set of surfaces

$$\pi'_{ik} \subset R_N : A_i(\xi, \omega_k) - a_i(\omega_k) = 0$$

and

$$\pi''_{ik} \subset R_N : \Phi_i(\xi, \omega_k) - \varphi_i(\omega_k) = 0 \quad (i = 1, \ldots, n; \quad k = 1, \ldots, K).$$

Different expressions may be used as a generalized measure of the simultaneous deviation of the point ξ from all surfaces π'_{ik} and π''_{ik}.

For example:

1) according to the least square solution

$$\psi(\xi) = \sum_{i=1}^{n} \sum_{k=1}^{K} (\Delta'^2_{ik}(\xi) + \Delta''^2_{ik}(\xi)), \tag{4}$$

2) according to Schteiner's point of the set of surfaces

$$\bar{\psi}(\xi) = \sum_{i=1}^{n} \sum_{k=1}^{K} (\Delta'_{ik}(\xi) + \Delta''_{ik}(\xi)), \tag{4'}$$

3) according to Chebyshev's point of the set of surfaces

$$\bar{\bar{\psi}}(\xi) = \max_{\substack{1 \le i \le n \\ 1 \le k \le K}} \{\max(\Delta'_{ik}(\xi), \Delta''_{ik}(\xi))\}. \tag{4''}$$

If there is a broadband spectrum $F(t)$, then problem (2) (or (3)) is to minimize the functional $\psi(\xi)$ in space R_N, that is

$$\psi(\xi) \to \min \tag{5}$$

where $\psi(\xi)$ can be expressed as (4), (4'), (4'').

It is easy to see that ξ^* is the minimizing point of the functional $\psi(\xi)$. Noting $S_F(\omega)$, $S_{X_i}(\omega)$ $(i = 1, \ldots, n)$ the energetic spectra of the right part $F(t)$ and components $X_i(t)$ of the vector solution respectively and using equality

$$S_{X_i}(\omega) = A_i^2(\xi, \omega) S_F(\omega) \quad (i = 1, \ldots, n)$$

([1] p. 437.) holding to be true for linear dynamic systems, we arrive at founding the basis for selecting the functional's (4) kind. This means that the minimization of the functional is equal to the selection of such system's parameters (2) (or (3)) for which the best identification of system (1) will be obtained in the energetic sense.* Identification of system (1) using functionals (4') and (4'') gives the best coincidence in the Chebysev and Schteiner sense of the system's amplitudo-frequence response and phase-frequence response with real characterisitics. In the following we will consider only the functional $\psi(\xi)$, i.e. equation (4), though all reasonings and the suggested method of solution can be used for (4') and (4'').

* Such interpretation of the functional's kind is valid if point $\xi^*(\bar{\xi}^*)$ satisfies equations

$$\Phi_i(\xi^*, \omega_k) = \varphi_i(\omega_k) \quad (\Phi_i(\bar{\xi}^*, \omega_k) = \varphi_i(\omega_k)) \qquad (i = 1, \ldots, n; k = 1, \ldots, K)$$

i.e. the phase-frequence response does not influence the determination of parameters.

Remark 2

Functional

$$\psi(\xi)=\sum_{i=1}^{n}\sum_{k=1}^{K}(\lambda'_{ik}\Delta'^{2}_{ik}(\xi)+\lambda''_{ik}\Delta''^{2}_{ik}(\xi)) \tag{6}$$

may be considered instead of functional (4) where

$$0\leq\lambda'_{ik},\lambda''_{ik}\leq 1,\quad \sum_{i=1}^{n}\sum_{k=1}^{K}(\lambda'_{ik}+\lambda''_{ik})=1 \tag{6'}$$

When solving the problem of identification of a dynamic system, the necessity of marking out any range of frequences in special (or amplitudo-frequence response and phase–frequence response of any components of vector $X(t)$), arises. For this purpose, weights λ'_{ik}, λ''_{ik} corresponding to these frequences and components should be fixed more than for other frequences and components. In the case when $\lambda'_{ik}=\lambda''_{ik}=1/2Kn$ $(i=1,\ldots,n;\ k=1,\ldots,K)$ we get (4) with common multiplier $1/2Kn$, which does not change the point of minimum $\xi^*(\bar{\xi}^*)$.

There are two basic approaches to identification. The first one includes identification of determined systems with filtered noise. Paper [3] can be applied to this method. In this paper, identification of the dynamic system is accomplished by using the integral surfaces of system (1), using paper [4]. The identification of linear systems according to amplitudo–frequence response and phase–frequence response is provided in [5]. Amplitudo-frequence response and phase-frequence response given in units 20 lg $A(\omega)$ and $\Phi(\omega)$ in lg ω may be presented as a spline which is not higher than the second order and the corresponding transfer function $G(S)$ is a product of some transfer functions of elements of the first and second order. The accuracy of such identification is not high as the approximate characteristics of links are used and, besides, phase–frequence response is necessary for an unambiguous identification. In [6], the case of identification of a linear system with the unknown matrix R is considered. Partial results of identification of non-linear systems are considered in [7], [8] and [9]. The second approach of the papers deals with the identification of determined systems with noise. One of the methods of identification of such systems is the usage of the Kálmán filtering method [5] and [10].

2. Existence and uniqueness of the solution of the identification problem

All the results of this part of the article refer to the case of sequence $\{\omega_k\}$ $(k=1,2,\ldots K)$ where corresponding surfaces π'_{ik} and π''_{ik} $(i=1,\ldots,n;\ k=1,2,\ldots,K)$ are linearly independent at a point ξ^{**} of the real set of parameters, i.e. of the coefficients of matrices M, R, C (of the vector).

Lemma 1

If the number of the addable sums of the right part of (6) is not less than N, then at any non–negative λ'_{ik} and λ''_{ik} the solution of problem (2)–(6)–(5) ξ^* will coincide with the real set of parameters ξ^{**} of dynamic system (1).

In fact, functional $\psi(\xi)$ of (6) is strictly convex and non–negative, where point ξ^{**} transforms in identities

$$A_i(\xi^{**}, \omega_k) = a_i(\omega_k), \qquad \Phi_i(\xi^{**}, \omega_k) = \varphi_i(\omega_k) \qquad (i = 1, \ldots, n; k = 1, 2, \ldots)$$

from which it follows that $\Delta'_{ik}(\xi^{**}) = \Delta''_{ik}(\xi^{**}) \equiv 0$; therefore point ξ^{**} is a point of global minimum of functional $\psi(\xi)$. So we obtain the identity $\xi^* \equiv \xi^{**}$. Let's note that functional $\psi(\xi)$ has only one extremum due to strict convexity.

Corollary

Instead of minimization of functional $\psi(\xi)$ of (6) we may carry out the minimization of functional

$$\hat{\psi}(\xi) = \psi(\xi) + \sum_{k=K+1}^{L} \sum_{i=1}^{n} (\lambda'_{iK} \Delta'^2_{ik}(\xi) + \lambda''_{ik} \Delta''^2_{ik}(\xi)) \tag{7}$$

with

$$\lambda'_{ik} = \lambda''_{ik} \sim \frac{1}{E} \qquad (i = 1, \ldots, n; \quad k = K+1, \ldots, L) \tag{8}$$

where E is a sufficiently large number.

Such modification of the functional will lead to the improvement of convergence of the minimization process.

According to Remark 1 and to the kind of functionals (4), (6) and (7), it is easy to see that they are twice differentiable in the vicinity of the points suspicious for the minimum of functionals (4), (6) and (7). The next lemma for functional (4) may be suggested.

Lemma 2

For any arbitrarily small $\varepsilon > 0$ there is a sufficiently large integer K so that the solution of problems (3)–(6)–(5) satisfies the inequality $\| \tilde{\xi}^* - \xi^{**} \|_{R_N} \leq \varepsilon$.

In fact, due to the smoothness of functions

$$A_i(\xi, \omega_k) \qquad \text{and} \qquad \Phi_i(\xi, \omega_k) \qquad (i = 1, \ldots, n; \quad k = 1, \ldots, K)$$

conditions (3′) and to the independence of random values $\tilde{a}_i(\omega_k)$ and $\tilde{\varphi}_i(\omega_k)$ when i and k are different, we may assume the distribution of hypersurfaces π'_{ik}, $\pi''_{ik} \subset R_N$ to be symmetrical. Therefore, for the minimum point $\tilde{\xi}^*$ of functional (6), according to the law of large numbers, $\tilde{\xi}^* \to \xi^{**}$ takes place if $K \to \infty$ which was to be proved.

Functional $\hat{\psi}(\xi)$ may be considered as perturbation of functional $\psi(\xi)$ of (6) by small addable sums the value of which depends on the selection of E. Thus functional $\hat{\psi}(\xi)$ may be represented by $\hat{\psi}(\xi)=\psi(\xi)+(1/E)Z(\xi)$ so according to theorem 6 (perturbation of optimum [11] p. 52) we obtain the vicinity of solutions $\tilde{\xi}^*$ of problems (3)–(7)–(5) and $\tilde{\tilde{\xi}}^*$ of problems (3)–(6)–(5). So the following theorem may be formulated on the basis of the functional properties (4), (6) and (7), lemmas 1, 2 and our reasonings.

Theorem

For any arbitrarily small $\varepsilon>0$ there is a sufficiently large integer L in a way that the solution $\tilde{\tilde{\xi}}^*$ of problems (3)–(7)–(5) satisfies the inequality $\| \tilde{\tilde{\xi}}^* - \xi^{**} \|_{R_N} \leq \varepsilon$ on condition of (8).

Remark 3

When solving problems (3)–(7)–(5) under systematic but independent noises in dimensions $\tilde{a}_i(\omega_k)$ and $\tilde{\varphi}_i(\omega_k)$ $(i=1, \ldots, n;\ k=1, \ldots, K)$ addition of additional summands to functional $\psi(\xi)$ improves the accuracy of solving the problem and makes the process of determining point $\tilde{\tilde{\xi}}^*$ more stable.

Remark 4

Functional (7) in the identification problem (3)–(7)–(5) may be substituted by functional

$$\psi(\xi)=\sum_{i=1}^{n_1}\sum_{k=K+1}^{L}(\lambda'_{ik}\Delta'^2_{ik}(\xi)+\lambda''_{ik}\Delta''^2_{ik}(\xi))$$

where $n_1<n$ and corresponds to the process of identification if there is no information about some generalized coordinates $X_{i_0}(t) i_0 \in (1, 2, \ldots, n)$. In this case, lack of information is partially compensated for by the increase of K, i.e. fuller utilization of information on amplitudo-frequence response and phase-frequence response.

Remark 5

The results described above are also true for the case when it is necessary to find only part of the parameters of system (1) (i.e. in the case, when some elements of matrices M, R, C are known.)

Remark 6

In some cases the following constraints

$$w_\gamma(\xi)\leq 0 \qquad \gamma=1, \ldots, \Gamma_1 \tag{9}$$

$$w_\gamma(\xi)=0 \qquad \gamma=\Gamma_1+1, \ldots, \Gamma_2$$

can be given a priori for the vector of the parameters that should be found. Then we have a limited problem of minimization of (3)–(4)–(9)–(5) (or (3)–(6)–(6′)–(9)–(5)).

3. The method of solving the identification problem and examples

Problem (3)–(4)–(9)–(5) (or (3)–(6)–(6′)–(9)–(5)) which is a limited-constrained problem of non-linear programming is suggested to be solved by the flexible tolerance method [12]. Expressions for amplitudo-frequence response and for phase-frequence response of system (1) can be obtained in the following way. Transform system (1) into a system of the first order $\dot{q}=Q(\xi)q+B$, where $(2n\times 2n)$—matrix $Q(\xi)$ and vector B of dimension $2n$ will be

$$Q(\xi)=\begin{bmatrix} 0 & 1 & 0 & & \ldots & & 0 \\ -\frac{c_{11}}{m_1} & -\frac{r_{11}}{m_1} & -\frac{c_{12}}{m_1} & -\frac{r_{12}}{m_1} & \ldots & -\frac{c_{1n}}{m_1} & -\frac{r_{1n}}{m_1} \\ 0 & 0 & 0 & 1 & 0 \ldots & & 0 \\ -\frac{c_{21}}{m_2} & -\frac{r_{21}}{m_2} & -\frac{c_{22}}{m_2} & -\frac{r_{22}}{m_2} & & -\frac{c_{2n}}{m_2} & -\frac{r_{2n}}{m_2} \\ \ldots & & & & & & \\ 0 & 0 & 0 & 0 & \ldots & 0 & 1 \\ -\frac{c_{n1}}{m_n} & -\frac{r_{n1}}{m_n} & -\frac{c_{n2}}{m_n} & -\frac{r_{n2}}{m_n} & & -\frac{c_{nn}}{m_n} & -\frac{r_{nn}}{m_n} \end{bmatrix}$$

$$B=(b_1, \ldots, b_{2n}) \begin{cases} b_i=0 & i\neq 2k-1 \\ b_i=F_k/m_k & i=2k \quad k=1, \ldots, n. \end{cases}$$

Let's introduce vector

$$C^i=(C_1^i, C_2^i, \ldots, C_{2n}^i): \begin{cases} C_j^i=0 & j\neq 2i-1 \\ C_j^i=1 & j=2i-1 \end{cases}$$

and assume $|[Q(\xi)-sI]| \neq 0$ for all ξ of sufficiently small vicinity of point ξ^{**} where I is the identity matrix. Then i-component of the vectors $G(s)$ is expressed through matrix $Q(\xi)$ and vectors B and C^i:

$$G_i(s)=C^i[-Q(\xi)+sI]^{-1}B$$

and

$$A_i(\xi, \omega_k)=|C^i[-Q(\xi)+j\omega_k I]^{-1}B|,$$

$$\Phi_i = (\xi, \omega_k) = \operatorname{Arg}\left[C^i[-Q(\xi)+j\omega_k I]^{-1}B\right]$$

$$i = 1, \ldots, n; \quad k = 1, \ldots, K .$$

Thus the problem of non-linear programming should be $\psi(\xi) \to \min$ with constraints (9).

Below, we are going to state the identification results of linear dynamic systems of the 1-st and 2nd order. The identification was performed in the case of one amplitudo–frequence response and $N = 2$.

Example 1. A system of the first order $\ddot{x} + r\dot{x} + cx = f$. The transfer function of the system will be $G(s) = 1/(s^2 + rs + c)$ and the amplitudo-frequence response will be $A(\omega) = 1/\sqrt{(c-\omega^2)^2 + r^2\omega^2}$.

In Table 1, values of the amplitudo-frequence response for $\xi^{**} = (m, r, c)$ are given, where $m = 1$, $r = 0.978$, $c = 264.87$.

Table 1

ω	5	14	15	16	17	18	25
$a(\omega)$	4.168×10^{-3}	1.48×10^{-2}	2.368×10^{-2}	5.55×10^{-2}	3.404×10^{-2}	1.168×10^{-2}	2.77×10^{-3}

In Fig. 1. c is plotted as a function of r for different ω.

Example 2. There is a system of the second order $M\ddot{x} + R\dot{x} + Cx = F$ where

$$M = \begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix} \qquad R = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} \qquad C = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}.$$

External excitation F is applied to m_2, transfer function for X_2 is

$$(m_1 s^2 + r_{11} s + c_{11}) / [(m_1 s^2 + r_{11} s + c_{11})(m_2 s^2 + r_{22} s + c_{22}) -$$

$$-(r_{12} s + c_{12})(r_{21} s + c_{21})] ,$$

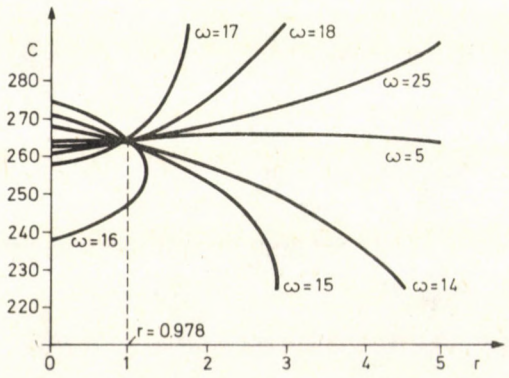


Fig. 1. System of the first order

the corresponding amplitudo-frequence response is

$$A(\omega)=\sqrt{d_1^2+d_2^2}/\sqrt{g_1^2+g_2^2},\ d_1=c_{11}-m_1\omega^2;\ d_2=r_{11}\omega;$$

$$g_1=m_1m_2\omega^4-(m_1c_{22}+m_2c_{11}+r_{11}r_{22}-r_{21}r_{12})\omega^2+c_{11}c_{22}-c_{21}c_{12};$$

$$g_2=(c_{11}r_{22}+c_{22}r_{11}-r_{12}c_{21}-r_{21}c_{12})\omega-(m_1r_{22}+m_2r_{11})\omega^3.$$

In Table 2, values of amplitudo-frequence response are given for

$$\xi^{**}=(m_1, m_2, r_{11}, r_{12}, r_{21}, r_{22}, c_{11}, c_{12}, c_{21}, c_{22}),$$

where

$$M=\begin{bmatrix}4 & 0\\ 0 & 0.4\end{bmatrix},\qquad R=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix},\qquad C=\begin{bmatrix}2000 & -1000\\ -1000 & 1000\end{bmatrix}.$$

Table 2

ω	10	15	19	30	50
$a(\omega)$	2.985×10^{-3}	1.1	1.061×10^{-2}	7.91×10^{-4}	8×10^{-3}

Matrix C is given in the form of

$$C=\begin{bmatrix}c_1+c_2 & -c_2\\ -c_2 & c_2\end{bmatrix}.$$

In Fig. 2, c_1 is plotted as a function c_2 for different ω.

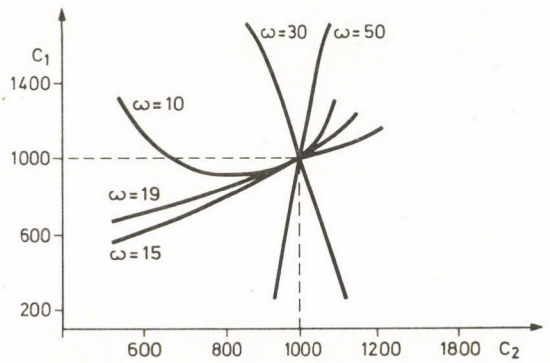


Fig. 2. System of the second order

References

1. Pugachev V. S.; Teorija sluchajnih funkcij, 1962
2. Remez E. Ja.: Osnovi chislennich metodov chebishevskovo priblizhenia, 1969
3. Gluharev K. K., Rozenberg D. E.: Metod sinteza uravnenij dvizhenia mechanicheskih sistem s izvestnim chislom stepenij svobodi. Maschinovedenie, 1973, No. **6,** 11-17
4. Erugin, N. P.: Postroenie vsevo mnozhestva sistem differencialnih uravnenij, imejushchih zadannuju integralnuju krivuju. Prikl. matem. i mechan. 1952, t. XVI., vip. 6, 659–670
5. D. Graupe: Identification of Systems. New York, 1976
6. P. Caravani, W. T. Thomson: Identification of damping coefficients from system response. 5-th World Conference on earthquake engineering, 1973
7. Senik, P. M., Sokil B. J.: Ob opredelnii parametrov nelinejnoj kolebatelnoj sistemi po amplitudo-chastochnój harakterisztike. Matem. Metodi i fiz. meh. polja 1977, No. **6,** 94–99
8. Griba V., J. Hergott, D. Laciak: Identifikacia nelinearnej mechanickej sustavy s jednym stupnom volnosti. Strejnicky casopis 1978, No. **5,** 521–528
9. N. Distefano, A. Rath: Sequential identification of hysteretic and viscous models in structural seismic dynamics. Computer Methods in Applied Mechanics and Engineering. 1975, 6, No. **2,** 219–232
10. Kraskovskij A. A.: Identifikacia i ocenivanie linejnih sistem pri nabljudenii proizvodnih. Techn. kibern. 1978, No. **5,** 150–157
11. A. V. Fiacco, G. P. McCormick: Nonlinearprogramming sequential unconstrained minimization techniques, 1968
12. D. M. Himmelblau: Applied nonlinear programming, 1972

Acta Technica Academiae Scientiarum Hungaricae, 98 (3—4), pp. 329—344 (1985)

SYMMETRICALLY EXCITED ARCHIMEDEAN TWO-WIRE SPIRAL ANTENNA

J. KAPOR*

[Received: 19 December 1983]

A symmetrically excited archimedean spiral antenna is discussed. The literature still owes us a comprehensive theoretical analysis of the spiral antenna and thus the design of a spiral antenna is difficult. In this paper, Burdine's Band Theory was further developed and the antenna geometrically analyzed to give simple and accurate design equations. The spiral windings are represented by radiating lines which, at any arbitrary radial plane, can be represented by point sources, their resultant array depending on the method of excitation: end-fire array, or broadside array. From the point source model, the typical antenna parameters can be determined.

1. Introduction

Since the mid fifties, E. M. Turner's archimedean spiral antenna [1] has been widely used in the 500–1200 MHz frequency range (in the aircraft industry, space communication, space transmission, in actinometers etc.). Despite of its wide use, there is no analytically elaborated exact theory to describe the operation of the antenna. It has been empirically developed, and most texts of the theory of mechanism of antenna operation are based on qualitative considerations [2–7]. In practice, among the different speculative analyses regarding antenna operation, the "Band Theory" of B. H. Burdine provides acceptable results [8]. B. H. Burdine considered the antenna as a simple band wound in the form of a spiral, and explained its radiation mechanism on the basis of the characteristic current distribution associated with the geometry of a spiral. The drawback of the "Band Theory" is that it gives neither exact design equations nor the radiation characteristics in a mathematically precise form, it merely gives a general outline.

W. L. Curtis demonstrated that the radiation properties of the spiral antennas could also be analytically calculated from the known current wave travelling in the arms of the spiral by writing the vector potential function [9]. This method, however, requires the solution of difficult and complicated integral equations. In this paper we present a more demonstrative alternative solution to determine radiation characteristics obtained by further development of the "Band Theory" and by substituting the spiral antenna with a point radiating system. We also give a geometrical analysis of the active range of the antenna, what can be considered very useful and generally applicable in practice. This work also gives the basic design equations of the spiral antenna.

* J. Kapor, H-1133 Budapest, Váci u. 88/a, Hungary

2. Antenna operation, geometrical analysis of active range

Operation

From the "Band Theory" it is known that intensive radiation from the antenna aperture as a result of the excitation at some frequency "f" occurs only from an annular surface from the windings in the interior of this surface. As the frequency is changed the active annular range decreases or increases proportionally to the actual wave length, which in such a way, fluctuates between the internal and external windings. At the upper operating frequency the internal windings radiate, while the external windings radiate at the lower operating frequency. The average diameter (D_a) and the direction of radiation of the active range depends on the polarity of the input supply.

In case of anti-phase excitation maximal radiation intensity occurs axially on both sides of the antenna (axial mode, $D_a=(2k-1)\cdot\frac{\lambda}{\pi}$). If the input excitation of the antenna is in-phase, the angle of the main direction of radiation is approximately 40° to the spiral plane, while absolute minimum of radiation intensity lies on the axis of the antenna (normal mode $D_a=2k\cdot\frac{\lambda}{\pi}$). Let us not deal with the formation of the higher harmonics ($k=2, 3, 4, \ldots$) but rather analyse the case of the fundamental harmonic ($k=1$).

Analysis

Geometrical analysis is made for the self-complementary antenna structure [10], which shows the most favourable wide-band properties, without loss in generality. In case of a self-complementary antenna, the conducting band (W_c) and the width of the insulated space between the conducting arms (W_i) are equal ($W_c=W_i=W$). The relationships written or proven during the analysis remain valid for any complementary two-armed structure spiral antenna by using the average width:

$$W=\frac{W_i+W_c}{2}.$$

The case of non-complementary antenna is not dealt with here. We assume that current waves travel in the spiral arms, and ignore the waves reflected at the ends. The propagation constant of the current waves is used to approximate the propagation constant of plane waves in free space [9].

The plane vector equation of the centre lines of the conducting arms of the symmetrically supplied two-armed archimedean spiral antenna shown in Figure 1:

$$\begin{aligned} r_1&=a\varphi+r_0\,,\\ r_2&=a(\varphi-2\pi)+r_0,\quad \varphi\geq 2\pi\,, \end{aligned} \tag{1}$$

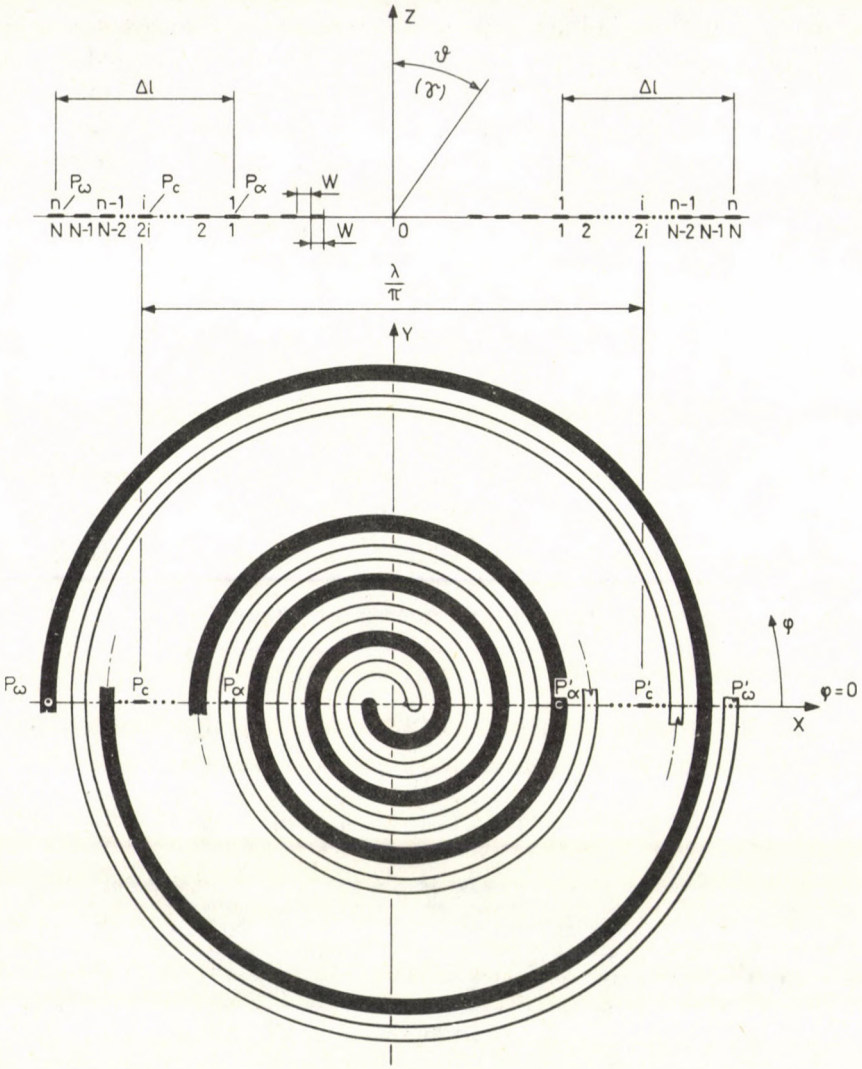


Fig. 1

where: a — *the spiral–constant*
φ — the azimuthal angle in radian
r_0 — the initial radius.

From (1) the pitch is given by

$$m = r[(n+1)2\pi] - r[2n\pi] = 2a\pi \tag{2}$$

On the basis of Fig. 1., and Eq. 2, arm width W can be calculated as follows:

$$W = \frac{a\pi}{2}. \tag{3}$$

Taking the band theory as a starting point and extending it, the active aperture area containing the radiating turns is then defined within the interval enclosed by turns on which the phases of instantaneous current elements (points P_α and P_ω) excised by the radius drawn in any arbitrary direction deviate $\pm\pi/2$ radians from the phase of current element (point P_c) which is excised by circular perimeter λ and which, together with the aforementioned radius, determines the active interval.

The width of the active interval referring to Fig. 1 is then:

$$\Delta l=(n-1)4W\approx 2NW, \tag{4}$$

where: n — number of turns of one of the spiral arms falling within the active interval.
N — number of turns of both spiral curves within the active region.

If we examine point P_c (which is found on a circular perimeter λ within the active area) at the very instant when the Phase of the current at P_c is zero, then within the area defined by points P_α and P_ω, the neighbouring current-elements along the diameter drawn in Fig. 1 will be equidirectional. At points P_α and P_ω, however, the currents start diverging.

According to the definition, the time function of the current at point P_c is given by

$$I_{P_c}(t)=I_0\cos(\omega t-\beta L), \tag{5}$$

where L-distance covered by the current wave from the point of induction to point P_c. Consequently the current at points P_α and P_ω is given as follows:

$$I_{P_{\alpha,\omega}}(t)=I_0\cos(\omega t-\beta L\pm\delta), \tag{6}$$

$$\delta=\frac{N}{2}\frac{\Delta s}{2}\beta=\frac{\pi}{2}, \tag{7}$$

where Δs — length between any two consecutive spiral turns.

A Phase difference π between instantaneous currents at points P_α and P_ω is a result of accumulation of Δs path-lengths. Reasonably, the value of Δs can be approximated as follows:

$$\Delta s=\int_{\varphi_0}^{\varphi_0+2\pi} a\varphi\,d\varphi-\int_{\varphi_0-2\pi}^{\varphi_0} a\varphi\cdot d\varphi=a(2\pi)^2. \tag{8}$$

Taking into consideration the number of turns n, the current phase difference between P_α and P_ω can be deduced on the basis of Eq. (7):

$$2Na\,\pi^2\cdot\beta=\pi. \tag{9}$$

Substituting Eq. (9) into Eq. (3) we obtain

$$4NW\beta=1. \tag{10}$$

Combining Eqs (4) and (10) the width of the active area becomes:

$$\Delta l = 2NW = \frac{1}{2\beta} = \frac{\lambda}{4\pi}. \tag{11}$$

The result obtained by Eq. (11) is, in fact, the fundamental relation which is necessary for designing the spiral antenna. The equation reveals that the width of the active area depends only on the wave length. In order to ensure that mainly travelling waves will develop along the spiral arms, a number-of-turns condition of $n > 3$ ($N > 6$) shall exist on the basis of practical experience, similarly to the multiturn helical antenna. In practice, the value of n for most cases falls between 3 and 8. From Eq. (11) and the minimum number of turns stipulation, it is then possible to determine the maximum value of arm width W enabling suitable antenna performance. When determining a still applicable maximum arm width, we, of course, make use of wave length λ_u associated with the designed upper limit frequency. Accordingly, the fundamental design equation takes the following shape:

$$W \leq \frac{\lambda_u}{16(n-1)\pi} \approx \frac{\lambda_u}{8N\pi} = \frac{1}{4N\beta}. \tag{12}$$

Condition for the number of turns: $N_{\min} = 6$. Substituting this into the above equation, the upper value of the arm-width is given as follows:

$$W_{\max} = \frac{\lambda_u}{48\pi} \cong \frac{\lambda_u}{150}. \tag{13}$$

Taking into consideration the path-length-increasing effect of the base plate supporting the antenna:

$$W_{\max} = \frac{\lambda_u}{150\sqrt{\varepsilon_{r_{\text{eff}}}}}. \tag{14}$$

where $\varepsilon_{r_{\text{eff}}}$ represents the effective relative dielectric constant which, for most substrates, is nearly unity.

Note that as arm-width W decreases, the value of damping factor α increases, which affects the transmission characteristics of the antenna unfavourably. In the knowledge of the quality characteristics of the supporting plate and in case of given gain requirements, it is also possible to give a minimum arm-width W.

Results of the analysis

In the course of geometrical analysis we defined specific limits for the active area containing the radiating turns. At the excitation frequency, this active area is separated from all the other turns operating as a transmission line by points P_α and P_ω. Of course,

this transition is indistinct in practice. The phase shift of currents at points P_α and P_ω is $\pm\pi/2$ radians with respect to the current at point P_c. On the basis of Eq. (9), the phase differences between the current-elements of neighbouring turns are equally π/N.

With the current-elements of width W shown in Fig. 1 considered to be points, Fig. 2 has then been plotted to show the phase and amplitude distribution of the

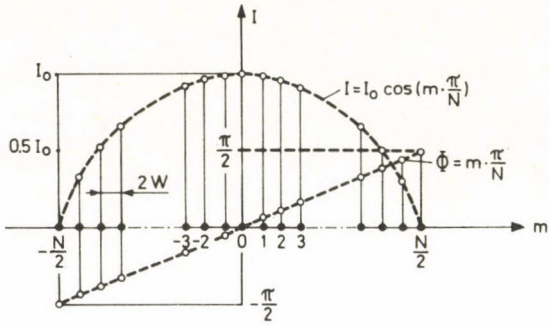


Fig. 2

momentary current associated with the point sources within the active region. For the sake of simplicity, we recorded the time when the value of current at the centre (P_c) of the active region was I_0 and the phase angle zero. At this instant the phase response in the active region is given by the following function:

$$\Phi = m \cdot \frac{\pi}{N}; \quad m = -\frac{N}{2} \ldots 0 \ldots + \frac{N}{2}, \tag{15}$$

where N must be an even number since the antenna has two arms. At the same time, it has to be stressed that, owing to the nature of the spiral curve, the turns falling within the active region are not concentric but open circles and therefore there is a phase uncertainty of $\pm\pi/2N$ within the examined interval.

The amplitude distribution function:

$$I = I_0 \cos\left(m \frac{\pi}{N}\right), \quad m = -\frac{N}{2} \ldots 0 \ldots + \frac{N}{2}. \tag{16}$$

The current distribution along the examined diameter ($\varphi = 0$ plane) changes in time in accordance with a $\cos \omega t$ function according to the excitation frequency:

$$I(t) = I_0 \cos\left(m \frac{\pi}{N}\right) \cdot \cos \omega t\,. \tag{17}$$

while the instantaneous current distribution shown in Fig. 2 turns off or spins in the direction of winding of the spiral according to the inducing angular frequency.

Antenna pattern

By analogy with the helical antenna the operation of which is based on travelling waves, we consider the turns falling within the active region to be transmission lines, where the combined effect of the turns corresponds to 'end-fire or 'broadside' arrays, depending on feed. In determining the radiation pattern, line width W is reduced to be infinitely thin and the propagation constant of the wave travelling along the line is approximated by the propagation constant of the planar wave travelling in free space.

Radiation pattern of the axial-mode antenna

The radiation pattern was determined according to Fig. 3. In the following calculations, the instant shown in Fig. 3 when the current at the centre of the active region has just reached its maximum was considered. At this time, the amplitude and phase distribution of the current associated with the point sources are identical to those

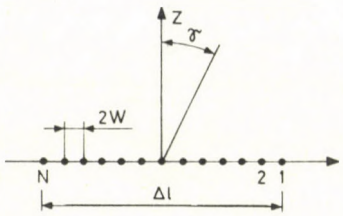


Fig. 3

given in Fig. 2. The currents of the point sources excising the $\varphi=0$ plane are approximated by the average value of the current amplitudes, which in practice gives a reasonable approximation for design purposes. On the basis of Fig. 4.:

$$I_a=\frac{I_0}{\pi}\int_{\alpha=-\frac{\pi}{2}}^{\frac{\pi}{2}}\cos\alpha\,d\alpha=\frac{2}{\pi}I_0\,. \tag{18}$$

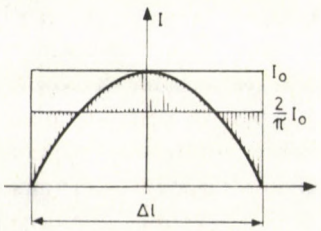


Fig. 4

The phase difference between the currents of the point sources is equally π/N. The directional factor [11] of the linear array of N isotropic point sources of equal amplitude and spacing is given by

$$f(\gamma)=\frac{\sin N\cdot\psi/2}{\sin\psi/2}, \tag{19}$$

where

$$\psi=2W\beta\sin\gamma+\frac{\pi}{N}=\frac{1}{2N}\sin\gamma+\frac{\pi}{N}. \tag{20}$$

Substituting Eq. (20) into Eq. (19), we obtain

$$f(\gamma)=\frac{\sin\left(\frac{1}{4}\sin\gamma+\pi/2\right)}{\sin\left(\frac{1}{4N}\sin\gamma+\frac{\pi}{2N}\right)}=\frac{\cos\left(\frac{1}{4}\sin\gamma\right)}{\sin\left(\frac{1}{4N}\sin\gamma+\frac{\pi}{2N}\right)}. \tag{21}$$

The directional factor has a maximum at $\gamma=0$ and $\gamma=\pi$:

$$f_{\max}=\frac{1}{\sin\frac{\pi}{2N}}\approx\frac{2N}{\pi},\quad N\geq 6.$$

Consequently, the relative directional factor is given by

$$f_r(\gamma)=\frac{\cos\left(\frac{1}{4}\sin\gamma\right)}{\frac{2N}{\pi}\sin\left(\frac{1}{4N}\sin\gamma+\frac{\pi}{2N}\right)}. \tag{22}$$

Taking into consideration the fact that $N\geq 6$, trigonometrical transformations result in

$$\sin\left(\frac{1}{4N}\sin\gamma+\frac{\pi}{2N}\right)\approx\frac{\pi}{2N}.$$

The error resulting from the above approximation will be 0% if $\gamma=0$. The error increases monotonously with γ and will reach a maximum of 12% at $\gamma=\pi/2$, if $N=6$. As will be seen later, this error is permissible since the direction corresponding to maximum error complies with the direction of minimum radiation of the antenna, which is of no interest. The error within the cone of main radiation remains between 0 and 9% and reaches a maximum at the edges of the main beam. With this error, both linear arrays arranged symmetrically around the origin in the $\varphi=0$ plane may be replaced by a single radiating point each, with a relative directional factor of

approximately

$$f_r = \cos\left(\frac{1}{4}\sin\gamma\right). \tag{23}$$

Two resultant radiation points separated by distance λ/π from each other have a relative directional factor (f_{re}) which may be determined from the principle of pattern multiplication as follows [12]:

$$f_{re} = f_r \cos\frac{\psi}{2}, \tag{24}$$

where

$$\psi = \frac{\lambda}{\pi}\beta\sin\gamma = 2\sin\gamma. \tag{25}$$

Substituting the value of ψ into Eq. (24) we obtain

$$f_{re} = f_r \cos(\sin\gamma). \tag{26}$$

Since f_r only slightly deviates from the isotropic radiating directional factor, ($f_{r\min} = 0.97$), the effect of the current elements excised by the $\varphi = 0$ plane is equivalent to a point source with a directional factor of approximately

$$f_{re} = \cos(\sin\gamma). \tag{27}$$

Equation (27) is the relative directional factor of an isotropic point source system which gives the directively relation of the field intensity in the $\varphi = 0$ plane (the origin is considered as the phase centre of the point source). Actually, in addition to the current elements in the region where the $\varphi = 0$ plane bisects the active range, the current elements (point sources) of the active annular surface in the $\varphi \neq 0$ plane also contribute to the formation of the antenna radiating field. According to our definition for the active range, for a diameter drawn in any direction $\varphi \neq 0$ we get two arrays of point sources symmetrically to the origin, similarly to that shown in Fig. 3, with a phase difference of π/N between the neighbouring current elements. The average current values calculated in a similar way as in (18) for the point radiating arrays along the radii drawn in different directions differ from each other. The combined effect of radiating lines in accordance with the current distribution at given instant can be characterized by the momentary current of the different point sources substituting the imaginary continuous linear arrays on radiating lines (turns). The equivalent point sources are located on the circular conductor of diameter λ/π shown in Fig. 5 where the instantaneous current distribution can be described by means of the following function:

$$\frac{2I_0}{\pi}\cos\varphi.$$

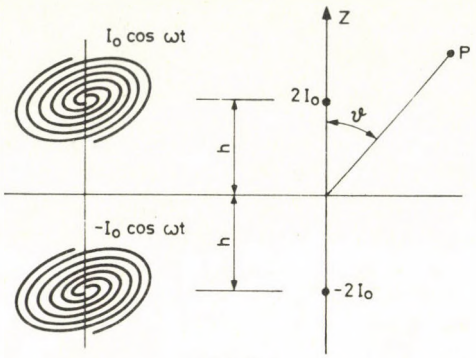


Fig. 5

On the basis of the momentary current distribution shown in Fig. 5, the circular conductor, considering the $\varphi = 0$ plane, can be substituted by a dipole antenna with an effective length of

$$h_y = K \frac{\lambda}{\pi} \cos (\sin \gamma).$$

For an arbitrary time t', depending on the direction of winding of the spiral the position of the dipole antenna replacing the circular conductor is such as to satisfy $\varphi = \omega t'$. In this case, the effective length of the dipole antenna for a left-wound spiral

$$h_x = K \cdot \frac{\lambda}{\pi} \cos (\sin \vartheta) \cos \varphi'$$

$$h_y = K \cdot \frac{\lambda}{\pi} \cos (\sin \vartheta) \sin \varphi'. \tag{28}$$

Accordingly, at any operating frequency, the spiral antenna can be modelled by an imaginary dipole antenna located in its geometrical centre, with although linear 'momentary polarization' but, as a function of time, rotating at the frequency of excitation in a direction depending on the direction of the spiral winding. Accordingly, for a right-hand spiral, the equivalent dipole antenna is right-hand while for a left-hand spiral left-hand circularly polarized. On the basis of the rotating dipole model, the relative directivity factor of the spiral antenna is given by

$$f = \cos (\sin \vartheta). \tag{29}$$

Characteristics of the directivity factor

The direction of minimum radiation of the antenna is the $\vartheta = \pi/2$ plane. $\vartheta = 0$ and $\vartheta = \pi$ are the direction of maximum radiation. There is no zero direction. In the $\vartheta < \pi/2$ and $\vartheta > \pi/2$ cone angle region the directions of rotation of polarization are opposite.

The value of the relative directional factor symmetrical to the $\vartheta = \pi/2$ plane in the direction of minimal radiation is $f(\vartheta) = \pi/2 = 0.54$. From relation $f = 0.707$, the half cone aperture angle for a 3 dB signal level reduction: $\vartheta = 52.5°$.

Antenna directivity

The directivity is determined by means of the complex effective length of the spiral antenna. On the basis of the rotating dipole model, the absolute value of the complex effective length:

$$|\bar{\mathbf{h}}| = K \frac{\lambda}{\pi} \cos(\sin \vartheta), \tag{30}$$

where K—a factor depending on the geometry of the antenna and the method of feed. The directivity, using (30) [13] is given by:

$$D = \frac{|\bar{\mathbf{h}}|^2}{\frac{1}{4\pi} \iint\limits_{4\pi} |\bar{\mathbf{h}}|^2 \, d\omega} = \frac{2 \cos^2(\sin \vartheta)}{\int\limits_{\vartheta=0}^{\pi} \cos^2(\sin \vartheta) \sin \vartheta \, d\vartheta}. \tag{31}$$

Numerically integrating the function of the denumerator:

$$\int\limits_{\vartheta=0}^{\pi} \cos^2(\sin \vartheta) \sin \vartheta \, d\vartheta \cong 0.98.$$

Therefore, the directivity function of the spiral antenna is given by:

$$D(\vartheta, \varphi) = 2.03 \cos^2(\sin \vartheta). \tag{32}$$

In the direction of maximum radiation:

$$D_{max} = 2.03 (3.08 \text{ dBi}).$$

Radiation impedance

Babinet's theory can be used to calculate the radiation impedance of the plane structured antennas. The theory used to determine the input impedance for slot radiators states that: if the radiation (input) impedance of an antenna cut from a thin metal plate is known, the following relationship can be written for a slot radiator of similar profile (complementary)

$$Z_1 \cdot Z_2 = \frac{Z_0^2}{4}, \tag{33}$$

where: Z_1: impedance of antenna cut from plate
Z_2: radiation impedance of complementary slot radiator
Z_0: specific impedance of free space (120π ohm)

In order to ensure optimum wide band properties, the antenna is generally prepared in the self complementary form. In this case $Z_1 = Z_2 = Z$ for which the appropriate radiating impedance is 188.5 ohm. The value obtained in practice differs from the above value which is understandable, since Eq. (33) is for a slot antenna immersed in a wide metal surface, and the slot antenna is assumed to be thin as compared with the wavelength. In fact, apart from the antenna's characteristic geometry, the actual radiation impedance depends on the substrate material and feeding geometry.

Constant K and other characteristics of the antenna

Knowing the radiation impedance, the value of constant "K" used in (30) can be simply determined for ideal self complementary spiral antennas. Using the complex effective length [13], we can write radiation resistance R_r as

$$R_r = \frac{30}{D}\left(\frac{2\pi}{\lambda}\right)^2 K^2 \left(\frac{\lambda}{\pi}\right)^2 = 120\frac{K^2}{D}.$$

Replacing R_r by the impedance (188.5 Ohm) obtained on the basis of Babinet's theory and writing the earlier determined directivity $D = 2.03$ we obtain $K = \sqrt{\pi}$. This leads to the absolute value of the effective length for the self complementary type spiral antennas

$$|\mathbf{h}| = \sqrt{\pi}\frac{\lambda}{\pi}\cos(\sin\vartheta).$$

Knowing $|\mathbf{h}|$, all the important characteristics of the self complementary antenna can be determined [13].

Modification of radiation characteristics with reflector

Placing a reflecting metal surface at a distance d behind the spiral antenna, the radiation of antenna can be directed in one direction. Vector **E** of the wave excited by the antenna undergoes a phase change of 180° during reflection at the reflector, because the tangential component of the electric field intensity is zero in the plane of a perfect conductor. As a result, we have to take into consideration the image effect of an oppositely directed current (Fig. 6) during calculation of the directivity factor. It has been proved that the radiation field of a spiral antenna is equivalent to the field of a dipole rotating at ω angular frequency as a result of which the effect of the reflector can

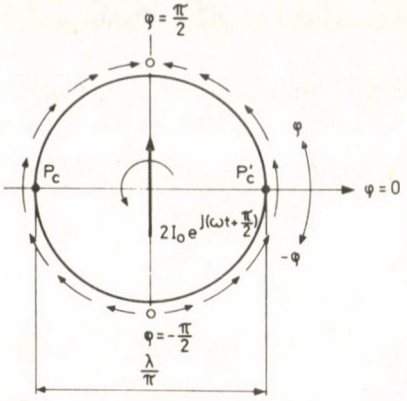


Fig. 6

be examined analogously to the behaviour of a dipole placed horizontally above the ground plane. Using this substitution, the dipole and its image can be considered as point sources in the plane placed perpendicularly to the substituted dipole, and their directivity on the basis of (27)

$$f_r = \cos(\sin\gamma).$$

The resultant relative directivity factor of the two point sources is given by

$$f'_r = f_r \cos \Psi/2, \tag{34}$$

where

$$\Psi = \pi - \beta \cdot d \cdot \cos\gamma. \tag{35}$$

Using relation (35) the relative directivity factor as a function of γ and reflector distance d:

$$f = \cos(\sin\gamma)\cos\left[\frac{\pi}{2}\left(1 - \frac{4d}{\lambda}\cos\gamma\right)\right]. \tag{36}$$

As for any plane indicated by an angle φ, we obtain an expression similar to (36), in place of γ we can write ϑ:

$$f = \cos(\sin\vartheta)\cos\left[\frac{\pi}{2}\left(1 - \frac{4d}{\lambda}\cos\vartheta\right)\right]. \tag{37}$$

According to equation (37) the radiation has its maximum intensity in the main direction of radiation ($\vartheta = 0$) when $d = \lambda/4$. In this case, the relative directivity factor of the spiral antenna supplied with a reflector operating in axial mode is given by

$$f = \frac{|\mathbf{h}|}{|\mathbf{h}_{max}|} = \cos(\sin\vartheta)\cos\left[\frac{\pi}{2}(1 - \cos\vartheta)\right]. \tag{38}$$

This means that the half cone aperture angle $\vartheta = 43°$ corresponds to a 3 dB reduction in signal level.

Theoretically available gain

On the basis of (38) and similarly to (31), the theoretically available gain of the spiral antenna supplied with a reflector operating in axial mode as compared with the isotropic antenna can be determined by the relationship below

$$G_{\max} = D = \frac{\cos^2(\sin\vartheta)\cos^2\left[\frac{\pi}{2}(1-\cos\vartheta)\right]}{\frac{1}{4\pi}\int_{\varphi=0}^{2\pi}\int_{\vartheta=0}^{\frac{\pi}{2}}\cos^2(\sin\vartheta)\cos^2\left[\frac{\pi}{2}(1-\cos\vartheta)\right]\sin\vartheta\,\mathrm{d}\varphi\,\mathrm{d}\vartheta} \tag{39}$$

where the value of the double integral in the denominator using numerical integrating method:

$$2\pi\int_0^{\frac{\pi}{2}}\cos^2(\sin\vartheta)\cos^2\left[\frac{\pi}{2}(1-\cos\vartheta)\right]\sin\vartheta\,\mathrm{d}\vartheta \cong \frac{\pi^2}{5}.$$

Writing the above result in (39), the directivity function:

$$D = \frac{20}{\pi}\cos^2(\sin\vartheta)\cos^2\left[\frac{\pi}{2}(1-\cos\vartheta)\right]. \tag{40}$$

Assuming a lossless antenna and a ground plane of infinite dimension, the theoretically available gain in the main direction:

$$G_{\max} \cong 6.37\ (8\ \mathrm{dBi}).$$

In practice, the gain of a spiral antenna printed on an average quality (average tan) substrate (FR4, G10), as compared with an isotropic antenna (using cylindrical reflecting cavity of approximately $2/5\lambda$ in diameter) is 5–6 dBi at mid-band frequency. According to the measured data published in [14], values of 5.6–6.9 dBi were measured using low-loss substrate in the operating band of the spiral antenna.

Summing up

We confirmed B. H. Burdine's qualitatively based "Band Theory" by a quantitative analysis of the radiation mechanism. Extending the "Band Theory", we accurately defined the range of the radiation channel and, with the antenna considered to be an elementary radiator forming a broadside array, we determined with reasonable accuracy the directional index and the theoretically available gain. On the basis of the point source model used to determine the radiation pattern, we proved that an antiphase-fed symmetrical (two-wire) spiral antenna is equivalent to a dipole which

spins around the axis of the antenna at angular frequency as a result of an $I_0 \cos \omega t$ excitation applied to the spiral input Depending on the direction of rotation of the substituting dipole, the antenna produces right or left circularly polarized field. With the equivalent rotating dipole we can also explain the practical experience that the spiral antenna keeps its property of circular polarization for a comparatively wide cone angle. In the geometrical analysis of the antenna, we proved practically useful and generally applicable basic design equations for the two-wire spiral antennas. We verified that the antenna design is based on the proper choice of only one parameter which is arm width W. The sharpness of the directional angle of the antenna, in accordance with the previous analysis, can be considered to be practically independent of the chosen value of W within the limits determined by the armwidth. We attributed the larger divergence in a given directional sharpness angle experienced only at the lower end of the band to the proximity and end effects of the reflector wall. The antenna gain depends not only on the quality of substrate but also on proper choice of arm width W. We experienced that if the chosen arm width was too thin, the gain decreased as a result of a significant increase in the attenuation coefficient. A typical radiation pattern is shown in Fig. 7. Besides the measured curves, the radiating pattern calculated on the basis of Eq. (38) is also drawn in dash lines in Fig. 7. The difference between the two curves with increasing ϑ is caused by the fact that for the calculated curve we assumed an infinite ground plane at a distance of $d/4$ from the antenna plane while in practice we mounted the antenna on the support of a metal cavity of approximately $25/\lambda_l$ in diameter and $d/4$ in depth (where λ_l is the wave length at the lower operating frequency).

Finally we note that, the relationships derived or given are also valid for a two-wire closely wound ($a \ll 1$) logspiral antenna. This can be easily proven mathematically by the series expansion of the equation of the logspiral curve:

$$e^{a\varphi} = 1 + a\varphi + \frac{a^2\varphi^2}{2} + \frac{a^3\varphi^3}{3!} + \ldots + \frac{a^n\varphi^n}{n!} + \ldots$$

For small values of "a", the following approximation is valid

$$e^{a\varphi} \approx 1 + a\varphi.$$

Consequently, we have proven that the archimedean spiral antenna can be considered as an approximation to closely wound logspiral antenna. Therefore, to a good approximation, our postulated relationships also apply to logspiral antenna.

Acknowledgement

I would like to express my sincere thanks to Dr. Csaba Ferencz and Dr. István Frigyes for the valuable discussions.

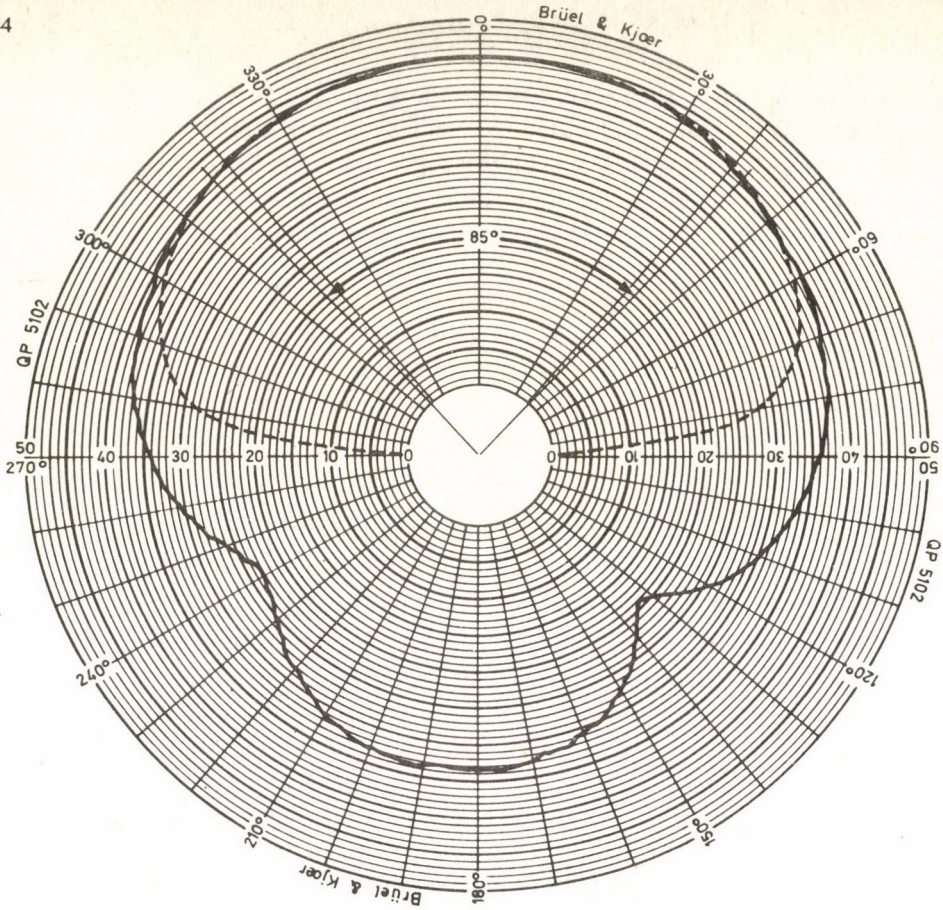


Fig. 7

References

1. E. M. Turner: Spiral Slot Antenna, Wright-Patterson AFB, Ohio, Techn. Note WCLR–55–8. WADC, June 1955
2. R. Bawer, J. Wolfe: The Spiral Antenna, IRE Nat. Conv. Rec., (1960) 84
3. P. Jones, El Taylor, W. Morrow: Design Techniques for a Light Weight High Power Spiral Antenna. IRE Wesc. Conv. Rec. 4 Part 1, (1960), 107
4. A. Kaiser: The Archimedean Two-Wire Spiral Antennas, IRE Trans. on Ant. and Prop. (1960), 32
5. R. Donellan: Second-Mode Operation of the Spiral Antenna. IRE Trans. on Ant. and Prop., (1960), 637
6. J. Kapor: Diplom's work (in Hungarian) Technical University of Budapest 1975
7. J. Kapor: Spiral Antenna. Radiotechnika, 32 (1982). (in Hungarian)
8. H. Burdine, M. McElvery: The Spiral Antenna. Massachusetts Inst. of Techn. Cambridge Res. Lab. of Electronics, Rept. Nos. 1 and 2.
9. L. Curtis: Spiral Antennas. IRE Trans. on Ant. and Prop., (1960), 298
10. H. Rumsay: Frequency Independent Antennas. Academic Press 1966.
11. E. Istvanffy: Waveguides, antennas and propagation. Tankönyvkiadó, Budapest 1979. (in Hungarian)
12. B. Szekeres: Antennas. Tankönyvkiadó, Budapest 1969. (in Hungarian)
13. J. Kapor: Characterization of Elliptically Polarized Antenna by Complex Effective Length. Híradástechnika 34. (1983) (in Hungarian)
14. E. Hörmann, R. Reitzig: Experimental Analysis and Selection of Airborne Antennas for Aircraft-to-Satellite Communication Systems. Frequenz 31 (1977), 11

Acta Technica Academiae Scientiarum Hungaricae, 98 (3—4), pp. 345—366 (1985)

EFFECT OF THE CHANGE OF CROSS SECTIONAL CHARACTERISTICS ON THE FORCE DISTRIBUTION OF VEHICLE FRAMES

KOVÁCS, M.—MICHELBERGER, P.*—NÁNDORI, E.

[Received: November 1984]

Studied here are the properties of the solution of the compatibility equation system determining the force distribution of statically multiply redundant vehicle frames. In doing so, the flexibility matrix containing the stiffness figures of the frame elements are modified in such a way that the cross sectional characteristics in the matrix are treated as variables, taking into consideration the permissible perturbation. Different methods are used to estimate the deviation occurring in the internal force distribution of the frame, and these methods are demonstrated by means of practical examples.

Symbols

$\mathbf{B}$ — matrix of size $(m \times n)$, $m > n$, with maximum rank ie rank $\mathbf{B} = n$
$\mathbf{R}$ — diagonal matrix of size $(m \times n)$, a simple diagonal matrix or diagonal hypermatrix, depending on the degree of load function
$\mathbf{a}$ — vector of elements m
$\mathbf{r}$ — vector of elements m with diagonal elements, $r^{(i)} > 0$, $i = 1, 2, \ldots m$, of $\mathbf{R}$
$\langle c \rangle$ — diagonal matrix with vector $\mathbf{c}$ in its diagonal
$\mathbb{R}$ — set of real numbers
$\mathbb{R}^s$ — real space of dimension s
$\bar{\mathbf{R}}$ — the reference matrix (nominal $\mathbf{R}$) with diagonal $\bar{\mathbf{r}}$: $\bar{r}^{(i)} > 0$
ΔR — $\{r \in \mathbb{R}^m : (1-\lambda)\bar{\mathbf{r}} \leq \mathbf{r} \leq (1+\lambda)\bar{\mathbf{r}},\ 0 \leq \lambda < 1\}$ the possible perturbance range of matrix $\mathbf{R}$
— maximum permissible perturbance parameter, $0 \leq \lambda \leq 1$.

1. Introduction

Like in any steel structure, there may be two types of dimensional inaccuracy in vehicle frames.

Any machine component is produced with tolerance, its size being close to the rated dimension but complying with it only seldom. If a close tolerance is used in production, then minor movements will occur in assembly and also the resulting stress is negligible. In the opposite case when components of a large tolerance zone are produced for e.g. economic reasons, major movements occur in assembly and

* P. Michelberger, H-1111 Budapest, Egry J. u. 19–21, Hungary

consequently, the resulting stresses are not negligible either. Dimensional inaccuracy may be considerable especially in case of large welded, riveted or bolted structural elements like vehicle frames.

Calculation of additional stresses resulting from inaccuracy in assembly and/or dimension in vehicle frames, called kinematic load, is not dealt with here because the methods to determine these stresses are well known from the literature [1].

The other type of dimensional inaccuracy results from the deviation in profile size of rolled steel sections built into the vehicle frames. The dimensions of rolled steel sections together with the applicable manufacturing tolerances are set out in standards. The permissible value of deviation in cross sectional dimensions (e.g. wall thickness, height) of rolled steel sections, the so called profile tolerance, varies in the range of about 0.5 to 2%. For cross sectional characteristics (e.g. equatorial moment of inertia), this value may amount to ± 10 to 12%.

In case of a highly valuable machine of unit production, determination of the dimensions of built–in structural elements and thus accurate calculation is possible in principle. However, in case of series–produced structures, the accurate measurement of any single supporting member and re–calculation for the unique constructions so obtained is not feasible indeed.

In the construction of up–to–date vehicle frames where a considerable gracialization of the frame takes place, it seems to be right to follow with attention the effect of cross sectional characteristics on the internal force distribution of the frame according to what has been said above in such a way that also changes due to dimensional deviation are followed up.

In the dimensioning practice of statically multiply redundant vehicle frames methods based on matrix force method rather than displacement method are preferred. Therefore, it is justified to take the method based on matrix force method as a starting point for investigations also in this study [2].

To simplify treatment, investigations are carried out for internal forces arising in intersections, of which ultimate stresses can be obtained by a simple superimposition.

By writing the flexibility matrix ($\bar{\mathbf{R}}$) produced with the rated cross sectional characteristics into the fundamental equation of the matrix force method, we obtain the so called reference equation

$$\mathbf{B}^T\bar{\mathbf{R}}\,\mathbf{B}\bar{\mathbf{x}} = \mathbf{B}^T\bar{\mathbf{R}}\mathbf{a} \tag{1}$$

with solution $\bar{x}$, and we investigate the relation of the solution of compatibility equation

$$\mathbf{B}^T\mathbf{R}\,\mathbf{B}\mathbf{x} = \mathbf{B}^T\mathbf{R}\mathbf{a} \tag{2}$$

to $\bar{\mathbf{x}}$ if $\mathbf{R}$ is an element of perturbance range ΔR.

Obviously, the relation of the elements of $\mathbf{R}$ to $\bar{\mathbf{R}}$ shall be qualified, and the investigation carried out accordingly.

2. Deterministic approach

In this case the norms $\|\mathbf{R}-\bar{\mathbf{R}}\|$ and $\|\mathbf{x}-\bar{\mathbf{x}}\|$ are used to measure the relationship between $\mathbf{R}$ and $\bar{\mathbf{R}}$ as well as $\mathbf{x}$ and $\bar{\mathbf{x}}$, respectively, where (and in the following everywhere) the vector norm is the Euclidean norm while the matrix norm the natural matrix norm corresponding to it that is the spectral norm [3].

Since

$$\bar{\mathbf{x}} = (\mathbf{B}^T\bar{\mathbf{R}}\mathbf{B})^{-1}\,\mathbf{B}^T\bar{\mathbf{R}}\mathbf{a}\,,$$

and

$$\mathbf{x} = (\mathbf{B}^T\mathbf{R}\mathbf{B})^{-1}\,\mathbf{B}^T\mathbf{R}\mathbf{a}\,,$$

thus

$$\begin{aligned}\mathbf{x}-\bar{\mathbf{x}} &= (\mathbf{B}^T\mathbf{R}\mathbf{B})^{-1}\,\mathbf{B}^T\mathbf{R}\mathbf{a} - (\mathbf{B}^T\bar{\mathbf{R}}\mathbf{B})^{-1}\,\mathbf{B}^T\bar{\mathbf{R}}\mathbf{a} = \\ &= [(\mathbf{B}^T\mathbf{R}\mathbf{B})^{-1} - (\mathbf{B}^T\bar{\mathbf{R}}\mathbf{B})^{-1}]\,\mathbf{B}^T\mathbf{R}\mathbf{a} + (\mathbf{B}^T\bar{\mathbf{R}}\mathbf{B})^{-1}\,\mathbf{B}^T(\mathbf{R}-\bar{\mathbf{R}})\mathbf{a} = \\ &= -(\mathbf{B}^T\mathbf{R}\mathbf{B})^{-1}\,\mathbf{B}^T(\mathbf{R}-\bar{\mathbf{R}})\,\mathbf{B}(\mathbf{B}^T\mathbf{R}\mathbf{B})^{-1}\,\mathbf{B}^T\mathbf{R}\mathbf{a} + \\ &+ (\mathbf{B}^T\bar{\mathbf{R}}\mathbf{B})^{-1}\,\mathbf{B}^T(\mathbf{R}-\bar{\mathbf{R}})\mathbf{a} = \\ &= -(\mathbf{B}^T\mathbf{R}\mathbf{B})^{-1}\,\mathbf{B}^T(\mathbf{R}-\bar{\mathbf{R}})\,[\mathbf{B}(\mathbf{B}^T\bar{\mathbf{R}}\mathbf{B})^{-1}\,\mathbf{B}^T\mathbf{R}\mathbf{a}-\mathbf{a}]\,.\end{aligned}$$

From here

$$\begin{aligned}\|\mathbf{x}-\bar{\mathbf{x}}\| &= \|(\mathbf{B}^T\mathbf{R}\mathbf{B})^{-1}\,\mathbf{B}^T(\mathbf{R}-\bar{\mathbf{R}})\,[\mathbf{B}(\mathbf{B}^T\bar{\mathbf{R}}\mathbf{B})^{-1}\,\mathbf{B}^T\mathbf{R}\mathbf{a}-\mathbf{a}]\| \le \\ &\le \|(\mathbf{B}^T\mathbf{R}\mathbf{B})^{-1}\|\,\|\mathbf{B}^T(\mathbf{R}-\bar{\mathbf{R}})\,[\mathbf{B}(\mathbf{B}^T\bar{\mathbf{R}}\mathbf{B})^{-1}\,\mathbf{B}^T\mathbf{R}\mathbf{a}-\mathbf{a}]\|\,.\end{aligned}$$

Since

$$\frac{\mathbf{u}^T(\mathbf{B}^T\mathbf{R}\mathbf{B})\mathbf{u}}{\mathbf{u}^T\mathbf{u}} = \frac{(\mathbf{B}\mathbf{u})^T\,\mathbf{R}(\mathbf{B}\mathbf{u})}{\|\mathbf{B}\mathbf{u}\|^2}\,\frac{\mathbf{u}^T\mathbf{B}^T\,\mathbf{B}\mathbf{u}}{\|\mathbf{u}\|^2}$$

and

$$0 < m_c\|\mathbf{u}\| \le \mathbf{u}^T\mathbf{C}\mathbf{u} \le M_c\|\mathbf{u}\|^2$$

for any positive semi-definite matrix $\mathbf{C}$, where m_c is the minimum and M_c the maximum eigenvalue of matrix $\mathbf{C}$, and $m_c > 0$ if $\mathbf{C}$ is non-singular. Thus, denoting the minimum eigenvalue of $\mathbf{B}^T\mathbf{R}\mathbf{B}$ by μ and the associated eigenvector by $\mathbf{u}$, we have

$$\mu = \frac{\mathbf{u}^T(\mathbf{B}^T\mathbf{R}\mathbf{B})\mathbf{u}}{\|\mathbf{u}\|^2} \ge m_R m_{B^TB} > 0\,,$$

and therefore the maximum eigenvalue of $(\mathbf{B}^T\mathbf{R}\mathbf{B})^{-1}$ lies below

$$\frac{1}{m_R m_{B^TB}}\,.$$

Since here $m_R = \min_i r^{(i)}$, we obtain the following estimation:

$$\|(\mathbf{B}^T\mathbf{R}\mathbf{B})^{-1}\| \le \frac{1}{m_R m_{B^TB}} = \frac{1}{\min_i r^{(i)} m_{B^TB}}\,.$$

On the other hand,

$$\|\mathbf{B}^T(\mathbf{R}-\bar{\mathbf{R}})\,[\mathbf{B}(\mathbf{B}^T\bar{\mathbf{R}}\mathbf{B})^{-1}\,\mathbf{B}^T\mathbf{R}\mathbf{a}-\mathbf{a}]\|\leq$$

$$\leq M_{BB^T}^{1/2}\|\mathbf{R}-\bar{\mathbf{R}}\|\,[\|\mathbf{B}(\mathbf{B}^T\bar{\mathbf{R}}\mathbf{B})^{-1}\,\mathbf{B}^T\mathbf{R}\mathbf{a}\|+\|\mathbf{a}\|]\leq$$

$$\leq M_{BB^T}^{1/2}\max_i|r^{(i)}-\bar{r}^{(i)}|\left(M_{B^TB}\frac{\max\limits_i r^{(i)}}{\min\limits_i \bar{r}^{(i)}m_{B^TB}}+1\right)\|\mathbf{a}\|.$$

Consequently,

$$\|\mathbf{x}-\bar{\mathbf{x}}\|\leq\frac{M_{BB^T}^{1/2}}{m_{B^TB}}\frac{\max|r^{(i)}-\bar{r}^{(i)}|}{\min\limits_i r^{(i)}}\left(\frac{M_{B^TB}}{m_{B^TB}}\frac{\max\limits_i r^{(i)}}{\min\limits_i \bar{r}^{(i)}}+1\right)\|\mathbf{a}\|\leq$$

$$\leq\frac{M_{BB^T}^{1/2}}{m_{B^TB}}\frac{\lambda\max\bar{r}^{(i)}}{(1-\lambda)\min\bar{r}^{(i)}}\left(\frac{M_{B^TB}(1+\lambda)\max\bar{r}^{(i)}}{m_{B^TB}\min\bar{r}^{(i)}}+1\right)\|\mathbf{a}\|.$$

However, the above obtained estimation is rather rough and disadvantageous in that it determines a considerable error limit for deviation $\mathbf{x}-\bar{\mathbf{x}}$ even in case the same solution is associated with the perturbed system as with the reference equation (e.g. $\mathbf{R}=c\bar{\mathbf{R}}$ or $\mathbf{Bx}-\mathbf{a}=\mathbf{0}$ can be solved).

It would be practicable to give an estimation where the identical solutions associated with the different perturbed $\mathbf{R}s$ were evaluated identically.

Let us denote the set of elements of the permissible perturbance range, resulting in the same element $\mathbf{x}$ as a solution if assumed for the perturbed system,

$$\mathscr{R}(\mathbf{x})=\{\mathbf{R}=\langle\mathbf{r}\rangle:\mathbf{B}^T\mathbf{R}(\mathbf{Bx}-\mathbf{a})=\mathbf{0},\quad\mathbf{r}\in\Delta R\},$$

$\mathscr{R}(\mathbf{x})$ being the section of a subspace with ΔR for any $\mathbf{x}$.

Since $\mathbf{R}$ is diagonal, thus equation (2) is equivalent to equation

$$\mathbf{B}^T\langle\mathbf{Bx}-\mathbf{a}\rangle\mathbf{r}=\mathbf{0}, \tag{3}$$

hence, the elements of $\mathscr{R}(\mathbf{x})$ are obtained by solving equation (3).

Let

$$\mathbf{B}=\begin{bmatrix}\mathbf{B}_1\\ \mathbf{B}_2\\ \mathbf{B}_3\end{bmatrix} \tag{4}$$

be a decomposition of $\mathbf{B}$ such as to result in dimension $-(n_1\times n)$ for $\mathbf{B}_1$, $-(n_2\times n)$ for

$\mathbf{B}_2$ and $(m-n_1-n_2)\times n$ for $\mathbf{B}_3$ and the following relations are satisfied

$$\begin{aligned}\mathbf{B}_1\mathbf{x}-\mathbf{a}_1&\neq\mathbf{0}\\ \mathbf{B}_2\mathbf{x}-\mathbf{a}_2&\neq\mathbf{0}\\ \mathbf{B}_3\mathbf{x}-\mathbf{a}_3&=\mathbf{0},\end{aligned}\tag{5}$$

and

$$\operatorname{rank}\begin{bmatrix}\mathbf{B}_1\\ \mathbf{B}_2\end{bmatrix}=\operatorname{rank}\ \mathbf{B}_1=n_1\leq n.$$

It is not excluded that $n_2=0$ or $n_1+n_2=m$ that is $\mathbf{B}_2$ or $\mathbf{B}_3$ are missing (all blocks can not be missing simultaneously because of the condition $m>n$). The vector

$$\mathbf{r}=\begin{bmatrix}\mathbf{r}_1\\ \mathbf{r}_2\\ \mathbf{r}_3\end{bmatrix}$$

will be a solution of the equation (3) if

$$\mathbf{r}_i\in\Delta R_i,\quad \Delta R_i=\{\mathbf{r}_i:(\mathbf{1}-\lambda)\bar{\mathbf{r}}_i\leq\mathbf{r}_i\leq(1+\lambda)\bar{\mathbf{r}}_i,\qquad i=1,2,3$$

$$\mathbf{B}_1^T\langle\mathbf{B}_1\mathbf{x}-\mathbf{a}_1\rangle\mathbf{r}_1+\mathbf{B}_2^T\langle\mathbf{B}_2\mathbf{x}-\mathbf{a}_2\rangle\mathbf{r}_2=\mathbf{0}\tag{6}$$

and $\mathbf{r}_3\in\Delta R_3$ is arbitrary.

If $n_2=0$ that is the block $\mathbf{B}_2$ is missing, then there will be only a trivial solution $\mathbf{r}_1=0$ of (6), the columns of $\mathbf{B}_1^T$ being linearly independent but, according to our conditions, $\mathbf{0}\notin\Delta R_i$. This means that $\mathscr{R}(\mathbf{x})$ is an empty set for any $\mathbf{x}$ where $\mathbf{B}_2$ is missing in the partitions (4)–(5) of $\mathbf{B}$.

If $n_2\neq0$, then there will exist a matrix $\mathbf{L}$ of dimension $(n_2\times n_1)$ resulting in $\mathbf{B}_2=\mathbf{L}\mathbf{B}_1$ that is

$$\mathbf{B}_1^T[\langle\mathbf{B}_1\mathbf{x}-\mathbf{a}_1\rangle\mathbf{r}_1+\mathbf{L}^T\langle\mathbf{B}_2\mathbf{x}-\mathbf{a}_2\rangle\mathbf{r}_2]=\mathbf{0}.$$

Since the columns of $\mathbf{B}_1^T$ are linearly independent, this is possible only if

$$\langle\mathbf{B}_1\mathbf{x}-\mathbf{a}_1\rangle\mathbf{r}_1+\mathbf{L}^T\langle\mathbf{B}_2\mathbf{x}-\mathbf{a}_2\rangle\mathbf{r}_2=\mathbf{0}.\tag{7}$$

Taking into consideration that $\langle\mathbf{B}_1\mathbf{x}-\mathbf{a}_1\rangle$ can be inverted, we have:

$$\mathbf{r}_1=-\langle\mathbf{B}_1\mathbf{x}-\mathbf{a}_1\rangle^{-1}\mathbf{L}^T\langle\mathbf{B}_2\mathbf{x}-\mathbf{a}_2\rangle\mathbf{r}_2.$$

Since ΔR is the Cartesian product of subsets ΔR_i, $i=1, 2, 3$, that is $\Delta R=\Delta R_1\times\Delta R_2\times\Delta R_3$, thus

$$\mathscr{R}(\mathbf{x})=\begin{cases}\{\mathbf{r}\in\Delta R:\mathbf{r}_i\in\Delta R_i,\quad i=1,2,3,\quad \mathbf{r}_1=\mathbf{G}(\mathbf{x})\mathbf{r}_2\},\\ \emptyset\qquad\text{if}\quad n_2=0,\end{cases}\tag{8}$$

where

$$\mathbf{G}(\mathbf{x})=-\langle\mathbf{B}_1\mathbf{x}-\mathbf{a}_1\rangle^{-1}\mathbf{L}^T\langle\mathbf{B}_2\mathbf{x}-\mathbf{a}_2\rangle.\tag{9}$$

3. Stochastic approach

Let $\mathbf{r}$ be a stochastic vector variable, its elements being random variables with density function $\mu^{(i)}(\mathbf{r}^{(i)})$. Then the density function of $\mathbf{r}$ will be $\mu(\mathbf{r}) = \prod_i \mu^{(i)}(r^{(i)})$. Assume that ΔR is the support of $\mu(\mathbf{r})$, that is,

$$\mu(\mathbf{r}) = \begin{cases} >0 & \text{if } \mathbf{r} \in \Delta R \\ \mu = 0 & \text{if } \mathbf{r} \notin \Delta R \end{cases}$$

and $\int_{\Delta R} \mu(\mathbf{r})\, d\mathbf{r} = 1$.

Let be $\boldsymbol{\rho} \in \mathscr{R}(\mathbf{x})$. Using the relationship $\boldsymbol{\rho}_1 = \mathbf{G}(\mathbf{x})\boldsymbol{\rho}_2$ we can write

$$\boldsymbol{\rho} = \begin{bmatrix} \boldsymbol{\rho}_1 \\ \boldsymbol{\rho}_2 \\ \boldsymbol{\rho}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{G}(\mathbf{x}) & \mathbf{0} \\ \mathbf{E} & \mathbf{0} \\ \mathbf{0} & \mathbf{E} \end{bmatrix} \begin{bmatrix} \mathbf{r}_2 \\ \mathbf{r}_3 \end{bmatrix},$$

that is $\boldsymbol{\rho}$ is a function of independent random variables $\mathbf{r}_2$ and $\mathbf{r}_3$ and therefore its distribution function can be written, as follows [4]:

$$\Phi(\mathbf{s}) = \iint_{D_s(\mathbf{x})} \mu_2(\mathbf{r}_2)\mu_3(\mathbf{r}_3)\, d\mathbf{r}_2\, d\mathbf{r}_3, \tag{10}$$

where

$$D_s(\mathbf{x}) = \left\{ \begin{bmatrix} \mathbf{r}_2 \\ \mathbf{r}_3 \end{bmatrix} : \mathbf{G}(\mathbf{x})\mathbf{r}_2 \leq \mathbf{s}_1, \mathbf{r}_2 \leq \mathbf{s}_2, \mathbf{r}_3 \leq \mathbf{s}_3, \begin{bmatrix} \mathbf{G}(\mathbf{x})\mathbf{r}_2 \\ \mathbf{r}_2 \\ \mathbf{r}_3 \end{bmatrix} \in \Delta R \right\}. \tag{11}$$

$\Phi(\mathbf{s})$ is the probability that the random variable $\boldsymbol{\rho}$ falls within the range

$$\{\mathbf{r} : \mathbf{r}_1 = \mathbf{G}(\mathbf{x})\mathbf{r}_2, (1-\lambda)\bar{\mathbf{r}}_1 \leq \mathbf{r}_1 \leq \mathbf{s}_1, (1-\lambda)\bar{\mathbf{r}}_2 \leq \mathbf{r}_2 \leq \mathbf{s}_2,$$
$$(1-\lambda)\bar{\mathbf{r}}_3 \leq \mathbf{r}_3 \leq \mathbf{s}_3\}.$$

If $\mathbf{s}_i$ moves to the boundary of the range ΔR, then this range will go over into $\mathscr{R}(\mathbf{x})$ and thus

$$\Phi(1+\lambda)\bar{\mathbf{r}}_1, (1+\lambda)\bar{\mathbf{r}}_2, (1+\lambda)\bar{\mathbf{r}}_3 = P(\mathbf{r} \in \mathscr{R}(\mathbf{x})) = p(\mathbf{x}) \tag{12}$$

$p(x)$ shows, with which probability there exist an flexibility matrix in the perturbation range such that the corresponding perturbed system would have solution $\mathbf{x}$, that is the probability of $\mathbf{x}$ being a solution to any possible perturbed system. Using (10), the

calculation of (12) requires the multiple integral

$$p(\mathbf{x}) = \iint_{D(\mathbf{x})} \mu_2(\mathbf{r}_2)\mu_3(\mathbf{r}_3)\,\mathrm{d}\mathbf{r}_2\,\mathrm{d}\mathbf{r}_3 \tag{13}$$

to be calculated, where $D(\mathbf{x})$ is the projection of $\mathscr{R}(\mathbf{x})$ on the subspace of dimension $n-n_1$, expanded by $\mathbf{r}_2$, $\mathbf{r}_3$. Making use of the independence of $\mathbf{r}_2$ and $\mathbf{r}_3$ we obtain

$$p(\mathbf{x}) = \int_{\Delta R_3} \mu_3(\mathbf{r})\,\mathrm{d}\mathbf{r} \int_{D_2(\mathbf{x})} \mu_2(\mathbf{r})\,\mathrm{d}\mathbf{r},$$

where

$$D_2(\mathbf{x}) = \{\mathbf{r}_2 \in \Delta R_2,\ G(\mathbf{x})\mathbf{r}_2 \in \Delta R_1\}. \tag{14}$$

Since the first integral is equal to 1 because of the density function behaviour, we obtain as the final form of our estimation:

$$p(\mathbf{x}) = \int_{D_2(\mathbf{x})} \mu_2(\mathbf{r})\,\mathrm{d}\mathbf{r}. \tag{15}$$

If $X_s = \{\mathbf{x} \in \mathbb{R}^n : p(\mathbf{x}) = S\} \neq 0$, then X_s will determine the solution set of probability s while $\bigcup_{\mathbf{x} \in X_s} \mathscr{R}(\mathbf{x})$ the possible perturbed flexibility matrices of probability s.

As will be seen in the example given, the analytical definition of the integral given in (15) is very complicated due to the dependence of the range on $\mathbf{x}$ even with a small number of dimensions.

4. Fuzzy approach

Let the degree of acceptability of $\mathbf{r}$ for the problem, with a value of $0 \leq \mu(\mathbf{r}) \leq 1$, be allocated to each $\mathbf{r}$. Here $\mu(\mathbf{r})$ need not necessarily be a concept taken from the probability theory although quite often some density function in the form of $f(\mathbf{r})/\sup f(\mathbf{r})$, normed to 1, can be selected to express this degree. $\mu(\mathbf{r})$ is often determined by expert estimate.

$$\mu_R = \{(\mathbf{r}, \mu(\mathbf{r})) : \mathbf{r} \in \mathbb{R}^m,\ \mu : \mathbb{R}^m \to [0, 1]\}$$

is the fuzzy set of matrices $\mathbf{R} = \langle \mathbf{r} \rangle$ while $\mu(\mathbf{r})$ the membership function of the elements.

μ_R can be given either with the common degree of acceptability of $\mathbf{r}$ as was defined above or with the membership function $\mu_i(r_i)$ for each element. In the latter case, μ_R is the Cartesian product of fuzzy sets μ_{r_i}

$$\mu_R = \bigtimes_i \mu_{r_i} = \{(\mathbf{r}, \mu(\mathbf{r})) : \mathbf{r} \in \mathbb{R}^m,\ \mu(\mathbf{r}) = \min_i \mu_i(r_i),\ \mu_i : \mathbb{R} \to [0, 1]\}. \tag{16}$$

Here $\mu(\mathbf{r})$ is not necessarily the common density function normed to 1 provided $\mu_i(r_i)$ is a density function normed to 1.

The Equation (2) defines an operator $F: \mathbb{R}^m \to \mathbb{R}^n$ wich maps every $\mathbf{r} \in \mathbb{R}^m$ to a solution $\mathbf{x} \in \mathbb{R}^h$.

The operator F generates a fuzzy transformation $\hat{F}$ transforming the fuzzy sets of $\mathbb{R}^m$ into the fuzzy sets of $\mathbb{R}^n$ with the membership function [5]

$$v(\mathbf{x}) = \sup_{\mathbf{r} \in F^{-1}(\mathbf{x})} \mu(\mathbf{r})$$

where $F^{-1}(\mathbf{x}) = \{\mathbf{r} \in \mathbb{R}^m : F(\mathbf{r}) = \mathbf{x}\}$.

Assume that ΔR is the support of μ_R that is

$$\mu(\mathbf{r}) = \begin{cases} \in [0, 1], & \text{if } \mathbf{r} \in \Delta R, \\ 0, & \text{if } \mathbf{r} \notin \Delta R. \end{cases}$$

In this case, $F^{-1}(\mathbf{x}) = \mathscr{R}(\mathbf{x})$, thus

$$v(\mathbf{x}) = \begin{cases} \sup_{\mathbf{r} \in \mathbb{R}(\mathbf{x})} \mu(\mathbf{r}) \\ 0 \qquad \text{if } \mathscr{R}(\mathbf{x}) = \emptyset. \end{cases} \tag{17}$$

If in the system of equations (2) $\mathbf{r}$ is considered to be an element of fuzzy set μ_R, then (2) will go over into a fuzzy equation system the solution of which is also a fuzzy set with a membership function (16). $v(\mathbf{x})$ gives in what degree can a point of the space be considered to be the solution of a perturbed system in which the perturbed $\mathbf{r}$ is acceptable in respect of the problem on the level $\mu(\mathbf{r})$.

If the membership function $\mu(\mathbf{r})$ fulfils the conditions

a) $\mu(\mathbf{r}) = 1$ only if $\mathbf{r} = \bar{\mathbf{r}}$,

b) $\mu(\mathbf{r})$ upper semicontinuous i.e.

$$\mu(\mathbf{r}^*) \geq \limsup_{\mathbf{r} \to \mathbf{r}^*} \mu(\mathbf{r})$$

for any limit point $\mathbf{r}^*$

c) $\Lambda_c = \{\mathbf{r} \in \mathbb{R}^m : \mu(\mathbf{r}) \geq C\}$ bounded and closed for any $0 \leq C \leq 1$,

then the fuzzy solution defined by (17) will have the following properties [6]:

1) If $\mathbf{x}_1$ and $\mathbf{x}_2$ are a solution each, associated with two different realizations $\mathbf{r}_1$ and $\mathbf{r}_2$, respectively and if $v(\mathbf{x}_1) \geq v(\mathbf{x}_2)$, then there exist $\mathbf{r} \in \mathscr{R}(\mathbf{x}_1)$ such that $\mu(\mathbf{r}) \geq \mu(\mathbf{r}_2)$. In other words, a 'better' solution can always be realized by a more 'acceptable' flexibility matrix.
2) $v(\mathbf{x}) = 1$ only if $\bar{\mathbf{r}} \in \mathscr{R}(\mathbf{x})$.

The optimization problem supplying the solution (17) is equivalent to the following conditional optimization problem

$$\sup_{(\mathbf{r}, \lambda) \in \mathbb{R}^m \times \mathbb{R}} \lambda \tag{18}$$

subject to

$$\mu(\mathbf{r}) \geq \lambda,$$

$$\mathbf{B}^T < \mathbf{Bx} - \mathbf{a} > \mathbf{r} = \mathbf{0},$$

$$\mathbf{r} \in \Delta R.$$

If the fuzzy set μ_R is given as a Cartesian product of the fuzzy set of the elements of $\mathbf{r}$ the conditional optimization problem may be given as follows

$$\sup_{(\mathbf{r}, \lambda) \in \mathbb{R}^m \times \mathbb{R}} \lambda$$

subject to

$$\mu_i(\mathbf{e}_i \mathbf{r}) \geq \lambda, \qquad i = 1, 2, \ldots, n, \tag{19}$$

$$\mathbf{B}^T < \mathbf{Bx} - \mathbf{a} > \mathbf{r} = \mathbf{0},$$

$$\mathbf{r} \in \Delta R,$$

which in the general case is a non–linear programming problem. Using relationships (8) and (14) the membership function of the fuzzy solution set defined by (17) has the form

$$v(\mathbf{x}) = \sup_{\substack{\mathbf{r}_2 \in D_2(\mathbf{x}) \\ \mathbf{r}_3 \in \Delta R_3}} \mu(G(\mathbf{x})\mathbf{r}_2, \mathbf{r}_2, \mathbf{r}_3) \quad \text{if} \quad \mathscr{R}(\mathbf{x}) \neq \emptyset. \tag{20}$$

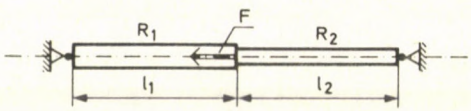


Fig. 1

5. Examples

5.1. Example

Here

$$\mathbf{B} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad \bar{\mathbf{R}} = \langle 90, 100 \rangle, \qquad \mathbf{a} = \begin{bmatrix} -F \\ 0 \end{bmatrix}.$$

Solution of the reference equation:

$$\bar{x} = \frac{9}{19} F.$$

With a perturbance constant $\lambda = 0.1$

$$\Delta R = \{\mathbf{r} : 81 \leq r^{(1)} \leq 99, 90 \leq r^{(2)} \leq 110\}.$$

a) Deterministic estimation:

Since

$$M_{BTB} = M_{BBT} = m_{BTB} = 2,$$

thus

$$|x - \bar{x}| \leq \frac{\sqrt{2}}{2} \frac{0.1}{0.9} \frac{100}{90} \left(\frac{2}{2} 1.1 \frac{100}{90} + 1 \right) F =$$

$$= \frac{\sqrt{2}10}{162} \frac{20}{9} F = \frac{100\sqrt{2}}{729} F \sim 0.1939936 \ldots F.$$

In the reality

$$\max |x - \bar{x}| = \frac{11}{21} F - \frac{9}{19} F = 0.0501253 \ldots F$$

$$\mathbf{r} \in \Delta R$$

that is, we have got a fourfold overestimate.

b) Stochastic estimation

Since $\mathscr{R}(F) = \emptyset$, it is enough to investigate the case where none of the equations of equation system $\mathbf{Bx} - \mathbf{A} = \mathbf{0}$ is fulfilled.

In this case

$$B_1 = [1], \qquad A_1 = -F,$$

$$B_2 = [1], \qquad A_2 = \quad 0,$$

that is, (7) $r_1 = \dfrac{x}{F - x} r_2$

takes the following shape:

$$D_2(\mathbf{x}) = \left\{ r_2 \in \Delta R_2 : \frac{x}{F - x} r_2 \in \Delta R_1 \right\}.$$

This means that the inequalities

$$90 \leq r_2 \leq 110, \tag{21}$$

$$81 \leq \frac{x}{F - x} r_2 \leq 99$$

must be fulfilled.

Therefore, it is obvious that if $x > F$ then $\mathscr{R}(x) = \emptyset$. As can be confirmed by calculation, there is no solution to inequality system (21) if $0 < x < (81/191)F$ and $x > (11/21)F$ either, that is $D_2(x)$ and $\mathscr{R}(x)$ are uniquely empty for these x. Also,

$$D_2(x)=\begin{cases}\left\{r_2:\dfrac{81(F-x)}{x}\le r_2\le 110\right\}, & \text{if } \dfrac{81}{191}F\le x\le\dfrac{9}{19}F,\\[2ex] \left\{r_2:90\le r_2\le\dfrac{99(F-x)}{x}\right\}, & \text{if } \dfrac{9}{19}F\le x\le\dfrac{11}{21}F,\end{cases}\tag{22}$$

and thus

$$\mathscr{R}(x)=\begin{cases}\left\{(r_1,r_2):r_2\in D_2(x),\ r_1=\dfrac{x}{F-x}r_2\right\},\\ \text{if } \dfrac{81}{191}F\le x\le\dfrac{11}{21}F,\\ \emptyset \quad \text{otherwise.}\end{cases}\tag{23}$$

1. Should each co-ordinate of $\mathbf{r}$ be uniformly distributed in ΔR, then, on the basis of (15), the following estimate will be obtained (Fig. 2):

$$p(x)=\begin{cases}\displaystyle\int_{\frac{81(F-x)}{x}}^{110}\frac{1}{20}\,dr_2=\frac{191x-81F}{20x}, & \text{if } \dfrac{81F}{191}\le x\le\dfrac{9}{19}F,\\[2ex] \displaystyle\int_{90}^{\frac{99(F-x)}{x}}\frac{1}{20}\,dr_2=\frac{99F-189x}{20x}, & \text{if } \dfrac{9}{19}F\le x\le\dfrac{11}{21}F,\\[2ex] 0 & \text{otherwise.}\end{cases}$$

2. Should each co-ordinate of $\mathbf{r}$ of normal distribution truncated to ΔR, with an expectable value of $\bar{\mathbf{r}}$ and a variation of $\sigma_i=\lambda\bar{r}_i/4$ (such a selection of the variation meaning that the perturbance range has been defined with four times the variation so

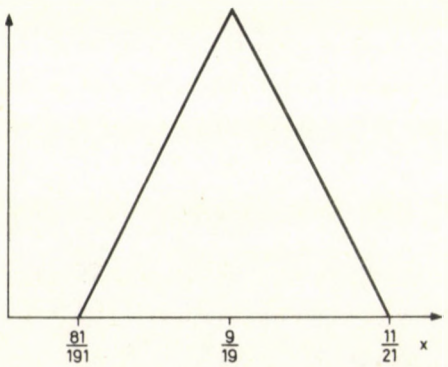


Fig. 2

that practically the truncation can no longer be observed), then, on the basis of (15), the following estimate will be obtained (Fig. 3):

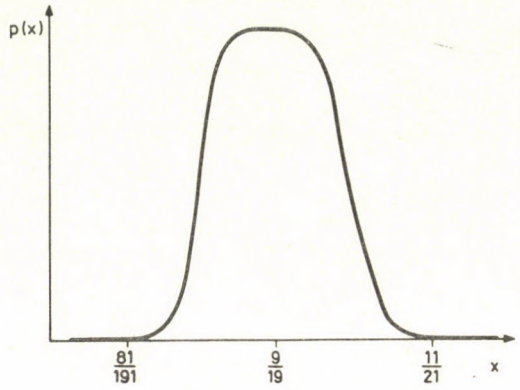


Fig. 3

x	$p(x)$	
	$\frac{1}{\sqrt{2\pi}\sigma_2}\int_{\frac{81(F-x)}{x}}^{110}\exp\left[-\frac{(r_2-100)^2}{2\sigma_2^2}\right]dr_2$	$\frac{1}{\sqrt{2\pi}\sigma_2}\int_{90}^{\frac{99(F-x)}{x}}\exp\left[-\frac{(r_2-100)^2}{2\sigma_2^2}\right]dr_2$
81/191	0	
0.43	0.0015	
0.44	0.1093	
0.45	0.6554	
0.46	0.9750	
0.47	0.9997	
9/19	1	1
0.48		0.9981
0.49		0.8875
0.50		0.3446
0.51		0.0356
0.52		0.0004
11/21		0

If, as compared with the variation, the perturbance range is too tight to permit the truncation to be neglected, then the maximum of $p(x)$ will remain below 1.

c) Fuzzy estimation

Let the common membership function $\mu(\mathbf{r})$ be the normal density function truncated to ΔR normed to 1, with expectable value $\bar{\mathbf{r}}$ and variation

$$\sigma_i = \frac{\lambda \bar{r}_i}{4}.$$

Then, according to (20), we have the following problem to solve:

$$\max_{r_2 \in D_2(x)} e^{-\frac{1}{2}\left[\frac{(x/(F-x)r_2-90)^2}{(9/4)^2}+\frac{(r_2-100)^2}{(10/4)^2}\right]}, \tag{24}$$

where $D_2(x)$ is the same as in the stochastic case that is, defined with (22) for any x for which it is not empty.

(24) is equivalent to problem

$$\min_{r_2 \in D_2(x)} \frac{\left(\frac{x}{F-x}r_2-90\right)^2}{81}+\frac{(r_2-10)^2}{100}. \tag{25}$$

Solution to (25):

$$r_2^* = \begin{cases} \dfrac{900(F-x)(x+9F)}{100x^2+81(F-x)^2} & \text{if } \dfrac{81}{191}F \le x \le 0.5199451\ldots F \\ 90 & \text{if } 0.5199451\ldots F \le x \le \dfrac{11}{21}F \end{cases}$$

Of this,

$$r_1^* = \begin{cases} \dfrac{900x/(x+9F)}{100x^2+81(F-x)^2} & \text{if } \dfrac{81}{191}F \le x \le 0.5199451\ldots F \\ \dfrac{90x}{F-x} & \text{if } 0.5199451\ldots F \le x \le \dfrac{11}{21}F \end{cases}$$

Thus we obtain the following fuzzy estimation (Fig. 4):

x	$v(x)$
81/191	0.000 ...
0.43	0.000 ...
0.44	0.000 ...
0.45	0.027 ...
0.46	0.299 ...
0.47	0.916 ...
9/19	1
0.48	0.774 ...
0.49	0.182 ...
0.50	0.012 ...
0.51	0.000 ...
0.5199451	0.000 ...
0.52	0.000 ...
11/21	

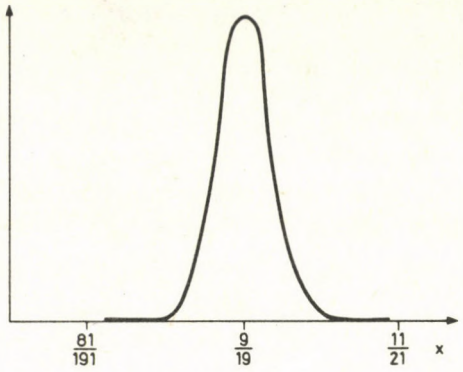


Fig. 4

5.2 Example

Here

$$\mathbf{B}=\begin{bmatrix} 1 & 0 \\ -\sqrt{2} & -1 \\ 1 & \sqrt{2} \\ 0 & 1 \end{bmatrix}, \qquad \bar{\mathbf{R}}=<90, 100, 100, 90>, \qquad a=\begin{bmatrix} 0 \\ -F \\ \sqrt{2}F \\ 0 \end{bmatrix},$$

Then, the solution of the reference equation:

$$\bar{\mathbf{x}}=\left[\frac{180\sqrt{2}}{721}F,\ \frac{370}{721}F\right].$$

Let the perturbance constant be $\lambda=0.1$, that is

$$\Delta R=\{r: 81\leq r^{(1)}\leq 99,\ 90\leq r^{(2)}\leq 110,\ 90\leq r^{(3)}\leq 110,\ 81\leq r^{(4)}\leq 99\}.$$

1) Deterministic estimation:

Since

$$M_{B^TB}=M_{BB^T}=4+2\sqrt{2},$$

$$m_{B^TB}=4-2\sqrt{2}.$$

thus

$$\|\mathbf{x}-\bar{\mathbf{x}}\|\leq\frac{(4+2\sqrt{2})^{1/2}}{4-2\sqrt{2}}\frac{10}{81}\left(\frac{4+2\sqrt{2}}{4-2\sqrt{2}}\frac{110}{90}+1\right)\sqrt{3}F\sim 0.576958\ldots F$$

Calculating with the actual dimensional variations this value is obtained as

$$\|\mathbf{x}-\bar{\mathbf{x}}\|=0.290826$$

i.e. the overestimation is twofold.

2) Stochastic estimation:

Assume that each co-ordinate of $\mathbf{r}$ is of uniform distribution.

Only the range of $\mathbf{x}>0$ will be investigated. With this condition, only one equality will be fulfilled in equation $\mathbf{Bx}-\mathbf{a}=\mathbf{0}$ if $\mathbf{x}$ is an element of set

$$T_1=\left\{(x_1,x_2)\in\mathbb{R}_+^2 : x_2=-\sqrt{2}x_1+F,\ x_1\in\left(0,\frac{\sqrt{2}}{2}\right)F\right\},$$

or

$$T_2=\left\{(x_1,x_2)\in\mathbb{R}_+^2 : x_2=-\frac{\sqrt{2}}{2}x_1+F,\ x_1\in(0,\sqrt{2}F)\right\}.$$

None of the equations of equation system $\mathbf{Bx}-\mathbf{a}=\mathbf{0}$ can be fulfilled in any other case.

Let it be designated $T_3=\mathbb{R}_+^2\setminus(T_1\cup T_2)$.

Tabulated below are the appropriate partitioning of $\mathbf{B}$ and $\mathbf{a}$, and matrices $\mathbf{L}$ and $\mathbf{G}(\mathbf{x})$ associated.

	$\mathbf{B}_3$	$\mathbf{B}_1$	$\mathbf{B}_2$	$\mathbf{L}$	$\mathbf{a}_3$	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{G}$
$\mathbf{x}\in T_1$	$[-\sqrt{2}\ \ -1]$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	$[1\ \ \sqrt{2}]$	$[1\ \ \sqrt{2}]$	$-F$	$\begin{bmatrix}0\\0\end{bmatrix}$	$\sqrt{2}F$	$\begin{bmatrix}-1\\ x_1/(x_1-\sqrt{2}F/2\end{bmatrix}$
$\mathbf{x}\in T_2$	$[1,\ \sqrt{2}]$	$\begin{bmatrix}1&0\\0&1\end{bmatrix}$	$[-\sqrt{2}\ \ -1]$	$[-\sqrt{2}\ \ -1]$	$\sqrt{2}F$	$\begin{bmatrix}0\\0\end{bmatrix}$	$-F$	$\begin{bmatrix}1\\ -x_1/(x_1-\sqrt{2}F)\end{bmatrix}$
$\mathbf{x}\in T_3$	–	$\begin{bmatrix}1&0\\0&1\end{bmatrix}\begin{bmatrix}-\sqrt{2}&-1\\1&\sqrt{2}\end{bmatrix}\begin{bmatrix}-\sqrt{2}&-1\\1&\sqrt{2}\end{bmatrix}$	–	$\begin{bmatrix}0\\0\end{bmatrix}\begin{bmatrix}-F\\\sqrt{2}F\end{bmatrix}$				$\begin{bmatrix}-\sqrt{2}g_1/x_1 & -g_2/x_1\\ g_1/x_2 & -\sqrt{2}g_2/x_2\end{bmatrix}$ where $g_1=-\sqrt{2}x_1-x_2+F$, $g_2=x_1+\sqrt{2}x_2-\sqrt{2}F$

If $\mathbf{x}\in T_1$, then the following unequalities should be fulfilled for the determination of $\mathscr{R}(\mathbf{x})$:

$$90\le r^{(2)}\le 110,$$

$$90\le r^{(3)}\le 110,$$

$$81 \leq r^{(3)} = r^{(1)} \leq 99,$$

$$81 \leq \frac{x_1}{x_1 - \frac{\sqrt{2}}{2} F} r^{(3)} = r^{(4)} \leq 99 .$$

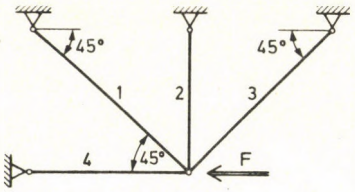


Fig. 5

Since the solution set of the above inequality system is empty, thus also $\mathscr{R}(\mathbf{x})$ and together with it $D_2(\mathbf{x})$ are empty. Therefore

$$p(\mathbf{x}) = 0 \qquad \text{if} \quad \mathbf{x} \in T_1^* .$$

If $\mathbf{x} \in T_2$, then $\mathscr{R}(\mathbf{x})$ will be determined by the system of conditions

$$90 \leq r^{(3)} \leq 110 ,$$

$$90 \leq r^{(2)} \leq 110 ,$$

$$81 \leq r^{(2)} = r^{(1)} \leq 99 ,$$

$$90 \leq \frac{-x_1}{x_1 - \sqrt{2}F} r^{(2)} = r^{(4)} \leq 110 .$$

From here,

$$D_2(\mathbf{x}) = \{r^{(2)} : \max(90, 90(\sqrt{2}F - x_1)/x_1) \leq r^{(2)} \leq \min(99, 110(\sqrt{2}F - x_1)/x_1)\} .$$

Thus

$$\mathbf{D}_2(\mathbf{x}) = \begin{cases} \left\{ r^{(2)} : \dfrac{90(\sqrt{2}F - x_1)}{x_1} \leq r^{(2)} \leq 99 \right\}, & \text{if} \quad x_1 \in \left[\dfrac{10\sqrt{2}}{21} F, \dfrac{\sqrt{2}}{2} F \right], \\ \{ r^{(2)} : 90 \leq r^{(2)} \leq 99 \}, & \text{if} \quad x_1 \in \left[\dfrac{\sqrt{2}}{2} F, \dfrac{10\sqrt{2}}{19} F \right], \\ \left\{ r^{(2)} : 90 \leq r^{(2)} \leq \dfrac{110(\sqrt{2}F - x_1)}{x_1} \right\}, & \text{if} \quad x_1 \in \left[\dfrac{10\sqrt{2}}{19} F, \dfrac{11\sqrt{2}}{20} F \right], \\ \emptyset & \text{otherwise} . \end{cases}$$

Consequently

$$p(\mathbf{x})=\begin{cases}\displaystyle\int_{\frac{90(\sqrt{2}F-x_1)}{x_1}}^{99}\frac{1}{20}\,dr^{(2)}=\frac{9}{20}\left(21-\frac{10\sqrt{2}F}{x_1}\right), & \text{if } x_1\in\left[\frac{10\sqrt{2}}{21}F,\ \frac{\sqrt{2}}{2}F\right],\\ \displaystyle\int_{90}^{99}\frac{1}{20}\,dr^{(2)}=\frac{9}{20}, & \text{if } x_1\in\left[\frac{\sqrt{2}}{2}F,\ \frac{10\sqrt{2}}{19}F\right],\\ \displaystyle\int_{90}^{\frac{110(\sqrt{2}F-x_1)}{x_1}}\frac{1}{20}\,dr^{(2)}=\frac{1}{2}\left(\frac{10\sqrt{2}F}{x_1}-19\right), & \text{if } x_1\in\left[\frac{10\sqrt{2}}{19}F,\ \frac{11\sqrt{2}}{20}F\right],\\ \emptyset & \text{otherwise.}\end{cases}$$

In this case

$$\mathscr{R}(\mathbf{x})=\left\{\mathbf{r}\in\Delta R: r^{(2)}D_2(\mathbf{x}),\quad r^{(1)}=r^{(2)},\quad r^{(4)}=\frac{\sqrt{2}F-x_1}{x_1}r^{(2)},\quad 90\le r^3\le 110\right\}.$$

Range T_3 can be divided into 3 disjunct parts designated

$$T_{31}=\left\{\mathbf{x}\in\mathbb{R}_+^2: x_2\le-\sqrt{2}x_1+F,\quad x_1\in\left(0,\frac{\sqrt{2}}{2}F\right)\right\},$$

$$T_{32}=\left\{\mathbf{x}\in\mathbb{R}_+^2: -\sqrt{2}x_1+F\le x_2\le-\frac{\sqrt{2}}{2}x_1+F,\quad x_1\in(0,\sqrt{2}F)\right\}$$

$$T_{33}=\left\{\mathbf{x}\in\mathbb{R}_+^2: x_2\ge-\frac{\sqrt{2}}{2}x_1+F\right\}.$$

Solution of the reference equation: $\bar{\mathbf{x}}\in T_{32}$. Let us investigate this range.

Introduce new variables

$$y_1=g_1r^{(2)},$$

$$y_2=g_2r^{(3)}.$$

If $x\in T_{32}$, then $g_1\le 0$, $g_2\le 0$, thus $y_1\le 0$, $y_2\le 0$.

In this case, $\mathscr{R}(x)$ is defined by the following inequalities:

$$h_1=110\,g_1\le y_1\le 90g_1=H_1, \tag{27}$$

$$h_2=110g_2\le y_2\le 90g_2=H_2, \tag{28}$$

$$81 \leq \frac{-\sqrt{2}}{x_1} y_1 - \frac{1}{x_1} y_2 \leq 99,$$

$$81 \leq \frac{1}{x_2} y_1 - \frac{\sqrt{2}}{x_2} y_2 \leq 99.$$

Of the latter two pairs of inequalities

$$-\frac{99\sqrt{2}}{2} x_1 - \frac{\sqrt{2}}{2} y_2 \leq y_1 \leq \frac{81\sqrt{2}}{2} x_1 - \frac{\sqrt{2}}{2} y_2,$$

$$81x_2 + \sqrt{2}y_2 \leq y_1 \leq 99 + \sqrt{2}y_2$$

define a rectangle C with apexes

$$C_1 = (-33\sqrt{2}x_1 + 33x_2, \quad -33x_1 - 33\sqrt{2}x_2),$$

$$C_2 = (-33\sqrt{2}x_1 + 27x_2, \quad -33x_1 - 27\sqrt{2}x_2),$$

$$C_3 = (-27\sqrt{2}x_1 + 27x_2, \quad -27x_1 - 27\sqrt{2}x_2),$$

$$C_4 = (-27\sqrt{2}x_1 + 33x_2, \quad -27x_1 - 33\sqrt{2}x_2).$$

In the positive range

$$\begin{aligned} &C_{41} \geq C_{11} \geq C_{31} \geq C_{21}, \quad \text{if} \quad x_2 \geq \sqrt{2}x_1, \\ &C_{41} \geq C_{31} \geq C_{11} \geq C_{21}, \quad \text{if} \quad x_2 \leq \sqrt{2}x_1, \end{aligned} \tag{29}$$

and

$$\begin{aligned} &C_{12} \leq C_{42} \leq C_{22} \leq C_{32}, \quad \text{if} \quad x_2 \geq \frac{\sqrt{2}}{2} x_1, \\ &C_{12} \leq C_{22} \leq C_{42} \leq C_{32}, \quad \text{if} \quad x_2 \leq \frac{\sqrt{2}}{2} x_1, \end{aligned} \tag{30}$$

furthermore

$$\begin{aligned} &C_{11} \geq 0 \quad \text{and} \quad C_{31} \geq 0, \quad \text{if} \quad x_2 \geq \sqrt{2}x_1, \\ &C_{11} \leq 0 \quad \text{and} \quad C_{31} \leq 0, \quad \text{if} \quad x_2 \leq \sqrt{2}x_1. \end{aligned} \tag{31}$$

The position of rectangles C and N is determined by the following relations:

$$h_1 \leq C_{21}, \quad \text{if} \quad x_2 \geq -\frac{77\sqrt{2}}{137} x_1 + \frac{110}{137} F,$$

$$h_1 \leq C_{11}, \quad \text{if} \quad x_2 \geq -\frac{7\sqrt{2}}{13} x_1 + \frac{10}{13} F,$$

$$
\begin{aligned}
&h_1 \le C_{31}, \quad \text{if} \quad x_2 \ge -\frac{83\sqrt{2}}{137}x_1 + \frac{110}{137}F,\\
&h_1 \le C_{41}, \quad \text{if} \quad x_2 \ge -\frac{83\sqrt{2}}{143}x_1 + \frac{10}{13}F,\\
&C_{41} \le H_1, \quad \text{if} \quad x_2 \le -\frac{21\sqrt{2}}{41}x_1 + \frac{30}{41}F,\\
&C_{31} \le H_1, \quad \text{if} \quad x_2 \le -\frac{7\sqrt{2}}{13}x_1 + \frac{10}{13}F,\\
&C_{11} \le H_1, \quad \text{if} \quad x_2 \le -\frac{19\sqrt{2}}{41}x_1 + \frac{30}{41}F,\\
&C_{21} \le H_1, \quad \text{if} \quad x_2 \le -\frac{19\sqrt{2}}{39}x_1 + \frac{10}{13}F,\\
&h_2 \le C_{12}, \quad \text{if} \quad x_2 \le -\frac{\sqrt{2}}{2}x_1 + \frac{10}{13}F,\\
&h_2 \le C_{42}, \quad \text{if} \quad x_2 \le -\frac{137\sqrt{2}}{286}x_1 + \frac{10}{13}F,\\
&h_2 \le C_{22}, \quad \text{if} \quad x_2 \le -\frac{143\sqrt{2}}{274}x_1 + \frac{110}{137}F,\\
&h_2 \le C_{32}, \quad \text{if} \quad x_2 \le -\frac{\sqrt{2}}{2}x_1 + \frac{110}{137}F,\\
&C_{32} \le H_2, \quad \text{if} \quad x_2 \ge -\frac{\sqrt{2}}{2}x_1 + \frac{10}{13}F,\\
&C_{22} \le H_2, \quad \text{if} \quad x_2 \ge -\frac{41\sqrt{2}}{78}x_1 + \frac{10}{13}F,\\
&C_{42} \le H_2, \quad \text{if} \quad x_2 \ge -\frac{39\sqrt{2}}{82}x_1 + \frac{30}{41}F,\\
&C_{12} \le H_2, \quad \text{if} \quad x_2 \ge -\frac{\sqrt{2}}{2}x_1 + \frac{30}{41}F.
\end{aligned}
\tag{32}
$$

It can be checked by calculation that the solution of the reference equation falls within a subrange, satisfying the above relations, where the following relationships apply to the relation of C to N:

$$C_{21} \leq h_1 \leq H_1 \leq 0 \leq C_{31} \leq C_{11} \leq C_{41}$$
$$h_2 \leq C_{12} \leq C_{42} \leq C_{22} \leq H_2 \leq C_{32} \leq 0\,, \tag{33}$$

and

$$81x_2 + \sqrt{2}H_2 = H_1\,,$$

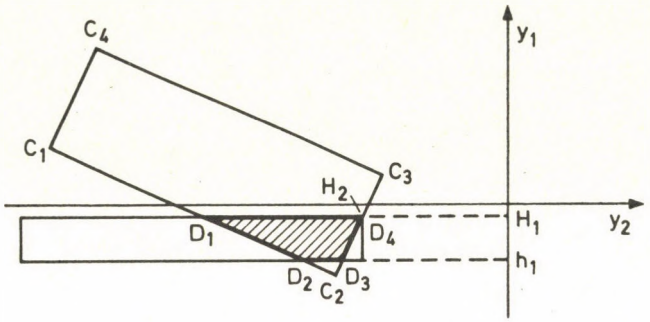


Fig. 6

and thus trapezium $D_1D_2D_3D_4$ is obtained from $D_2(x)$ by transformation (26), therefore

$$p(x) = \frac{1}{400g_1g_2} \int_{110g_1}^{90g_1} \int_{\frac{\sqrt{2}}{2}y_1}^{-\sqrt{2}y_1 - 99x_1} dy_2 dy_1 =$$

$$= \frac{300\sqrt{2}g_1 + 198x_1 - 81\sqrt{2}x_2}{40g_2} =$$

$$= \frac{-402x_1 - 381\sqrt{2}x_2 + 300\sqrt{2}F}{40(x_1 + \sqrt{2}x_2 - \sqrt{2}F)}\,.$$

Specifically for the solution of the reference equation with $F = 1$

$$p(\mathbf{x}) = 0.4342\ldots$$

$\mathscr{R}(\mathbf{x})$ under conditions (33):

$$\mathscr{R}(\mathbf{x}) = \{\mathbf{r} \in \Delta R : (r^{(2)}, r^{(3)}) \in D_2(\mathbf{x})\,,$$

$$r^{(1)} = -\frac{\sqrt{2}g_1}{x_1} r^{(2)} - \frac{g_2}{x_2} r^{(3)}\,,$$

$$r^{(4)} = \frac{g_1}{x_1} r^{(2)} - \frac{\sqrt{2}g_2}{x_2} r^{(3)}\Big\}$$

and

$$D_2(\mathbf{x})=\left\{(r^{(2)}, r^{(3)}): 90\leq r^{(2)}\leq 110\,,\quad \frac{\sqrt{2}}{2}\,g_1 r^{(2)}-\frac{81\sqrt{2}}{2}\,x_2\leq\right.$$

$$\left.\leq r^{(3)}\leq -\sqrt{2}g_1 r^{(2)}-99x_1\right\}. \tag{34}$$

Conditions (29) and (32) define additional subranges where $v(\mathbf{x})$ as a function of $\mathbf{x}$ can be calculated in each independently.

Using normal distribution instead of uniform distribution the complexity of calculations would have increased additionally because with no primitive function being available, $v(\mathbf{x})$ could have been determined only for the concrete $\mathbf{x}$ instead of for each range separately.

3) Fuzzy estimation

Let the common membership function $\mu(\mathbf{r})$ be given in the shape of

$$\mu(\mathbf{r})=\begin{cases}\prod_{i=1}^{4}\mu_i(r^{(i)}), & \text{if}\quad \mathbf{r}\in \Delta R\,,\\ \emptyset\end{cases}$$

where $\mu_i(r_i)$—normal density function normed to 1 with expectable value $\bar{r}^i$ and variation $\sigma_i=\gamma_i\bar{r}_i$:

$$\mu_i(r^{(i)})=e^{-\frac{(r^{(i)}-\bar{r}^{(i)})^2}{\sigma_i^2}}$$

$\mu(\mathbf{r})$ being maximum if quadratic function

$$\sum_{i=1}^{4}\frac{(r^{(i)}-\bar{r}^{(i)})^2}{\sigma_i^2}$$

is minimum. Here also like in the stochastic case, it is first of all the environment of the solution of the reference equation that we are interested in. Using (20) and transformation (26), the following problem shall be solved under conditions $(y_1\,, y_2)\in D_1D_2D_3D_4\,,$

$$\frac{\left(-\frac{\sqrt{2}}{x_1}y_1-\frac{1}{x_1}y_2-\bar{r}^{(1)}\right)^2}{\sigma_1^2}+\frac{\left(\frac{y_1}{g_1}-\bar{r}^{(2)}\right)^2}{\sigma_2^2}+\frac{\left(\frac{y_2}{g_2}-\bar{r}^{(3)}\right)^2}{\sigma_3^2}+$$

$$+\frac{\left(\frac{y_1}{x_2}-\frac{\sqrt{2}}{2}y_2-\bar{r}^{(4)}\right)^2}{\sigma_4^2}\to\min\,,$$

this is a simple quadratic programming problem for the solution of which quite a number of machine programmes are available. E.g. it can be solved by means of the projected conjugate gradient method described in [7].

References

1. Michelberger, P.,–Keresztes, A.: The estimation of stresses due to production inaccuracies by means of higher order moments. Acta Technica Hung., Tom. **79** (1–2). pp. 63–72 (1974).
2. Beermann, H. J.: Joint deformation and stresses of commercial vehicle frame under torsion. International Conference on Vehicle Structures. Proceedings C **178/84,** pp. 171–180. Crainfield (1984).
3. Rózsa, P.: Lineáris algebra és alkalmazásai. (Linear algebra and its applications.) Műszaki Könyvkiadó, Budapest (1974).
4. Gnedenko, B. V.: The theory of probability. Moscow, Mir (1973).
5. Zadeh, L. A.: Fuzzy sets. Inform. and Control, **8.** (1965). pp. 338–353.
6. Kovács, M.: Mathematical modelling under uncertainties (to be published)
7. Michelberger, P.,–Nádori, E.,–Kovács, M.: Design of vehicle structures taking nonlinearities into consideration. (to be published in Periodica Polytechnica).

Acta Technica Academiae Scientiarum Hungaricae, 98 (3—4), pp. 367—380 (1985)

USE OF MAXWELL BODY IN GAS PRESSURE MEASUREMENT BY MEANS OF CRUSHER

Z. KOVÁTS*

[Received: 3 May 1983]

After a brief survey of the history of gas pressure measurement, this paper describes a method where gas pressure is measured by means of crushers. A fundamental problem of this method is that statically calibrated crushers (copper cylinders) are used for measurement under dynamic conditions and therefore the measured values are lower than the actual values of pressure. To cope with this conflict, an empirical solution was found by Lamothe. Taking this as a basis, Sutterlin used the Maxwell body to model the copper cylinder, and obtained thus a theoretical solution resulting in fair agreement of the data so calculated with the results of both static and dynamic laboratory measurements. In this paper the results calculated on the basis of Sutterlin's theory are compared with the results of gas pressure measurements by piezoelectric methods.

1. Introduction

Gas pressure measurement by means of crusher (copper cylinder) is at present the most frequently used, and in some experts' opinion, most accurate method to measure the gas pressure developing during firing in the barrel of firearms. Firearms are essentially gas-operated machines where the high-pressure, high-temperature gases of the powder burning in explosion behind the piston eject the piston (bullet, shot column) from the barrel. The knowledge of the value of gas pressure is important also for strength calculations of the barrel and bullet and for calculation of the movement of the bullet in the barrel. Measurement of the gas pressure prevailing for 10^{-3} to 10^{-4} second only is not a simple problem indeed.

The value of gas pressure was determined indirectly by Russian general Majewski in 1867. He measured initial velocity v_0 [m/s] of the bullet of mass m [kg] by means of Le Boulengé's falling chronograph to obtain muzzle energy E_0 [J], and inferred the magnitude of gas pressure from relationship

$$E_0 = \frac{mv_0^2}{2} = A \int_0^s p \, dx, \tag{1.1}$$

where A cross-sectional area of barrel [m^2]
x bullet trajectory [m]
s length of barrel [m]
p gas pressure [Pa]

* Kováts, Z. H-1084 Budapest, József u. 26–28, Hungary

With function $p = p(x)$ unknown, information was obtained by the relationship on the average value of gas pressure only [1].

In direct methods, the effect of gas pressure (or more precisely, of thrust) resulting in motion or deformation was measured. The basic element in the so called knife method was a piston ending in an edge at the outer end, closely fitting while moving in a radial hole in the wall of the barrel. The edge of the piston ('knife') was intended as a result of gas pressure into the copper plate (Radman), bronze plate (Uchatius) or zinc plate (Schatzl) firmly fixed before the piston. By means of a materials testing machine, the magnitude of force required to obtain the indentation in the plate upon firing was then determined. In the knowledge of the cross-sectional area of the piston, the value of pressure could be calculated from this force.

However, the knife method was displaced by a measurement method invented by British captain Noble in about 1870. In principle, the method is similar to the knife method, but here both ends, thus also the outer end, of the piston moving in a radial hole (possibly extended) in the wall of the barrel are flat plates, the outer end bearing against a supported lead or copper cylinder, and the gas pressure is calculated from upset or compression of this cylinder. The process was called crusher after the English word. The crusher method displaced the knife method because it involved less uncertainty due to material defect as it measured compression throughout the entire copper cylinder. (In the knife method, the measurement results will be completely falsified if the knife cuts into a material defect in the plate.)

2. Gas pressure measurement by means of crusher

Gas pressure measurement by means of crusher can be used for any kind of firearms. For pressure measurements of small-calibre weapons and/or ammunitions, a separate measuring tube with screw-on crusher is required while in cannons, the crusher is placed into the cartridge case, under the powder (Fig. 1).

The procedure of measurement is the same in both methods. Gas pressure acts upon the face of a piston closely fitting while moving in its hole. The volume of the hole before the face is filled with a plastic mass in order to avoid interfering effects. The inner face of the piston within the fixture is surfaced and finished at right angles to the longitudinal axis i.e. to the direction of motion. The measuring element which is today a copper cylinder exclusively bears against this inner face. The copper cylinder is fixed in its place by the face of the cap screw (also surfaced and finished).

Compression of the copper cylinder during firing is measured to an accuracy of 0.01 mm, and the value of pressure associated with this compression is found in the so called tare table up in the course of calibration of the copper cylinders. (Of course, the tare table can be used only for given piston cross section.)

The copper cylinders are usually calibrated statically. In one of the calibration methods, the copper cylinder is supported, then the calibration body (gauge) of definite

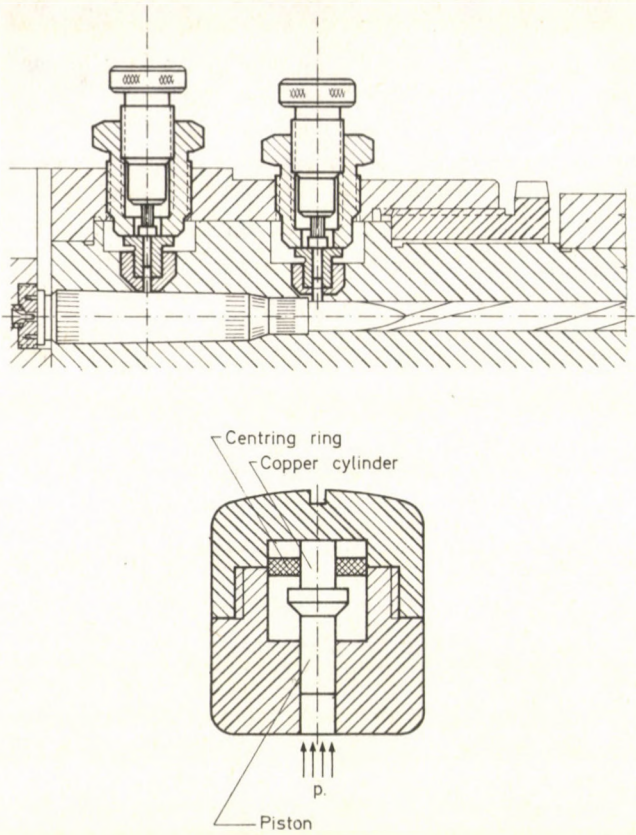


Fig. 1. Crusher measuring heads: a) screw-on; b) place-in

mass is quickly (within about 1 ms) applied to the cylinder and kept there for 30 s. A quick relief follows then. In another method, calibration takes place in the so called free-piston manometer. Here the calibration body is applied to the copper cylinder gradually, at a slow rate (within about 80 s). Other, dynamic calibration methods were also attempted without, however, finding wide use [2].

As has been said earlier, the important point here is that the magnitude of pressure is determined from tare tables in which the values tabulated have been obtained under static conditions and on the basis of compression of the copper cylinders as a result of dynamic impacts, a fundamental conflict. Double tare tables have therefore been set up by Polain, a Belgian [3]*. In Polain's method, the pressure is measured by the crusher located in the chamber statically while by crushers located in other points of the barrel dynamically. Thus in a tare table for the piston of a diameter

* In the numerical example given below, atm is used as the unit of measurement according to Polain's original double table.

of 6.08 mm (and of a surface of 29 mm^2) of copper cylinders of a height of 4.88 mm and a diameter of 3 mm, a static pressure of 837 atm and a dynamic pressure of 535 atm is specified for a compression of 0.96 mm while a static pressure of 535 atm and a dynamic pressure of 362 atm for an upset of 0.48 mm (1 atm = 0.980665 bar).

The conflict between static calibration and dynamic measurement can be resolved only by theoretical studies of the compression of crushers. The first such attempt was made in the 1920s [2] according to which a thrust of $F = A \cdot p$ acted upon the crusher, resulting in displacement (upset) x against resistance R of the copper cylinder. With the mass displaced designated m,

$$m\frac{d^2x}{dt^2} = F - R. \tag{2.1}$$

Although the equation seems to be simple, it involves different variables. E.g. the load is a function of time: $F(t) = Ap(t)$ because pressure changes in time. On the other hand, resistance R may be a function of time, upset, deformation rate, temperature, or of a combination of these factors. At that time, it was believed that

$$R = k_0 + kx, \tag{2.2}$$

where k_0 and k are constant. With this, a differential equation of the following shape is obtained:

$$m\frac{d^2x}{dt^2} = Ap(t) - (k_0 + kx). \tag{2.3}$$

In a static case, the rate of loading is very slow, approximately zero, and acceleration is negligible.

Total compression x_{max} is obtained from maximum pressure p_{max} in the following way:

$$x_{max} = \frac{Ap_{max} - k_0}{k}. \tag{2.4}$$

In a dynamic case, the rate of loading is rather high with maximum pressure occurring almost immediately. If the time of increase of the load is considered to be zero, then

$$m\frac{d^2x}{dt^2} = Ap_{max} - k_0 - kx \tag{2.5}$$

is obtained from (2.3).

By substitution of $\omega^2 = k/m$ and $x_0 = (Ap_{max} - k_0)/k$ we obtain the following differential equation:

$$\frac{d^2x}{dt^2} = \omega^2(x_0 - x). \tag{2.6}$$

Solution of this equation:

$$x_0 - x = x_0 \cos \omega t, \qquad \text{i.e.} \qquad x = x_0(1 - \cos \omega t). \tag{2.7}$$

A maximum of x is obtained at $\cos \omega t = -1$:

$$x_{max} = 2x_0 = 2 \frac{Ap_{max} - k_0}{k} \tag{2.8}$$

which is exactly twice as much as the value obtained statically (2.4) and thus it verifies Polain's table.

3. Lamothe's empirical solution [4]

The compression of copper cylinders of a height of 13 mm and a diameter of 8 mm serving as so called artillery crushers calibrated in free-piston manometer was measured as a function of load and rate of load (or rate of deformation) by French engineer and general Lamothe in the 1930s. Then an empirical equation was set up by Lamothe, supplying data that fairly agreed with the measured value.

The empirical formula shows (with Lamothe's symbols in it) that Lamothe accepted Volterra's theory on the so called hereditary phenomena, on the influence of the 'prehistory' of the material in that he assumed load F to be identical with resistance R of the copper cylinder (action = reaction) so that the equation took the shape of

$$F = R = f(e) - e\chi(t) \tag{3.1}$$

where e measured compression

$\chi(t)$ 'remembering' function, this latter given as

$$\chi(t) = \frac{at^{\alpha}}{b + t^{\alpha}} \tag{3.2}$$

where the values of $a = 0.188$, $b = 1.742$ and $\alpha = 0.25$ were calculated for the tested crusher on the basis of the results of measurement. It can be seen that if $t = 0$ then $\chi(t) = 0$ that is

$$R = f(e). \tag{3.3}$$

According to Lamothe, this case represents infinitely high deformation rate, the tendency in case of dynamic load. Static load tends towards infinitely low deformation rate. In this case that is if $t \to \infty$, then $\chi(t) \to a$ and thus the load will be

$$R = f(e) - ae. \tag{3.4}$$

It is worth mentioning that Lamothe studied also crushers compressed in advance. These copper cylinders are loaded statically to experience a deformation e_1, then in this deformed condition they are used for pressure measurement when they experience again a deformation designated e_2. According to Lamothe, if a load applied to the copper cylinder for time t_1 brought about a deformation e_1 of the copper cylinder, then in the knowledge of second deformation e_2, load R_2 acting upon the

same cylinders for time t_2 can be calculated, that is

$$R_2 = f(e_2) - e_1 \left[\frac{t_1 + t_2}{t_1} \chi(t_1 + t_2) - \frac{t_2}{t_1} \chi(t_2) \right] - (e_2 - e_1)\chi(t_2). \tag{3.5}$$

Had the copper cylinders been compressed in advance under static conditions indeed, then $t_2 \ll t_1$ since pressure measurement takes always place under dynamic conditions, that is the following approximating formula is obtained from equation (3.5):

$$R_2 = f(e_2) - e_2 \chi(t_2) \tag{3.6}$$

which is formally identical with equation (3.1). That means that according to Lamothe, the load for copper cylinders compressed or not compressed in advance can be determined in the same way in the knowledge of compression.

Lamothe's empirical formula can be expressed also in other form.

If

$$\chi(t) = \frac{K_2(t)}{t} \tag{3.7}$$

and

$$K_2 = \int_0^t K_1(t)\,\mathrm{d}t, \qquad K_2(0) = 0 \tag{3.8}$$

then, after substitution of the values into equation (3.1) and adopting independent variable $t - \tau$, the formula is obtained as

$$F = R = f(e) - \int_0^t K_1(t - \tau) \frac{\mathrm{d}e}{\mathrm{d}\tau}\,\mathrm{d}\tau. \tag{3.9}$$

4. Sutterlin's solution: generalized Maxwell body [5]

It was worth recalling Lamothe's empirical formula not only because of their historical curiosity but also because R. Sutterlin, French engineer and general, had obtained essentially the same result in a theoretical way. All the measurement results obtained after the publication of Lamothe's study were taken into consideration by Sutterlin, among others, Rougier's (1938) static experiments, the shooting experiments run in the same period in Versailles and by the Gâvre-committee, and the dynamic measurements of Habib (USA) started in 1943 and continued for several years, but even Charbonnier's experiments to measure warming-up of the crushers in 1900 quite forgotten since.

In Sutterlin's study from 1967, it is pointed out above all that the crusher has not become obsolete but it is still the best instrument to measure maximum gas pressure, first of all in cannons where piezoelectric measurements are difficult for different reasons.

Let a visco-elastic element n be connected in parallel with a spring of elasticity G_∞ (Fig. 2). Stress σ should act upon the body.

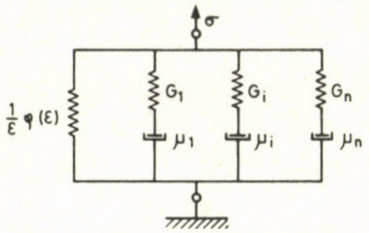


Fig. 2. Maxwell body containing visco-elastic elements of number n in addition to the element of variable elasticity

The value of σ:

$$\sigma = G_\infty \varepsilon + \sum_{i=1}^{i=n} G_i \varepsilon_i. \tag{4.1}$$

The time derivative:

$$\frac{d\sigma}{dt} = G_\infty \frac{d\varepsilon}{dt} + \sum_{i=1}^{i=n} G_i \frac{d\varepsilon_i}{dt}. \tag{4.2}$$

It may be written that

$$\sum_{i=1}^{i=n} G_i \varepsilon_i = \sum_{i=1}^{i=n} \mu_i \frac{d(\varepsilon - \varepsilon_i)}{dt}. \tag{4.3}$$

By removal of

$$\sum_{i=1}^{i=n} \varepsilon_i \qquad \text{and} \qquad \sum_{i=1}^{i=n} \frac{d\varepsilon_i}{dt}$$

from the above equation, differential equation

$$\sigma + \frac{d\sigma}{dt} \sum_{i=1}^{i=n} \frac{\mu_i}{G_i} = G_\infty \varepsilon + \frac{d\varepsilon}{dt} (G_\infty + \sum_{i=1}^{i=n} G_i) \sum_{i=1}^{i=n} \frac{\mu_i}{G_i} \tag{4.4}$$

and/or an equivalent integral equation

$$\sigma = G_\infty \varepsilon + \sum_{i=1}^{i=n} G_i \int_0^t \exp\left[- \frac{G_i}{\mu_i} (t - \tau) \right] \frac{d\varepsilon}{d\tau} d\tau \tag{4.5}$$

is obtained.

Assume that the spring of elasticity G_∞ is of variable elasticity with a modulus of elasticity of

$$\frac{\varphi(\varepsilon)}{\varepsilon}. \tag{4.6}$$

With this, equation (4.1) will take the shape of

$$\sigma = \varphi(\varepsilon) + \sum_{i=1}^{i=n} G_i \varepsilon_i \tag{4.7}$$

while the integral equation (4.5)

$$\sigma = \varphi(\varepsilon) + \sum_{i=1}^{i=n} G_i \int_0^t \exp\left[-\frac{G_i}{\mu_i}(t-\tau) \right] \frac{\mathrm{d}\varepsilon}{\mathrm{d}\tau} \mathrm{d}\tau , \tag{4.8}$$

or

$$\sigma = \varphi(\varepsilon) + \sum_{i=1}^{i=n} G_i \varepsilon - \sum_{i=1}^{i=n} G_i \int_0^t \left\{ 1 - \exp\left[-\frac{G_i}{\mu_i}(t-\tau) \right] \right\} \frac{\mathrm{d}\varepsilon}{\mathrm{d}\tau} \mathrm{d}\tau . \tag{4.9}$$

A comparison of this equation with (3.9), taking also (3.4) into consideration, shows that Eqs (4.9) and (3.9) are equivalent since

$$a = \sum_{i=1}^{i=n} G_i \tag{4.10}$$

and by proper selection of G_i and μ_i, function $K_1(t-\tau)$ can be made equal to a function of the following shape:

$$\sum_{i=1}^{i=n} G_i \left\{ 1 - \exp\left[-\frac{G_i}{\mu_i}(t-\tau) \right] \right\} .$$

Lamothe's symbols have been left unaltered by Sutterlin but the results he obtained on the basis of the Maxwell body have been marked with:

$$\chi^*(t) = \frac{K_2^*(t)}{t} ; \qquad K_2^*(0) = 0$$

where

$$K_2^*(t) = \int_0^t K_1^*(t) \, \mathrm{d}t .$$

Of this,

$$\sigma = f(\varepsilon) - \int_0^t K_1^*(t-\tau) \frac{\mathrm{d}\varepsilon}{\mathrm{d}\tau} \mathrm{d}\tau , \tag{4.11}$$

where

$$f(\varepsilon) = \varphi(\varepsilon) + \sum_{i=1}^{i=n} G_i \varepsilon$$

and function $K_1^*(t)$ consists of the sum of exponential terms.

The results of measurement taken into consideration by Lamothe are reproduced to an adequate accuracy by the following function:

$$K_1^*(t) \cong 20(1-e^{-10^4 t}) + 25(1-e^{-10^2 t}) + 55(1-e^{-t}) +$$
$$+ 50(1-e^{-10^{-2} t}) + 25(1-e^{-10^{-4} t}) + 13(1-e^{-10^{-6} t}).$$

The loads calculated on the basis of Eqs (3.1) and (4.11) suggest that the values of pressure read from the tare tables are lower than the actual values. Calculated values are always higher than measured values. As seen in both Lamothe's and Sutterlin's formula, the value of load highly depends on the rate of load.

5. Correction factors to increase the measured values of gas pressure

The difference between the values of gas pressure calculated on the basis of equations obtained from the results of experimental measurement and given in the tare table has been expressed in per cents of the tabulated values of pressure and diagrammatically illustrated by Sutterlin. Since the difference mentioned and thus the correction factor expressed in per cents are considerably affected by the rate of load, this fact had to be taken into consideration. In the diagram shown in Fig. 3, this is illustrated as the time of gas pressure build-up (load transfer) for four different cases. According to Sutterlin, the time of load transfer to the copper cylinder in the impact test is $t = 10^{-4}$ while build-up of maximum gas pressure takes a time of 3.10^{-4} s in small arms, 10^{-3} s in light guns, and 3.10^{-3} s in cannons.

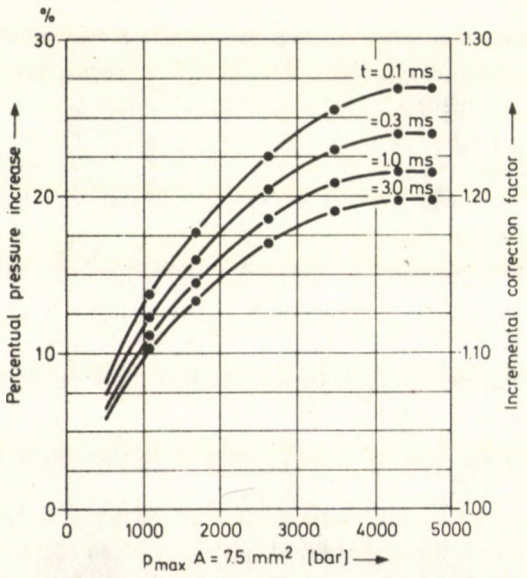


Fig. 3. Incremental correction factors for gas pressure measurement by means of a Ø 3 × 4.9 mm copper cylinder and a piston of a cross sectional area of 7.5 mm²

Note that the values of load calculated from the equations are modified by Sutterlin in accordance with the reduction in strength due to warm-up of the crusher. (The strength of copper warmed up reduces, the same deformation being brought about by smaller load due to the reduced strength). A local maximum of the curve of correction factor occurs therefore e.g. in case of a deformation of about 24% for guns (namely, the effect of warm-up will be considerable only in case of larger deformations).

6. Comparison of theory and results of measurements

There is a fair agreement between the calculated results of Sutterlin and the results of measurements obtained in static tare tests and dynamic impact tests, a matter of fact since the numerical values of the coefficients in the Maxwell body equation have been obtained on the basis of these results, the loading force and the rate of loading being known in both cases, while the compression of the copper cylinder measurable. Since the rate of deformation (rate of loading) during firing falls within the range between both values, Sutterlin believed that his theoretical results (and the numerical coefficients) could be interpolated also for actual gas pressure measurement, and obtained the correction factors in this way. To decide whether or not this idea is justified, the simplest way is to measure the gas pressure built up in the barrel by some other method less sensitive to the rate of loading like the use of an apparatus operating with piezoelectric quartz crystal.

The data given below are based on several hundreds of simultaneous gas pressure measurements made piezoelectrically or by means of crushers. Let us now study the 'correction factors' obtained in case the values of pressure measured piezoelectrically are considered 'actual pressure'. Described below are the conditions and results of measurements.

6.1 Ball cartridges

Measurements were made in the chamber, at a distance of about 25 mm from the bolt face. The values of pressure measured by means of a $\emptyset$ 4 × 6.5 mm crusher (the average of 10 firings each) fell in the range of 2735 to 2991 bar with correction factors of 21% to 21.5% at a pressure build-up time of $t = 3.10^{-4}$ s according to the diagram (Fig. 3). Pressure values between 3242 bar and 3603 bar were measured piezoelectrically. The 'correction factor' calculated as the quotient of the averages of 10 firings ranged between 18.6% and 22.3%.

These results show a fair agreement with Sutterlin's results, especially if we consider that the diagram was plotted for a $\emptyset$ 3 × 4.9 mm crusher while we used a $\emptyset$ 4 × 6.5 mm copper cylinder for the measurements (however, the copper cylinders were rather similar since 3/4.9 = 0.6122 and 4/6.5 = 0.6153). The pressure build-up time

ranged from $4.4 \cdot 10^{-4}$ s to $4.9 \cdot 10^{-4}$ s in the piezoelectric gas pressure measurements (this time was obtained by approximation as shown in Fig. 4).

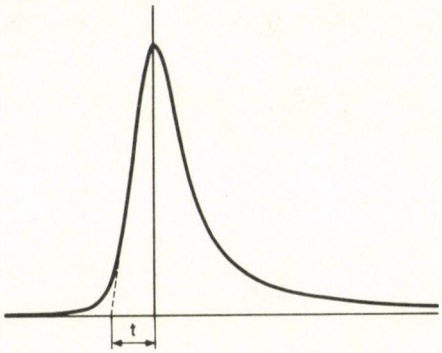


Fig. 4. Gas pressure curve showing the approximative determination of pressure build-up time (t)

6.2 Shotshells

Gas pressure was measured in the chamber also in this case, at a distance of 17.5 mm from the bolt face. A $\emptyset$ 3 × 4.9 mm crusher was used for the measurement, the same as in plotting the Sutterlin diagram, but with a piston of a cross sectional area of 30 mm^2 instead of 7.5 mm^2 because of the low pressure values. The average values of gas pressure for 12/70 calibre shotshells, measured by means of crusher, ranged from 508 bar to 530 bar, corresponding to 2032 bar and 2120 bar respectively, in case of a piston of a surface of 7.5 mm^2. Associated with these pressures as well as with the pressure build-up time of $t = 3 \cdot 10^{-4}$ on the basis of the diagram is a correction factor of about 18%. Piezoelectric measurements resulted in a pressure build-up time of $4.2 \cdot 10^{-4}$ and $4.9 \cdot 10^{-4}$ and in pressure of 558 to 588 bar, the value of correction factor ranging from 5.2% to 12.8%. These values lie well below the values read in the diagram.

For 16/70 calibre shotshells, pressures of 526 to 689 bar were measured by crusher, while 568 to 751 bar by piezoelectric pressure gauge. On this basis of the diagram, the associated values are 2104 bar and 2756 bar for pressure, while 18.6% and 20.7% for correction factor respectively. However, correction factors of only 7.7 to 13.7% are obtained for the pressures measured piezoelectrically. Hence, the difference is considerable in spite of the fact that the pressure build-up time ranged from $3.5 \cdot 10^{-4}$ s to $3.9 \cdot 10^{-4}$ s that is the conditions of measurement were more similar to those of the diagram than for the 12/70 calibre.

A value of 2000 bar has been specified for the upper measuring limit of the piezoelectric pressure gauge by the manufacturer. However, measurements were made also using a pressure gauge of a measuring limit of 8000 bar. In this case, the pressures measured by the piezoelectric system lay 31–33% above the values measured by the crusher, a value lying well above the values of the diagram.

6.3 Rim fire cartridges

In case of both ball cartridges and shotshells, load was applied to the measuring head through the holes of case mantle in gas pressure measurements both by crushers and piezoelectric gauges that is the pressure of powder gases was followed from the beginning of combustion by the measurement. At the same time, the gas pressure of rim fire cartridges was measured with the measuring head placed in front of the mouth. The measuring hole became free only after the bullet had been discharged, hence, a considerable pressure had been prevailing in the barrel at the beginning of measurement. Accordingly, the gas pressure curves showed a pressure build-up time of 3.10^{-5} s to 6.10^{-5} s that is, by an order of magnitude shorter than in the previous cases.

$\emptyset$ 3 × 4.9 mm crushers with a piston surface of 12 mm^2 were used to measure the gas pressure of calibre 22 long rifle cartridges. The measurements resulted in an average pressure of 1352 bar for which a correction factor of about 18% was given by the diagram (in case of $t = 3.10^{-4}$ s). The piezoelectric pressure gauge measured an average pressure of 1139 bar that is 84.2% of that measured by the crusher. That means that the 'incremental factor' became a 'decremental factor' surprisingly enough because according to Sutterlin's theory, the shorther the pressure build-up time that is the higher the rate of loading, the higher the value of the correction factor.

6.4 Evaluation of measurements

If the value measured piezoelectrically is considered to be the actual pressure, then

— satisfactory results are obtained by the 'remembering function' determined numerically for a $\emptyset$ 3 × 4.9 mm copper cylinder according to Sutterlin's theory in case of ball cartridge measurements,

— the measured values are lower than the calculated values in case of shotshell gas pressure measurements, while

— in case of measurement in front of the mouth, the values of pressure measured piezoelectrically lie below the values measured by crusher although the piezoelectric values should have been considerably higher according to the theory. (If Polain's double table were used where a 'static' pressure of 1352 bar corresponds to a 'dynamic' pressure of 855 bar, a correction factor of 33.2% would be obtained).

It is important that after compression, the copper cylinders became barrel-shaped with the maximum increase in diameter occurring at half length even in the last case where the pressure build-up time was of an order of magnitude of 10^{-5} (that is no local deformations were observed). This fact is all the more interesting because the copper cylinders experienced different longitudinal compression or deformation in the different measurements: 25.5% to 33.8% for ball cartridges, 9.4 to 12.9% (calibre 12/70) and 13.9 to 18.4% (calibre 16/70) for shotshells while 9.4 to 13.3% for rim fire cartridges.

Another important fact is that Sutterlin had plotted the correction factor curves for pressures measured by $\emptyset\ 3 \times 4.9$ mm crushers with a piston of a cross sectional area of 7.5 mm^2 and a mass of about 1 g, while we used copper cylinders of the same size but of a mass of about 3 g. Presumably, the mass of the piston moving during pressure measurement affects the deformation of copper cylinders.

7. Conclusions, problems

Sutterlin's theoretical approach has been confirmed by gas pressure measurements, and the numerical values of coefficients given by Sutterlin have been found correct, in case of ball cartridges (infantry ammunition). However, the same theory failed in case of pressure measurements before the mouth that is in case the measuring heads were exposed to surge pressure instead of a pressure built up gradually.

No definite opinion can be formed on the results of shotshell pressure measurements. Although the Maxwell body could possibly be used also here to model the copper cylinder but the numerical values of the coefficients are inadequate. (The assumption that the deviation resulted from the relatively small deformation of the copper cylinder is unacceptable, because a correction factor of 4.2% to 4.6% has been obtained in the measurement of gas pressure of high-pressure cartridges (so called force-testing cartridges) while the relative deformation of the copper cylinder ranged between 27% and 38%).

A significant difference between the measurements serving as a basis for theoretical approach and those made to confirm the theory lay in the mass of the piston. Hence, the influence of the mass of the piston and/or the ratio between the mass of piston and copper cylinder on the results of measurement shall be investigated.

Throughout the above investigations, the assumption that the piezoelectric gauge measures actual pressures has been taken as the starting point. However, this cannot be confirmed. Namely, it was found that measuring heads (and measuring circuits) of different layout measured different pressure values. The values measured piezoelectrically are also affected by the plastic material in the measuring hole. (To say nothing of the fact that an average pressure of 570 bar was measured by the piezoelectric gauge of the same type, with an upper measuring limit of 2000 bar, and 677 bar by the gauge of a measuring limit of 8000 bar, as compared with the crusher pressures of 508 bar and 509 bar respectively, because the latter measuring head was not recommended for pressures below 800 bar by the manufacturer.) However, it is important to mention that pressure gauges with a strain gauge glued on the barrel as the measuring element have also been used recently, and the values measured by means of these gauges differ from the pressures measured either piezoelectrically or by means of crushers. Hence, Sutterlin's theory can not be confirmed by direct measurements yet.

Similar problems are encountered in the measurement of pressure build-up time. Namely, the time can be read from the piezo-pressure curve only. The difficulty lies in

that a thin membrane displaces over a distance of 10^{-6} m before the quartz crystal (in our case), while a piston of a mass of sam grams over a distance of 10^{-4} m before the crusher. Here the question arises whether or not the crusher is 'delayed' as compared with the piezoelectric gauge.

In summing up, we might be right in saying that the use of the Maxwell body to model the crushers is a promising approach, but no reassuring answer is given by Sutterlin's described theory to all the questions arisen.

References

1. Massart: Erforschung des Druckes bei Feuerwaffen. Premier congrès international des Bancs d' Épreuves des Armes à feu. Goemaere, Bruxelles 1910
2. Ottenheimer, J.: Balistique intérieure. Librairie Armand Colin, Paris 1926
3. Commission Technique Internationale des Bancs d'Épreuves des Armes à feu. Compte-Rendu des Séances. Imprimerie Vaillant–Carmanne, Liège 1911
4. Lamothe, A.: Essai sur la théorie des crushers en cuivre. Mémorial de l'artillerie française, t. **XIV.** 1935. Paris, Imprimerie nationale
5. Sutterlin, R.: Sur l'application d'un modèle visco-elasto-plastique à la théorie des crushers. Mémorial de l'artillerie française, t. **41,** 1967. Paris, Imprimerie nationale

Acta Technica Academiae Scientiarum Hungaricae, 98 (3—4), pp. 381—386 (1985)

A SIMPLE AL–THIN SiO_2–pSi MIS SOLAR CELL

B. PŐDÖR*,**–K. O. OGUNKOYA**–V. A. WILLIAMS**

[Received: 29 November 1984]

First results concerning the fabrication and characterization of Al–thin SiO_2–pSi MIS solar cells are presented. Devices with an open–circuit voltage of 0.24–0.26 V, a short-circuit current of 3–4 mA/cm^2 and a power conversion efficiency of 1.5–2.0 per cent under nominal 20 mW/cm^2 irradiance have been fabricated.

1. Introduction

The reduction of solar cell fabriaction costs in one of the major goals of research and development in the area of photovoltaic energy conversion. The production of the most advanced cells reported so far, based either on Si (~17.5% AM1) [1, 2, 3] or on GaAs (~22% AM1) [4, 5] is still on a small laboratory scale due to the sophisticated processing required to achieve these high efficiencies. On the other hand Schottky barrier solar cells are suitable for large scale terrestrial photovoltaic conversion because they represent a potentially low-cost, low-temperature fabrication technology [6, 7]. The structure of Schottky barrier type solar cells makes possible the application of a homogeneous technological process, which can be implemented economically, and which can be fully automatized. A further potential advantage is their adaptability to polycrystalline materials.

If an ultrathin insulator (e.g. oxide) layer is sandwiched between the semiconductor and metal contact, the performance of the MIS solar cell is considerably enhanced over the corresponding MS structure [6, 8, 9, 10], while retaining the basic technological simplicity. The incorporation of a thin film interfacial oxide layer increases both the open–circuit voltage and the efficiency of the devices [9, 10]. E.g. MOS solar cells based on p–type Si with semitransparent barriers formed with Cr, Ti, and Al can exhibit open–circuit voltages in the range of 0.50–0.58 V [8], close to the theoretical limits achievable with pn junctions in Si [11]. According to the literature [8, 9] the open-circuit voltage in Si MOS cells reaches its maximum for SiO_2 thicknesses in the range of 1.3 to 2.0 nm. A recent comparative study of the fabrication, performance,

* B. Pődör, H-1131 Budapest, Jász u. 90A, Hungary. On leave of absence from the Research Laboratory for Inorganic Chemistry of the Hungarian Academy of Sciences, Budapest, Hungary.

** Department of Electronic and Electrical Engineering, University of Ife, Ile-Ife, Oyo State, Nigeria

and stability of Al– and Cr–MOS solar cells on p–Si substrates indicates that Al metallization is the better choice in many respects and leads to more stable devices [12].

The main technological advantages of MIS solar cells over the pn junction devices can be summarized as [13]

i) lower temperature of processing;

ii) collecting junction located right at the surface exposed to sunlight;

iii) applicability to polycrystalline materials.

The above discussed features of MIS solar cells make them a promising target for technological research and development for application in third world environment.

We have fabricated experimental Al–thin SiO_2–pSi solar cells, with the aim to develop a simple fabrication technology on the one hand, and to produce solar cells which can be used as detectors in monitoring solar irradiance at Ile–Ife, Oyo State, Nigeria, and at other similar places in the country where no systematic data for solar irradiance are yet available [14].

The study of solar irradiance at different places in Nigeria has been commenced only recently, see e.g. [15, 16]. Nigeria lies in the equatorial belt roughly between 4°30 and 13°50 latitude north. Using standard worldwide maps for daily means of total solar irradiation (beam and diffuse) incident on a horizontal surface [17], the daily means of solar radiation over Nigeria as a whole can be estimated as about 3.5 to 5.5 kWh/m^2. The number of average hours of sunshine shows a marked latitudinal variation not only in Nigeria but in the whole subregion [18], e.g. 8.8 h/day in Sokoto (13°02N, 5°16 E) and 6.3 h/day at Cotonou, Benin (6°20 N, 2°25 E), which in lieu of more detailed data can be taken as representative data for the northern and southern parts of the country.

In this paper we present the first results on the characterization and properties of the solar cells fabricated in our laboratory.

2. Solar Cell Fabrication

The solar cell structure consists of a p–type Si slice with a vacuum deposited Al ohmic contact layer on the back surface, a thermally grown SiO_2 layer on the front surface covered with a vacuum deposited semitransparent Al Schottky barrier layer. A thick Al contact finger completes the structure. The fabrication steps are detailed below.

p-type Si slices, (111) oriented, 5–20 ohmcm resistivity, 250 μm thickness are lapped on one surface and polished on the other.

After standard chemical cleaning the slices are immediately loaded into the vacuum chamber of the evaporator and a 50 to 100 nm Al layer is evaporated onto the lapped surface, (vacuum 3×10^{-8} Pa).

The back contact is heat treated at about 460 °C in room air for about 30 min. Simultaneously the top surface is annealed and an estimated 1 to 3 nm thin oxide layer is formed on it.

The slice is loaded again into the vacuum chamber of the evaporator and a 10 to 15 nm Al Schottky barrier layer is deposited onto the top surface (vacuum 3×10^{-8} Pa). Then finally a 100 nm thick contact layer is deposited through a mask.

The effective area of the cells is about 1 cm^2.

3. Characterization of the Solar Cells

The finished devices were characterized by measurement of I–V characteristics in dark and in artificial light. Besides these the light transmission vs Al layer thickness was also measured on Al layers deposited onto glass substrates simultaneously with the deposition of the barrier layer.

Typical dark I–V characteristics of the devices measured at 300 K are shown in Fig. 1.

The forward bias characteristics were analysed assuming non-ideal Schottky diode behaviour, taking into account also the effect of series resistance, R_s, as

$$J = A^* T^2 \exp(-e\Phi_B/kT)(\exp((eV - IR_s)/nkT) - 1)$$

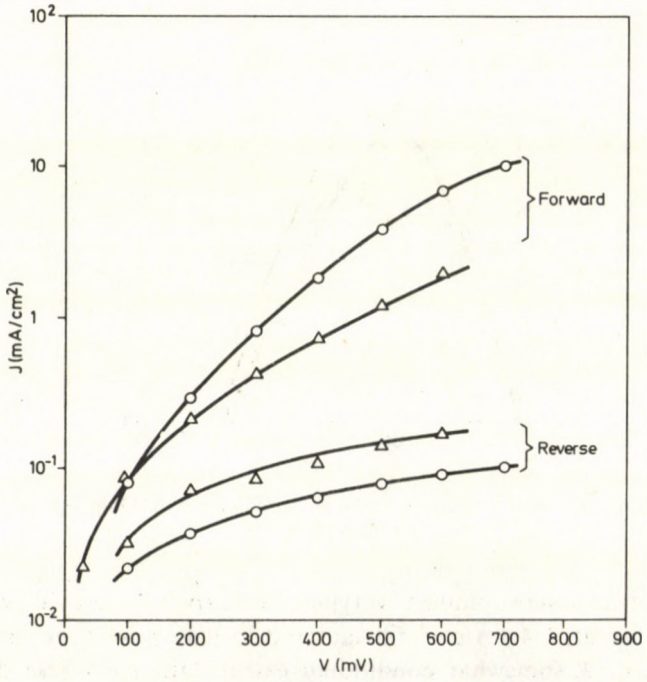


Fig. 1. Dark I–V characteristics of the devices. $J_0 = 5 \times 10^{-2}$ mA/cm^2, $n = 4.0$, $R_s = 15$ ohm cm^2, $\Delta - J_0 = 3 \times 10^{-2}$ mA/cm^2, $n = 5.2$, $R_s = 140$ ohm cm^2

to obtain the barrier height and the series resistance. In the analysis the Richardson constant was taken as $A^* = 120(m^*/m_0)$ A/cm^{2}K^2 with $m^*/m_0 = 0.55$ being the density–of–states effective mass of holes in Si. Barrier heights in the range of 0.65–0.69 eV were obtained, with diode nonideality factors, n, ranging from 4 to 6. These latter values correspond closely to those obtained on Pd–thin SiO_2–Si diodes with oxide thicknesses of 1.5 to 3.0 nm [19]. Making use of Norde's modified current–voltage analysis [20, 21] to evaluate the barrier height, we got 0.69 ± 0.02 eV, which agrees well with the results derived from the forward current extrapolated to zero voltage. Both methods of analysis resulted in a series resistance ranging from 20 to 80 ohms.

The I–V characteristics of the cells under illumination were measured using a nominal ~ 20 mW/cm^2 light input from a tungsten lamp. This illumination level corresponds to about 20 per cent of the standard AM1 irradiance level. The black body equivalent temperature of the radiation source was estimated as 2620 K using standard values of the resistance vs temperature dependence of tungsten.

Typical I–V characteristics of the devices under the above specified illumination are shown in Fig. 2. From the analysis of the I–V characteristics under illumination the

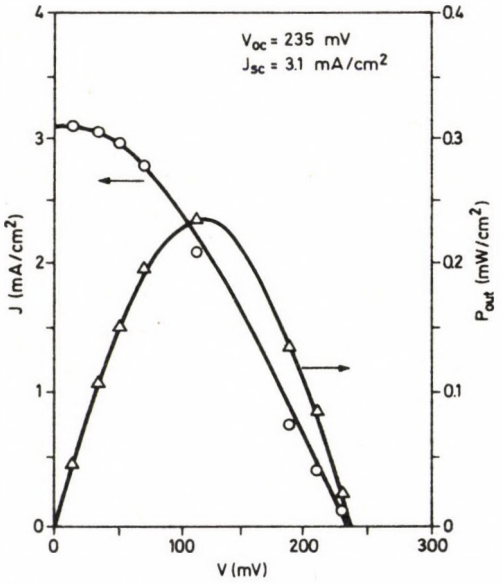


Fig. 2. I–V characteristics of solar cells under ~ 20 mW/cm^2 illumination. Output power vs voltage curve is also shown

following parameters were obtained for typical cells: open–circuit voltage: 0.24–0.26 V, short–circuit current: 3–4 mA/cm^2, fill–factor: 0.30–0.35, power conversion efficiency: 1.5–2.0 per cent. A somewhat conditional extrapolation of these data for AM1 conditions (100 mW/cm^2), which has been based on the proportionality of the short–circuit current to the irradiation level and on the logarithmic dependence of the open–

circuit voltage on the short–circuit current, neglecting the difference in the spectral distributions, would give the following values: open–circuit voltage: 0.30–0.35 V, short–circuit current: 14–18 mA/cm^2 [14].

4. Discussion

The extrapolated value of the open–circuit voltage is lower than the value of about 0.6 V given by MOS theory [9]. A comparison with similar measurements, [9], indicates that the average oxide thickness in our devices is about 1.0–1.2 nm, which is lower than the value of 1.6 to 2.0 nm necessary to reach the maximum value of the open–circuit voltage. The short–circuit current which is proportional to the light flux reaching the silicon material is chiefly limited by the reflection and absorption losses in the Al barrier layer. These losses according to our measurements on the Al layers deposited onto glass substrates can be as high as about 65 per cent in an Al layer of 12 nm thickness. A further limiting factor seems to be the excessive series resistance, which affects both the dark and illuminated I–V characteristics and also the fill–factor.

Nevertheless the parameters obtained on our cells (e.g. energy conversion efficiency) can be compared with those reported in the literature for similar structures (i.e. without antireflection coating) [22, 23]. According to the theoretical calculations of Pulfrey and McOuat [24], the conversion efficiency of Schottky barrier type solar cells depends strongly on the actual barrier height. Using their curves, the theoretical efficiency for Si based cells with a barrier height of about 0.7 eV is estimated as 3–4 per cent. The values measured on our cells compare not unfavourably with this theoretical calculations.

The data for the irradiance and mean sunshine hours in Nigeria referred to in the first part of this paper imply an average irradiation level of 50 to 60 mW/cm^2. Actually, recent measurements, [16], performed in the period from January to July 1982 at Birnin Kebbi (12°30 N, 4°20 E) recorded a maximum solar irradiance of about 90 mW/cm^2 and this normally occured between 12 noon and 1 p.m. Therefore performance data of solar cells referred to AM1 irradiation might serve the purpose of intercomparison, but for the purposes of field assessment a lower reference level might also be useful, which should lie somewhat inbetween the 20 mW/cm^2 used in this work and the 100 mW/cm^2 corresponding to AM1.

Work is in progress in our laboratory to develop further and to optimize the device structure.

References

1. Bae, M. S., D'Aiello, R. V.: Appl. Phys. Lett. **31** (1977), 285
2. Young, R. T., Wood, R. F., Christie, W. H.: J. Appl. Phys. **53** (1982), 1178
3. Young, R. T., Wood, R. F., Narayan, J., White, C. M., Christie, W. H.: IEEE Trans. Electr. Dev. **ED-27** (1980), 807

4. Woodall, J. M., Hovel, H. J.: Appl. Phys. Lett. **30** (1977), 492
5. Fan, J. C. C., Bozler, C. O., Chapman, R. L.: Appl. Phys. Lett. **32** (1978), 390
6. Singh, R., Shewchun, J., In: Sharing the Sun, Solar Technology in the Seventies, Conference Proceedings, Winnipeg, Canada, 1976, Vol. 6, Photovoltaics and Materials, p. 146
7. Godfrey, R. B., Green, M. A.: Appl. Phys. Lett. **34** (1979), 790
8. Kim, J. K., Anderson, W. A., Hyland, S.: IEEE Trans. Electr. Dev. **ED-26** (1979), 1777
9. Rajkanan, K., Singh, R., Shewchun, J.: IEEE Trans. Electr. Dev. **ED-27** (1980), 250
10. Srivastava, A. K., Guha, S., Arora, B. M.: Appl. Phys. Lett. **40** (1982), 43
11. Hovel, H. J., In: R. K. Willardson and A. C. Beer (eds.) Semiconductors and Semimetals, Vol. **11,** Solar Cells, Academic Press, N. Y. 1975
12. Dey, S. K., Tan, S. Y.: Proc. Conf. ENERGEX '82, August 23–29, 1982, Regina, Saskatchewan, Canada, Ed. F. A. Curtis, The Solar Energy Society of Canada, Vol. **1,** p. 243
13. Godfrey, R. B., Green, M. A.: Appl. Phys. Lett. **34** (1979), 790
14. Ogunkoya, K. O.: M. Sc. Thesis, Department of Electronic and Electrical Engineering, University of Ife, 1983
15. Adegboyega, G. A.: Proc. Conf. ENERGEX '82, August 23–29, 1982, Regina, Saskatchewan, Canada, Ed. F. A. Curtis, The Solar Energy Society of Canada, Vol. **1,** p. 77
16. Gulma, M. A., Bajpai, S. C.: Nigerian Journal of Science and Technology, **1** (1983), 11
17. Boes, E. C.: Fundamentals of Solar Radiation, In: Solar Energy Handbook, Ed. J. F. Kreider, F. Kreith, McGraw-Hill, 1981
18. Ojo, O.: The Climates of West Africa, London, Heinemann, 1977
19. Keramati, B., Zemel, J. N.: J. Appl. Phys. **53** (1982), 1091
20. Norde, H.: J. Appl. Phys. **50** (1979), 5052
21. Schwartz, G. P., Gualteri, G. J.: Appl. Phys. Lett. **42** (1983), 265
22. Green, M. A., King, F. D., Shewchun, J.: Solid State Electronics **17** (1974), 551
23. Anderson, V. A., Milano, R. A.: Proc. IEEE **63** (1975), 206
24. Pulfrey, D. L. McOuat, R. F.: Appl. Phys. Lett. **24** (1974), 167

Acta Technica Academiae Scientiarum Hungaricae, 98 (3—4), pp. 387—397 (1985)

ALGORITHM FOR DETERMINING THE NEAR OPTIMAL CENTRUM LOCATIONS IN LARGE GRAPH STRUCTURES

D. SINGER,* J. ELEK

[Received: 11 November 1984]

The paper gives an algorithm for the near optimal solution of large multi-centre problems by the simplification of the original problem. The method avoids nodes having a priori no chance to become a centre and in this manner diminishes the number of computational steps. The paper gives a description of the computer program and refers to experiments with typical graph structures.

1. Introduction

The multi-centrum problem of the graph theory can be formulated as follows: Find a location of M centers on the graph so that the distance (or more generally the weighted distance) required to reach the most remote point of the graph, from any of the centers, becomes a minimum. In the category of the M-center problems the optimal location of the feeding points of power networks, telephone switching centers, plant locations, public service stations etc. are included. The primary impulse for us is to deal with the problem and the task of locating the feeding point of a large municipal gas net.

The theoretical aspects of the multicenter location problem is treated in several papers and modern textbooks [1–4]. Some of these also give little or more convenient algorithms to solve practical problems. One of the recent contributions, that of Christofides and Viola reduces the problems to a set covering one and is computationally relatively effective for solving medium size problems [1]. As a whole, one can say about the computational methods worked out so far, is as follows: With problems of computational character the computing time generally rises very rapidly with the dimension of the problem.

The same is true for finding the optimal center location by increasing the number of centers. The upper limit which can be effectively solved with the known location methods are systems with about 150 nodes and 50 centers. The existing methods are, therefore, not appropriate to solve such common practical problems, as the optimal location of the feeding points of energy nets with some hundreds or thousands of nodes and some tenths or hundreds of feeding points.

To increase the capacity of the centrum location methods, the actual paper introduces the idea of replacing the original graph structure with its simplified version

* D. Singer, H-1021 Budapest, Nyéki u. 9, Hungary

guaranteeing a relatively good agreement with the exact solution. The simplified model is obtained in such a way that all nodes are a priori eliminated from the set of nodes having no opportunity to be a centre. From the remaining subset of nodes—which we will call in the followings "favorite" nodes—being generally smaller than the original set, the M centers can be found with one of the known methods or with a special method used in the following.

It must be emphasized that the centers obtained after the simplification can be, but must not be exactly the same as those of the original graph structure, therefore, in the following, we do not generally speak about centers, but about "dominant nodes". The saving of computer time using the method, can be very considerable, mainly by large problems. The little uncertainty of the results relative to the exact solution, is the tax to be paid for the augmented efficiency and for avoiding the danger of the combination explosion.

2. Main steps of the new algorithm

According to the foregoings, the method consists of two main parts, from the simplification of the original problem and from the procedure of finding the dominant nodes. The single steps of the algorithm are as follows:

— Finding the shortest path between all N nodes i, j of the graph (may be through to be intermediate nodes),

— Choosing the favorite nodes having the opportunity to be accepted as dominant nodes (centers). The favorite nodes will be obtained, ranging the nodes in ascending order, according to the sum of the path lengths (path weights) incident with the node and choosing from these the first $c \cdot M$ nodes. M is the number of centers, c is a constant.

— Determining the regions belonging to the $c \cdot M$ favorite nodes and calculating the weights G_i of these regions.

— The choice of the M dominant nodes from the $c \cdot M$ favorite ones. Each of the possible choices is characterized by the vector $\mathbf{G}$ of its G_i-values.

As the dominant node configuration is considered, the vector $\mathbf{G} = \hat{\mathbf{G}}$, where $\hat{\mathbf{G}}$ denotes the $\mathbf{G}$-vector with minimal relative mean square difference of the vector elements.

We give now a more detailed explanation to the single steps.

3. The shortest path between all nodes of weighted graphs

The most straightforward way of determining the shortest path between all nodes of the graph is to start with the node-to-node matrix $\mathbf{D}^0$ of the graph. $\mathbf{D}^0$ represents the shortest path of the graph for that case when intermediate nodes on the path are not allowed.

The element d_{ij}^0 of $\mathbf{D}^0$ is simply the length—or in the general case the weight—of the branch joining nodes i and j; Fig. 1.

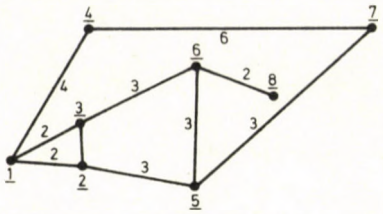


Fig. 1. Determining the shortest path matrix

From matrix $\mathbf{D}^0$ a matrix $\mathbf{D}^1$ can be derived whose elements d_{ij}^1 represent the path lengths with maximally one intermediate node. Similarly, $\mathbf{D}^2$ represents a path with maximally two intermediate nodes. One can continue deriving the matrices $\mathbf{D}^3$, $\mathbf{D}^4$ etc. until a full matrix $\mathbf{D}$ results having no zero elements. $\mathbf{D}$ is the shortest path matrix of the graph.

Matrix $\mathbf{D}^m$ can be derived from matrix $\mathbf{D}^{(m-1)}$ using the recursive algorithm of Floyd starting with the node-to-node matrix $\mathbf{D}^0$ [5]. $\mathbf{D}^0$ can be directly read from the graph. Contrary to common use, the zero entries of $\mathbf{D}^0$ must be changed to ∞. The Floyed algorithm has the following form:

$$d_{ij}^m = \min \{(d_{im}^{m-1} + d_{mj}^{m-1}), d_{ij}^{m-1}\}; \qquad m = 1, 2, \ldots, N. \tag{1}$$

For $m = 1$ and 2, one gets from (1)

$$d_{ij}^1 = \min \{(d_{i1}^0 + d_{1j}^0), d_{ij}^0\},$$

$$d_{ij}^2 = \min \{(d_{i2}^1 + d_{2j}^1), d_{ij}^1\}.$$

Example: The shortest path matrix $\mathbf{D}$ of the weighted graph in Fig. 1 is determined according to (1).

	i \ j	1	2	3	4	5	6	7	8		$d_i = \Sigma d_{ij}$	
	1	0	2	2	4	5	5	10	7		35	
	2	2	0	1	6	3	4	6	6		28	
	3	2	1	0	6	4	3	12	5		33	
D	4	4	6	6	0	9	9	6	9	→	49	(2)
	5	5	3	4	9	0	3	3	5		32	
	6	5	4	3	9	3	0	6	2		32	
	7	10	6	12	6	3	6	0	8		51	
	8	7	6	5	9	5	2	8	0		42	

4. Choice of the favorite nodes

The main idea on which the concept of the method is based, is to radically reduce the number of nodes from which the centrum nodes can be chosen. By choosing all nodes N for determining the M centers of the graph, one has totally:

$$\binom{N}{M} = \frac{N!}{(N-M)!M!} \tag{3}$$

possibilities. If one determines the centers only from $c \cdot M$ favorite nodes, one has only

$$\binom{cM}{M} = \frac{cM!}{M![(c-1)M]!} \tag{4}$$

possibilities; c is a small integer constant. It depends on the value of c how large the reduction of possible alternatives is. The number of combinations is radically diminished with diminishing c, as can be seen according to (3). The reduction constant c can be varied between the limits $1 < c < N/M$.

From a practical point of view, the reduction of c has some limits. By small values of c, one has lost the possibility to model the original graph with a simpler one with sufficient accuracy, because diminishing c means omitting some structural details. The choice of c must be, therefore, a result of compromise depending on the "regularity" of the graph. It is, therefore, not advisable to work with c-values < 2. Likewise increasing value c over 5, to increase the accuracy of the results is generally not advisable because the strong rise of machine time.

The favourite nodes can be obtained evaluating the weight sums of the rows in the all-path matrix **D**

$$d_i = \sum_{j=1}^{N} d_{ij} \qquad i = 1, 2, \ldots, N \tag{5}$$

and rearranging this according to ascending d_i values. The first $c \cdot M$ rows in **D** constitute its "favourite submatrix" $\mathbf{D}_F$. The appropriate nodes are the favourite ones.

Example: The graph in Fig 1 should have 3 centers; $N = 8$, $M = 3$. We chose for c the value $c = 2$. According to the last column of (2), containing the weight sums d_i, the favourite matrix $\mathbf{D}_F$ becomes

$$\mathbf{D}_F = \begin{array}{c|cccccccc} i \backslash j & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline 2 & 2 & 0 & 1 & 6 & 3 & 4 & 6 & 6 \\ 5 & 5 & 3 & 4 & 9 & 0 & 3 & 3 & 5 \\ 6 & 5 & 4 & 3 & 9 & 3 & 0 & 6 & 2 \\ 3 & 2 & 1 & 0 & 6 & 4 & 3 & 12 & 5 \\ 1 & 0 & 2 & 2 & 4 & 5 & 5 & 10 & 7 \\ 8 & 7 & 6 & 5 & 9 & 5 & 2 & 8 & 0 \end{array} \tag{6}$$

The favourite nodes are here:

$$i=\{2, 5, 6, 3, 1, 8\};$$

see vector d_i in (2).

5. Determining the dominant nodes

From the $c \cdot M$ favourite nodes the best M-member combination should be selected as "dominant" node. There exist totally $\binom{cM}{M}$ possibilities to select from the $c \cdot M$ favourite nodes the dominant ones. The best set of these must fulfil the following requirements.

To each of the M dominant nodes, there belong "satellite" nodes their distances (weights) from the dominant node in question is minimal relative to the remaining $(M-1)$ other dominant nodes. According to this, the "regions" of the dominant nodes can be defined. To each of the $\binom{c \cdot M}{M}$ alternatives to select M dominant nodes belong in this manner the same number of possible decomposition of the graph in the regions.

We *define* now M favourite nodes as dominants, if their regions are of near equal weight. The equality is meant here in a relative manner, relative to all other M-combinations of the $c \cdot M$ favourite nodes. Weight $\mathbf{G}_i$ of a region will be defined as the mean value of all inverse distances (weights) d_{ij}^{-1} of the region nodes from node i

$$\mathbf{G}_i=\frac{1}{N_i}\sum_j d_{ij}^{-1}; \qquad j=\alpha_1, \alpha_2, \ldots, \alpha N_i. \tag{7}$$

N_i is the number of nodes in the region.

Because the equality of the region weights has in our case no direct geometric sense, we measure it with the relative mean square difference of the appropriate region weights $\mathbf{G}_i$.

For the k-th alternative of the M centrum locations the relative mean square difference of the M region weights Δ_k is

$$\Delta_k=\sum_i\left[\frac{\mathbf{G}_i-\frac{1}{M}\sum_i G_i}{\frac{1}{M}\sum_i \mathbf{G}_i}\right]^2. \tag{8}$$

The Δ_k – s evaluated, according to (6) and (7), serve as criterions for the choice of the M dominant nodes from the $c \cdot M$ favourite ones. As the best of all alternatives, one choses that with the minimal Δ_k value.

Starting with the $\mathbf{D}_F$ matrix, the calculation process of the Δ_k values of the $\binom{c \cdot M}{M}$ alternatives can be organized as follows. For illustration purposes we show this on the foregoing example. According to (4), the number of alternatives for choosing $M=3$

dominant centers from $c \cdot M = 2.\ 3 = 6$ favourite ones are:

$$\frac{(2.3)!}{3![(2-1)3]!} = 20\,.$$

Each of these is characterized by a submatrix of $\mathbf{D}_F$. We denote these submatrices with the appropriate node indices i of the assumed (dominant) center nodes. For the alternative with the center nodes $i = 1, 2, 3$, the submatrices will be:

	i \ j	1	2	3	4	5	6	7	8	
$\mathbf{D}_F^{(1,2,3)}$	1	0	2	2	4	5	5	10	7	(9)
	2	2	0	1	6	3	4	6	6	
	3	2	1	0	6	4	3	12	5	
		1	2	3	1	2	3	2	2	

The first task is to determine to which center $i = \{1, 2, 3\}$ the satellite node j belongs. This can be done by finding the entries with the minimal values in each column j and notifying the appropriate i; this i values are notified in (9) under the double line.

One can see that the region of center $i = 1$ consists of nodes $j = 4$, that of $i = 2$ from $j = 5, 7, 8$ and that of $i = 3$ from $j = 6$.

The d_{ij}-s to calculate the weights $\mathbf{G}_i$ according to (6) can be read from (9)

$$\mathbf{G}_1 = \frac{1}{N_1} d_{14}^{-1} = \frac{1}{2}\frac{1}{4} = \frac{1}{8}\,,$$

$$\mathbf{G}_2 = \frac{1}{N_2}(d_{25}^{-1} + d_{27}^{-1} + d_{28}^{-1}) = \frac{1}{3}\left(\frac{1}{3} + \frac{1}{6} + \frac{1}{6}\right) = \frac{2}{9}\,,$$

$$\mathbf{G}_3 = \frac{1}{N_3} d_{36}^{-1} = \frac{1}{2}\left(\frac{1}{3}\right) = \frac{1}{6}\,.$$

The Δ_k relative mean square difference for the centrum combination $k = \{1, 2, 3\}$ is according to (8)

$$\Delta_{1,2,3} = 0.161\,.$$

In the same manner the Δ_k-s for the 19 other centrum combinations can be calculated.

6. Structure of the realized program

The flowsheet of the program (CENTRUM) working according to the described algorithm can be seen on Fig. 2. The scheme is self-evident and we add some remarks to the blocks only, where the purposes are not evident from the above considerations.

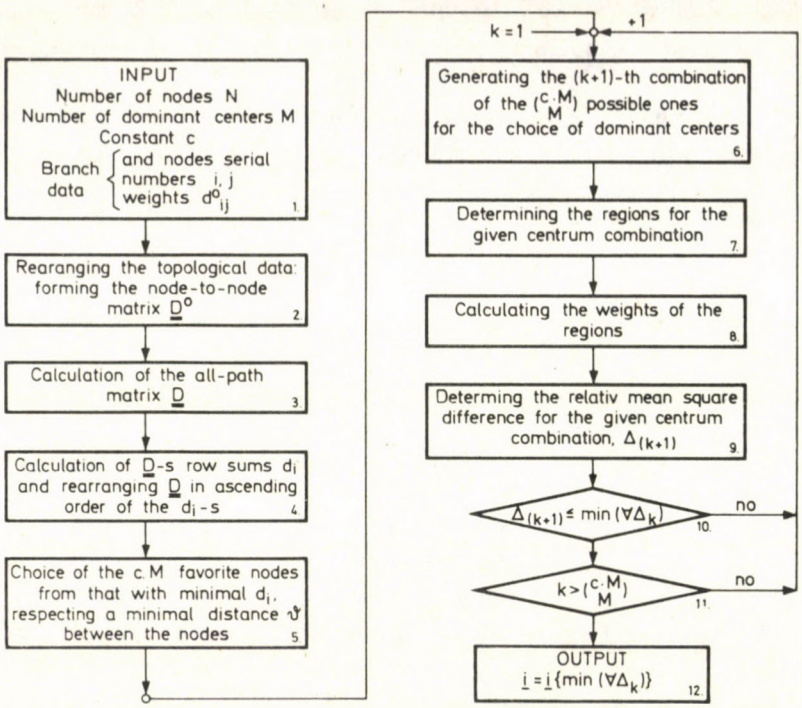


Fig. 2. The flow sheet of the CENTRUM program

The given graph structure can have a relatively high order of symmetry and many row sums of **D** can be equal or nearly equal. In such cases, it is advisable to renumber the nodes accidentally. This can be done by block 2. The renumbering of the nodes can also advisable be for other reasons. If one has some rough estimation concerning the location of the centers, one can diminish the value of the constant c and can diminish in this way the computing time too. One number in this case, first of all M presumed center nodes and the nodes in their vicinity. The remaining nodes will be, thereafter, numbered randomly.

The use of block 5 is optional and is motivated by the following: The program has to choose $c \cdot M$ favourite nodes from the N total number ones according to the ascending values d_{ij} of the **D**-row sums. It is advisable that the chosen d_{ij}-s should have minimal values. On the other hand, the subgraph with the $c \cdot M$ nodes should be a simplified model of the original graph and, therefore, it is not appropriate to chose the $c \cdot M$ favourite nodes with the same or near the same d_{ij} values. Block 5 serves to maintain a minimal distance δ between the d_{ij}-s of the subgraph nodes. This δ distance is maintained by block 5, solving the following inequality:

$$[\max(d_{ij}) - \min(d_{ij})] \geq c \cdot M\delta > \left[\frac{\max(d_{ij}) - \min(d_{ij})}{2}\right]. \tag{10}$$

The inequality is solved for δ by an iteration process. The meaning of (10) is evident: the sum of all $c \cdot M$ distances should not be larger than the difference of the max and min d_{ij} values.

On the other hand, the value of $c \cdot M\delta$—for the reason given below—should be almost as large as the half of this difference.

7. Conclusion

There exist principally three ways to determine the accuracy of heuristic algorithms:

— by comparison of the results with the results obtained with an exact algebraic method solving the same problem,
— by comparison of the results with the results obtained with the same algorithm using other starting values,
— using the algorithm for graph problems, where the location of the centers are evident according to the symmetry conditions.

In our investigations concerning the ability of the program CENTRUM, we used the last two possibilities. For demonstration purposes we will here show three centrum location tasks, where the accuracy of the results are evident from the symmetry of the topology and the edge-weight relations; see Figs 3, 4 and 5. Numbers with underlining denote node numbers, those without underlining edge weights.

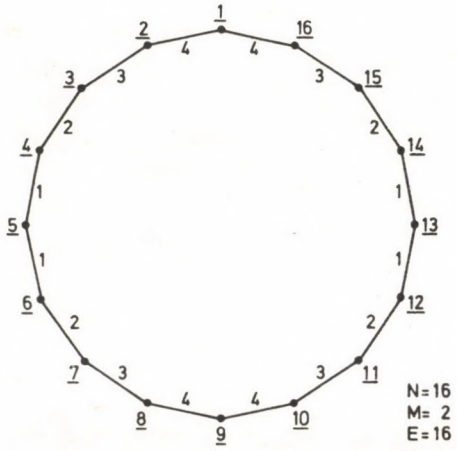


Fig. 3. Centrum location task

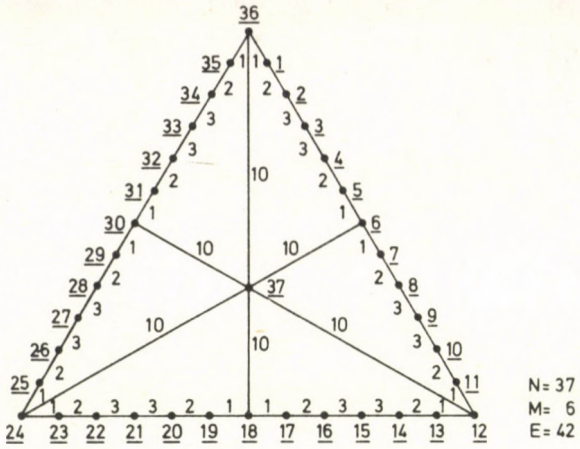


Fig. 4. Centrum location task

The data and results of the given sample tasks are summarized in Table I.

Table I

Task	No. of nodes N	No. of centers M	No. of edges E	c	Computed centrum locations, node numbers					
Fig. 3.	16	2	16	3	6	13				
Fig. 4.	37	6	42	4	6	12	18	24	30	36
Fig. 5.	68	4	131	3	1	34	51	68		

One can see that the computed centrum locations are, except Fig. 3 precisely the same as can be observed directly from the figures. The exact coincidence of the computed and true centrum locations is, however, not a rule, less or more large differences can be present.

According to the large number of computer experiments and in accordance with logical considerations, one can state the followings:

The method represents a relative efficient approximate solution of the multi-centrum problem for large graph structures, where the existing methods become—because of the combinatorial explosion—very ineffective.

The accuracy of the algorithm increases with increasing randomness of the topology and the edge weight distribution. The best the symmetry and the more little the dispersion of the edge weights, the less accurate are generally the results. The accuracy can be in all cases augmented by increasing the value of constant c, however, with the expense of a radical increase in computing time. The optimal working condition for the program is a compromise between the needed accuracy and the allowed computing time.

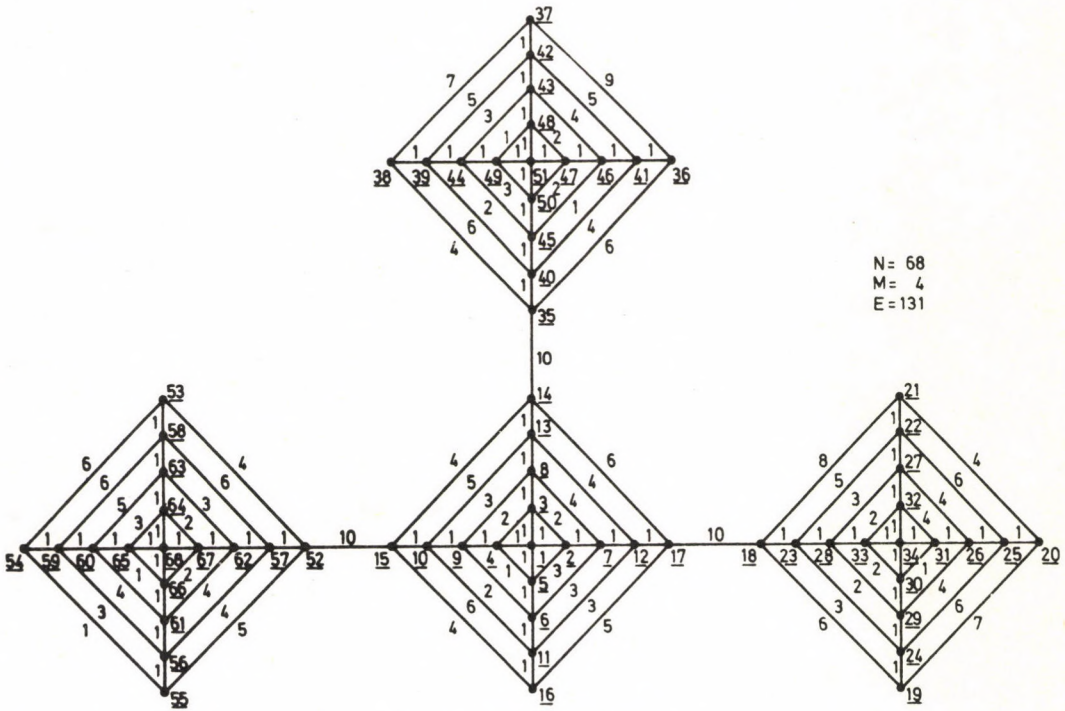


Fig. 5. Centrum location task

References

1. Christofides, M. and Viola, P.: Ops. Res. Quart. **22** (1971), 145–154
2. Goldman, A. J.: Transp. Sci. **4** (1969), 352–356
3. Hakimi, S. L.: ORSA **13** (1965), 462–475
4. Minieka, E.: Optimization Algorithms for Networks and Graphs Marcel Dekker, New York, (1978)
5. Floyd, R. W.: Comm. ACM **5** (1962), 345

Acta Technica Academiae Scientiarum Hungaricae, 98 (3—4), pp. 399—402 (1985)

BOOK REVIEWS

G. DALLOS–C. SZABÓ: *Random Access Methods of Telecommunication Channels* (in Hungarian)

One of the Hungarian Book publishing companies, the Academic Press has recently launched a new series of books under the heading: "Latest results in electronics". Using the well known photoprint method, the forthcoming issues in this series have a significantly shorter publishing time, than conventional editions. It is the intention of the Academic Press to publish books on special topics in electronics, either in Hungarian or in foreign languages such as English, Russian.

The authors start off with a survey of the random access methods used in telecommunications, the characteristics of these methods and those of the simple and slotted ALOHA channels. Before the various reservation schemes are tackled or thorough discussion follows on stability and control problems.

A special chapter is devoted to those access methods, which rely on carrier sensing and the possible procedures of resolving message collisions are also covered in detail. The authors give an introducing piece of theoretical analysis of an error correcting method developed for unslotted ALOHA channels. In conclusion some practical aspects of implementing terrestrial interactive terminal networks are dealt with.

This book is a very great help for project design engineers as well as for research personal involved in data communication problems.

P. Ferenczy

G. FRANZ (Schriftleiter): *Beton-Kalender 1985.* Taschenbuch für Beton-, Stahlbeton- und Spannbetonbau, sowie für die verwandten Fächer. Ernst und Sohn, Verlag für Architektur und technische Wissenschaften, Berlin 1985.

The manual — the 74the volume of the Beton-Kalender, — consists of two parts the first of which has 1030 and the second 1027 pages.

The first part of the manual contains, in accordance with many years' practice, the theoretical knowledge necessary for designing structures. The behaviour of the concrete (J. Bonzel), the different steel (D. Bertram) and the asbestos cement products (H. Pösch) are presented. It treats of the statics of bar structures (H. Duddeck–H. Ahrens), designing of r.c. units for particular loadings (E, Grasser), buckling safety of slender structures (K. Kordina–U. Quast). The last chapter presents the designing of r.c. units also including the case of partially prestressed structures (H. Kupfer).

Part 2 discusses the building regulations valid in the German Federal Republic (H. Goffin). A special chapter deals with different ways of supporting of structural units (K. Rahlwes), the problems of r.c. high-rise blocks (G. König–S. Liphardt). The application of prestressed concrete structures (H. Kupfer–H. Hochreiter), as well as the problems of different scaffoldings are also discussed (F. Nather).

The editor of the manual, Prof. G. Franz, the distinguished and generally esteemed expert of r.c. constructions entrusted the most outstanding persons with the writing of the different chapters. The authors resolved their task with due diligence, taking the newest results of science into account. A manual, as a matter of course cannot cover the full knowledge of a subject, nevertheless, it comprises all the information which a civil engineer need in practising his profession on the subject in question. The manual presents all the data in an easily intelligible manner accompanied by numerous demonstration figures. Many informative diagrams and valuable tables make the work of the designing engineer and the constructor easier.

Eventually, the newest volume of the Beton–Kalender is just as useful as the preceding valuable editions and, by the richness of its contents may increase the reputation of the editor and authors of this work.

P. Csonka

I. HAJNAL–J. MÁRTON–Z. REGELE: *Construction of Diaphragm Walls.* Publishing House of the Hungarian Academy of Sciences, Budapest 1984

Countless examples all over the world demonstrate that—provided correctly designed and applied—the diaphragm wall technology copes with requirements. Several publications have already been devoted to various problems of diaphragm

walls, but until recently, systematization, recapitulation and criticism were missing. This bock is expected to fill the gap, to further knowledge of experts of theory and practice in this subject, all over the world.

This book of 400 pages, complete with great many figures and a subject index, has been concerned with problems of theory, design, and practice-construction of diaphragm walls of keen actualness, world-wide applied in civil engineering. A valuable feature of the book is to embrace all of the three important fields of use of diaphragm walls (road an railway engineering, hydraulic structures, civil engineering foundations), proportionately, without overlapping, thanks to the complementarity of the Authors.

Introductorily, a survey is given—among others—on the history and phases of arise and development of the diaphragm wall construction method in the USA and in Europe. Chapter 2 classifies diaphragm wall structures, and illustrates the actual realizations in each field of application by rich graphic matter. Chapter 3 has been concerned with one of the most essential and most debated problems of diaphragm wall construction: the theory of fluid support; beside theoretical fundamentals, it describes the most up-to-date design methods taken either from the international technical literature or from research results of the Authors, and illustrates them on numerical examples.

Chapters 4 and 5 are spent on practical construction. Internationally applied cutting machinery, auxiliary equipment, tools and implements are presented in particulars, followed by that of the main phases of technology, with due consideration to supporting slurry composition and mixing, as well as to knowledge in the trench fill matter. Another important feature is description of the cutting operation itself, including detail problems such as repair of unavoidable construction defects.

Chapter 6 handles design: from preliminary soil mechanical tests, through selection of the structure, to various method of dimensioning. Methods are presented for determining bottom resistance and mantle friction—two important components of load capacity—illustrated on numerical examples. Finally, Chapter 7 expounds labour safety and environmental aspects. In Annex, computer programs are given to determine trench stability and load capacity, keeping demands of present-day engineers in mind.

Each chapter is concluded with a detailed list of references.

L. Rétháti

M. MAJOR: *Geschichte der Architektur* (History of Architecture) Vol. 3. Publishing House of the Hungarian Academy of Sciences, 1984, p. 606

The third volume of the new version of this huge work, surveying the universal history of architecture, is concerned with the history of recent architecture. The subject has been treated in two parts, a shorter one concerning the 19th century architecture, the foundation laid by capitalist society, and a longer one separately treating 20th century architecture of capitalist and socialist societies. Parts are introduced with the historical survey of the period, followed by listing its achievements in cultural history, involving not only attendant arts creations but literary and philosophical works as well. What is more, by outlining the development and achievements of technical, biological and other disciplines, the architecture was integrated into its own age. Development of the building technology itself emerges in the analysis of works of architecture, and so does urbanology considered as a synthesis of architecture.

When M. Major wrote the first Hungarian version of this fundamental work nearly thirty years ago, there was little comprehensive work on the history of architecture concerned with the creations of modern times, also relationships were controlled by subjective aspects. The pioneering work "Space, Time and Architecture" by Giedion started out from his connection to Le Corbusier, offering it in an entertaining, in a lively way all that belongs to the development of modern architecture, without taking note of the main body of historicism, rather typical of European urbanization late in the 19th century. Of course, Hitchcook's work centered on British conditions. Benevolo offered a real, wide-range analysis, even mentioning the Hungarian CIRPAC. In his work adjusted to track development in consecutive editions, Joedicke wrote the most concise survey of the history of modern architecture. The work by M. Major essentially differed and differs from them, all even in this edition, by projecting architecture on a social background, and fitting all typical achievements to that age. Thereby architecture is not self-contained any more but integer to a social and cultural development.

A special merit of this book is that he devotes an extensive chapter to each to the architectures periods of the USSR and Hungary in the part on socialist architecture. A suggestive description is given of the avantgardism of the young Soviet state, on the way to historicism in the '30s. It fades after the '60s and seems a bit short-cut, but since the whole

volume ends with this time, it is understandable and acceptable. Representative illustrations have been selected from recent Hungarian architecture.

It is difficult to assign data to extend to these days of such a mighty work, and to track the variable trends of architectural conception, so it is understandable that—in conformity with its subtitle—the book lasts till the mid-'20th century. And certainly, the survey of modern architecture becomes increasingly difficult now when—rather than the scattered creations emerging from a sea of historicism—entire towns are made up typically of modern architecture, becoming, in turn, grey their mass or even with some trends. The volume ends at this, fanning out of modern architecture, so to say, it tracks booming of modern architecture up to its zenith. Understandably, no mention is made of individual trends, proliferating since the '60s.

M. Kubinszky

A., JOAN: *Cavitatia*, Vol. I., Editura Acadamiei, Bucuresti 1984, 337 p.

"Cavitation is a most unpleasant hydrodynamic phenomenon, the harmful effects of which are both widespread and obvious and seriously handicaps many phases of science and engineering. Conversely, its basic nature has long been veiled in mystery and only recently is it beginning to be understood" *(Knapp)*.

The phenomenon was observed long ago. The hissing of water flowing through a constricted tube was observed by Reynolds in 1873 and attributed to an internal boiling of water under diminished pressure. Cavitation on the back of the blades of a propeller was first observed in 1894 by Thornycroft and Barnaby. It is not surprising to find that, for a long period after its identification, rather divergent views were expressed concerning the physical nature of cavitation. This divergency appeared in the theories of the hydrodynamic process and reached a maximum in the various attempts to concoct plausible descriptions of the process through which cavitation produces damage on solid boundaries.

Over the past half century an extensive bibliography on cavitation has developed. Practically all treatises are discussions of isolated of the phenomena. This is to be explained by the complexity of the manifestation. The published few books only cover some fields of research. Such books are e.g. Pernik: Problemi kavitacii (Leningrad 1966) which summarizes the theoretical literature of bubble dynamics, scale effects and some basic hydrodynamics. Karelin (Moscow 1963) presented a book on the cavitation problems of centrifugal pumps. Noskievič: Kavitace (Praha 1969) discusses on bubble dynamics, the types of cavitation, erosion of solid surfaces and cavitation in the rotodynamic machines (turbines, pumps, propellers).

The first book covering the subject as a whole, a modern treatise in English is Knapp–Daily–Hammitt: Cavitation (McGraw–Hill 1970). The presentation of this volume covers four general topic areas such as. 1) The basic characteristics and physical mechanics of hydrodynamic cavitation. 2) Cavitation damage from the viewpoint of both hydrodynamic process and the reaction of particular materials. 3) Cavitation study methods and euipment for research and making tests. 4) Cavitation effects in flow passages and hydraulic equipment and on fixed and free bodies. Each topics is the subject of particular parts of the book, because there is much interrelation between the areas.

The multiplicity of the research prevented the elaboration of all branches of the investigations. For instance the detection of cavitation by acoustic methods when the research in the subject began one decade earlier (Rata: Bruit de Cavitation (1960), Cormault (1962), Huguenin (1964), Pearsall (1966) etc.). Therefore, 14 years after the publication this valuable book, it seems desirable to have a new and complete summary of this research field.

It is to be hoped that this requirement will be realized by the books (Vol. I and II) of professor Anton. The contents of Vol. I are: Cavitation nucleation and inception, stressing of liquids, dynamics of the cavitation bubble, cavitation coefficients, standard cavitating bodies, similitude at cavitation, scale effects, destruction processes of solid materials through their impact with raindrops and liquid jets, mechanism of the cavitation induced breakdown of the materials. The well selected contents, the excellent treatment of the matter, moreover the outlined contents in the preface of the second volume suggests that a comprehensive treatise is being created from the complete territory of investigations.

These expectations are supported by the circumstance that voluminous review of this kind can be performed only by a scientist who himself is a successful researcher in the field of cavitation. All these are generally known from professor and academician Anton.

In the interest of completeness of the work it is necessary to mention, that in Vol. I there are only short references from the acoustic methods of detecting the incipient cavitation. It is desirable to complete this subject, mentioned earlier as a lack in the book of Knapp et alii.

It is to be hoped that the large-scale work of professor Anton will promote the cavitation research activity and it is therefore desirable to publish these volumes in English. Looking forward to the publication of the second volumes with great anticipation.

J. J. Varga

NOTICE TO CONTRIBUTORS

Papers in English* are accepted to the condition that they have not been previously published or accepted for publication.

Manuscripts in two copies (the original type-written copy plus a clear duplicate one) complete with figures, tables, and references should be sent to the

Acta Technica
Münnich F. u. 7. I. 111A
Budapest, Hungary
H-1051

Although every effort will be made to guard against loss, it is advised that authors retain copies of all material which they submit. The editorial board reserves the right to make editorial changes.

Manuscripts should be typed double-spaced on one side of good quality paper with proper margins and bear the title of the paper and the name(s) of the author(s). The full postal address(es) of the author(s) should be given in a footnote on the first page. An abstract of 50 to 100 words should precede the text of the paper. The paper should not exceed 25 pages including tables and references. The approximate locations of the tables and figures should be indicated on the margin. An additional copy of the abstract is needed. Russian words and names should be transliterated into English.

References. Only papers closely related to the author's work should be referred to. The citations should include the name of the author and/or the reference number in brackets. A list of numbered references should follow the end of the manuscript.

References to periodicals should mention: (1) name(s) and initials of the author(s); (2) title of the paper; (3) name of the periodical; (4) volume; (5) year of publication in parentheses; (6) number of the first page. Thus: 5. Winokur, A., Gluck, J.: Ultimate strength analysis of coupled shear walls. American Concrete Institute Journal 65 (1968), 1029.

References to books should include: (1) author(s) name; (2) title; (3) publisher; (4) place and year of publication. Thus: Timoshenko, S., Gere, J.: Theory of Elastic Stability. McGraw-Hill Company, New York, London 1961.

Illustrations should be selected carefully and only up to the necessary quantity. Black-and-white photographs should be in the form of glossy prints. The author's name and the title of the paper together with the serial number of the figure should be written on the back of each print. Legends should be brief and attached on a separate sheet. Tables, each bearing a title, should be self-explanatory and numbered consecutively.

Authors will receive proofs must be sent back by return mail.

Authors are entitled to 50 reprints free of charge.

* Hungarian authors should submit their papers also in Hungarian.

Periodicals of the Hungarian Academy of Sciences are obtainable at the following addresses:

AUSTRALIA
C.B.D. LIBRARY AND SUBSCRIPTION SERVICE
Box 4886, G.P.O., *Sydney N.S.W. 2001*
COSMOS BOOKSHOP, 145 Ackland Street
St. Kilda (Melbourne), Victiria 3182

AUSTRIA
GLOBUS,Höchstädtplatz 3, *1206 Wien XX*

BELGIUM
OFFICE INTERNATIONAL DE LIBRAIRIE
30 A venue Marnix, *1050 Bruxelles*
LIBRAIRIE DU MONDE ENTIER
162 rue du Mindi, *1000 Bruxelles*

BULGARIA
HEMUS, Bulvar Ruszki 6, *Sofia*

CANADA
PANNONIA BOOKS, P.O. Box 1017
Postal Station "B", *Toronto, Ontario M5T 2T8*

CHINA
CNPICOR, Periodical Department, P.O. Box 50
Peking

CZECHOSLOVAKIA
MAD'ARSKÁ KULTURA, Národní třida 22
115. 66 Praha
PNS DOVOZ TISKU, Vinohradská 46, *Praha 2*
PNS DOVOZ TLAČE, *Bratislava 2*

DENMARK
EJNAR MUNKSGAARD, Norregade 6
1165 Copenhagen K

FEDERAL REPUBLIC OF GERMANY
KUNST UND WISSEN ERICH BIEBER
Postfach 46, *7000 Stuttgart 1*

FINLAND
AKATEEMINEN KIRJAKAUPPA, P.O. Box 128 *SF-00101*
Helsinki 10

FRANCE
DAWSON-FRANCE S. A., P. 40, *91121 Palaiseau*
EUROPÉRIODIQUES S. A., 31 Avenue de Versailles, *78170 La Celle St. Cloud*
OFFICE INTERNATIONAL DOCUMENTION ET
LIBRAIRIE,48 rue Gay-Lussac
75240 Paris Cedex 05

GERMAN DEMOCRATIC REPUBLIC
HAUS DER UNGARISCHEN KULTUR
Karl Liebknecht-Straße 9, *DDR-102 Berlin*
DEUTSCHE POST ZEITUNGSVERTRIEBSAMT Straße der
Pariser Kommüne 3 4, *DDR-104 Berlin*

GREAT BRITAIN
BLACKWELL'S PERIODICALS DIVISION
Hythe Bridge Street, Oxford OX1 2ET
BUMPUS, HALDANE AND MAXWELL LTD.
Cowper Works, *olney, Bucks MK46 4BN*
COLLET'S HOLDINGS LTD., Denington Estate *Wellingborough, Northants NN8 2QT*
WM. DAWSON AND SONS LTD., Cannon House *Folkstote, Kent CT19 5EE*
H. K. LEWIS AND CO., 136 Gower Street
London WCIE 6BS

GREECE
KOSTARAKIS BROTHERS INTERNATIONAL
BOOKSELLERS, 2 Hippokratous Street, *Athens-143*

HOLLAND
MEULENHOFF-BRUNA B. V., Beulingstraat 2,
Amsterdam
MARTINUS NIJHOFF B.V.
Lange Voorhout 9 11, *Den Haag*
SWETS SUBSCRIPTION SERVICE
347b Heereweg, *Lisse*

INDIA
ALLIED PUBLISHING PRIVATE LTD., 13/14
Asaf Ali Road,*New Delhi 110001*
150 B-6 Monunt Road, *Madras 600002*
INTERNATIONAL BOOK HOUSE PVT. LTD.
Madame Cama Road, *Bombay 400039*
THE STATE TRADING CORPORATION OF INDIA LTD.,
Books Import Division, Chanralok 36 Janpath, *New Delhi 110001*

ITALY
INTERSCIENTIA, Via Mazzé 28, *10149 Torino*
LIBRERIA COMMISSIONARIA SANSONI, Via Lamarmora 45,
50121 Firenze
SANTO VANASIA, Via M. Macchi 58
20124 Milano
D. E. A., Via Lima 28, *00198 Roma*

JAPAN
KINOKUNIYA BOOK-STORE CO. LTD.
17-7 Shinjuku 3 chome, Shinjuku-ku, *Tokyo 106-91*
MARUZEN COMPANY LTD., Book Department, P.O. Box
5050 Tokyo International, *Tokyo 100-31*
NAKUA LTD. IMPORT DEPARTMENT
2-30-19 Minami Ikebukuro, Toshima-ku, *Tokyo 171*

KOREA
CHULPANMUL, *Phenjan*

NORWAY
TANUM-TIDSKRIFT-SENTRALEN A.S., Karl Johansgatan
41 43, 1000 Oslo

POLAND
WEGIERSKI INSTYTUT KULTURY, Marszalkowska 80,
00-517 warsawa
CKP-I W, ul. Towarowa 28,*00-958 Warsawa*

ROUMANIA
D.E.P., *Bucuresti*
ILEXIM,Calea Grivitei 64-66, *Bucuresti*

SOVIET UNION
SOJUZPECHAT IMPORT, *Moscow*
and the post offices each town
MEZHDUNARODNAYA KNIGA, *Moscow G-200*

SPAIN
DIAZ DE SANTOS, Lagasca 95, *Madrid 6*

SWEDEN
GUMPERTS UNIVERSITETSBOKHANDEL AB
Box 346, *40125 Göteborg 1*

SWITTZERLAND
KARGER LIBRI AG, Petersgraben 31, *4011 Basel*

USA
EBSCO SUBSCRIPTION SERVICES
P.O. Box 1943, *Birmingham, Alabama 35201*
F.W. FAXON COMPANY, INC.
15 Soutwest Park, *Westwood Mass. 02090*
READ-MORE PUBLICATIONS, INC.
140 Cedar Street, *New York,N.Y. 10006*

YUGOSLAVIA
JUGOSLOVENSKA KNJIGA,Terazije 27, *Beograd*
FORUM,Vojvode Mišića 1, *21000 Novi Sad*

HU ISSN 0001-7035 Index: 26.025