Erhard Godehardt

Graphs as Structural Models

Advances in System Analysis
Editor: Dietmar P. F. Möller

Volume 1: Emil S. Bücherl (Ed.)
Proceedings of the Second World Symposium
Artificial Heart

Volume 2: Dietmar P. F. Möller (Ed.)
System Analysis of Biological Processes

Volume 3: Kiichi Tsuchiya and Mitsuo Umezu
Mechanical Simulator of the Cardiovascular System:
Design, Development and Application

Volume 4: Erhard Godehardt
Graphs as Structural Models

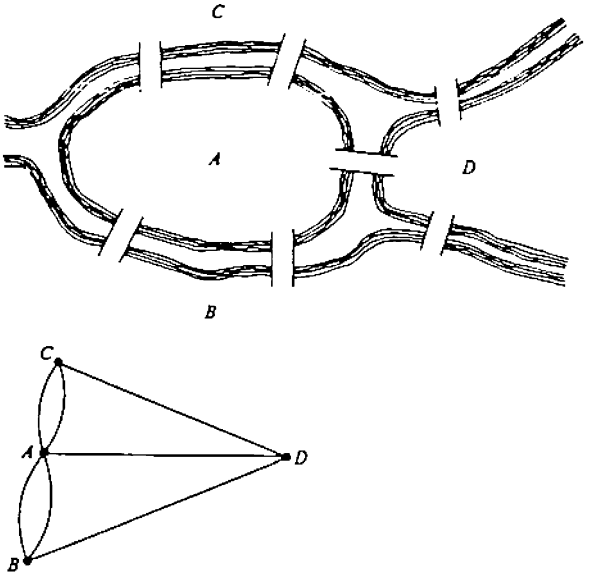
Manuscripts submitted to *Advances in System Analysis* must be original, pointing out the advancement of the contribution with respect to the actual a-priori knowledge.

Manuscripts or exposés should be sent to the Editor of the Series:
Dietmar P. F. Möller, Johannes Gutenberg Universität Mainz, Physiologisches Institut, Saarstr. 21, D-6500 Mainz 1, W.-Germany.

Erhard Godehardt

Graphs as Structural Models

The Application of Graphs and Multigraphs in Cluster Analysis



Springer Fachmedien Wiesbaden GmbH
Braunschweig / Wiesbaden

CIP-Titelaufnahme der Deutschen Bibliothek

Godehardt, Erhard:
Graphs as structural models: the application
of graphs and multigraphs in cluster analysis/
Erhard Godehardt. – Braunschweig; Wiesbaden:
Vieweg, 1988
(Advances in system analysis; Vol. 4)
ISBN 978-3-528-06312-2 ISBN 978-3-322-96310-9 (eBook)
DOI 10.1007/978-3-322-96310-9

NE: GT

Vieweg is a subsidiary company of the Bertelsmann Publishing Group.

ISBN 978-3-528-06312-2

ISSN 0932-593X

PREFACE

The advent of the high-speed computer with its enormous storage capabilities enabled statisticians as well as researchers from the different topics of life sciences to apply multivariate statistical procedures to large data sets to explore their structures. More and more, methods of graphical representation and data analysis are used for investigations. These methods belong to a topic of growing popularity, known as "exploratory data analysis" or EDA.

In many applications, there is reason to believe that a set of objects can be clustered into subgroups that differ in meaningful ways. Extensive data sets, for example, are stored in clinical cancer registers. In large data sets like these, nobody would expect the objects to be homogeneous. The most commonly used terms for the class of procedures that seek to separate the component data into groups are "cluster analysis" or "numerical taxonomy". The origins of cluster analysis can be found in biology and anthropology at the beginning of the century. The first systematic investigations in cluster analysis are those of K. Pearson in 1894. The search for classifications or typologies of objects or persons, however, is indigenous not only to biology but to a wide variety of disciplines. Thus, in recent years, a growing interest in classification and related areas has taken place. Today, we see applications of cluster analysis not only to biology but also to such diverse areas as psychology, regional analysis, marketing research, chemistry, archaeology and medicine.

These applications indicate not only the importance of the existing classification tools. They also stress the need for further development and investigation into classification theory as a mathematical topic. This progress in the development of mathematical procedures and models is not only stimulated indirectly by the advances in electronic data processing (since lots of data are waiting to be evaluated), but also directly inspired directly by the improvements in electronic computers which allow thorough study and simulation of more complex models. The spectrum of mathematical models for classification reaches from multivariate analysis to the theory of graphs and random graphs.

The roots of graph theory are obscure. The famous eighteenth-century Swiss mathematician Leonard Euler was perhaps the first to solve a problem using graphs when he was asked to consider the problem of the Königsberg bridges (in the 1730s). Problems in (finite) graph theory are often enumeration problems, and thus can become rather intricate to solve. However, in the late 1950s and early 1960s the Hungarian mathematicians Paul Erdős and Alfred Rényi founded the theory of random graphs and used

probabilistic methods (limit theorems) to by-pass enumeration problems. (These problems also became secondary with the emergence of powerful computers.) Thus, perhaps no topic in mathematics has enjoyed such explosive growth in recent years as graph theory. This stepchild of combinatorics and topology has emerged as a fascinating topic for research in its own right. Moreover, during the last two decades, calculus of graph theory has proved to be a valuable tool in applied mathematics and life sciences as well. Using graph-theoretic concepts, scientists study properties of real systems by modelling and simulation. The aim of graph-theoretic investigations is, in fact, the simplest topological structure after that of isolated points: The structure of a graph is that of "points" or "vertices", and "edges" or "lines". A graph can be conveniently pictured as a diagram where the vertices appear as small circular dots and the edges are indicated with line segments joining two appropriate dots (or arcs instead of lines if the direction is of relevance).

These graphs can be used to model many real systems. This usually works if the vertices can be identified as physical objects, and the edges can be identified as relations or equations between the objects. In fact, every binary structural relation can be described by a graph. Thus it was a natural consequence to use graphs in classification theory; here, the similarity between the objects defines the relation. Structures in data sets can also be depicted very simply and impressively by dendrograms, a special kind of graph. The advantage of using graphs when describing and modelling structures is the fact that the analyst can rely on a plethora of mathematical theorems and results to gain insight into the structure of a real system. Graph theory provides researchers of different topics with a single (mathematical) language. Thus, it promotes the exchange of ideas to a degree which probably would be impossible if the scientists relied only on the technical terms of their own subject. This is reflected in the fact that a lot of theoretical results in graph theory were not found by pure mathematicians but by scientists working in the field of applied mathematics. Like every mathematical model, graph theory is only useful for applications so long as it describes our scientific observations with a certain amount of accuracy. Its immediate popularity, however, shows that it provides the researcher with a lot of good ideas — at least within the limitations of our actual scientific knowledge.

In many topics in the natural and life sciences, the interpretation of quantitative results from data is tied to mathematical ideas and models. This, however, implies an obligation for the true biometrician: In contrast to pure mathematics and statistics, mathematical concepts in data analysis not only have to be "true" on a theoretical basis, but also must fit into different environments like physiology and biology. And to be applicable, they must stay as simple as possible. This is impossible without constant discussions between the applied mathematicians and the researchers from life sciences. The classification model proposed in this book has been discussed with many physicians and biologists, among others with Professor K.-E. Richard (Department of Neural Surgery of the University of Köln), Priv.-Doz. Dr. H. Borberg (Department of Internal Medicine I of the University of Köln), Priv.-Doz. Dr. H. Feltkamp (Department of Medical Research of Bayer Pharmaceutics, Wuppertal) and Dr. E. Rehse (Institute of Medical Documentation and Statistics of the University of Köln).

The idea of applying results from the theory of random multigraphs rose in 1980 from my dissertation in pure mathematics on probability models for random multigraphs (a generalization of the concepts of P. Erdős and A. Rényi). I am grateful to Professor H. Klinger from the Institute of Mathematical Statistics and Documen-

tation of the University of Düsseldorf, where I wrote this thesis. He and Professor J. Steinebach (now at the Statistical Institute of the University of Hannover) not only discussed the mathematical proofs with me, but, together with Professors J. Krauth (Psychological Institute of the University of Düsseldorf) and O. Richter (now at the Division of Statistics of the Agricultural Institute of the University of Bonn) encouraged me to apply my theoretical results in practice. I also wish to thank Professors V. Weidtman and P. Bauer from the Institute of Medical Documentation and Statistics of the University of Köln where I wrote my habilitation thesis in biometrics on the application of graph-theoretic models in exploratory statistics, and where I was employed until 1986 after I moved from Düsseldorf to Köln. We had very interesting and stimulating discussions on multivariate data analysis and on the problems of the application of graph-theoretic models and concepts to medicine.

It is a great pleasure to acknowledge the generous help of Professors D. Matula (Department of Computer Science, Southern Methodist University, Dallas) and J. Wierman (Mathematical Sciences Department, Johns Hopkins University, Baltimore). I am especially greatful to Professors J.W. Kennedy and L.V. Quintas (Dyson College, Pace University New York) and to Professor M. Karoński, Dr. Z. Palka and Dr. J. Jaworski (Mathematical Institute of the University of Poznań), who not only gave valuable hints for my work, but also invited me for research work to New York and to Poznań. (In Poznań, 1987, the work on this book started: I began to translate my habilitation thesis into English.)

In this monograph, I review the principles and properties of numerical classification. The different chapters deal with its possibilities as well as with its limitations. I emphasize the application of graph-theoretic concepts as tools in the structural analysis of data sets which are composed of "mixed data" as they often occur in medicine. I further propose a graph-theoretic classification model which I developed on the basis of multigraphs.

Chapter 0 deals with the mathematical symbols and notations which are necessary for the understanding of this book. In Chapter 1, a short review of the elementary ideas of mathematical modelling, graph theory, and exploratory statistics are given. They are followed by a short description of the basic ideas of cluster analysis. In Chapter 2, I give a survey of the current methods and different algorithms of cluster analysis. Chapter 3 deals with graph-theoretic concepts in classification theory, especially with the single-linkage and k-linkage procedures. An algorithm for the uncovering of clusters is proposed which has been developed on the basis of my multigraph model.

Up to this point, judging the "relevance" of a classification of a data set is not considered. The following two chapters are devoted to probabilistic models for evaluating the validity of the suggested clusters. In Chapter 4, investigations in the development of different statistical tests are reported. Again, I focus on graph-theoretic based probability models, which are discussed in detail. I propose a conditional statistical test for the homogeneity of a data set which is derived from random multigraphs. In Chapter 5, a probability model for random multigraphs is derived. This part is written as an independent section in this book. It also can be read as a mathematical theory on its own right. The main test principles reported in the previous chapter, however, are founded upon this theory.

The last chapter then deals with three examples, all from medicine. The first two examples are the pharmacokinetics of Urapidil and Lidocaine. In these medical trials, our interest is focussed on studying whether differences in the kinetics of these drugs can

be explained by certain external criteria, namely by the occurrence of additional kidney or liver impairments. That is, we want to find whether the a priori groups (according to diagnoses) are reflected by the a posteriori groups (according to the kinetic data). The third example is from the habilitation thesis of Priv.-Doz. Dr. H. Feltkamp. It is part of a long-term study of the significance of pregnancy-induced hypertensive disorders as a prognostic index for a manifestation of an essential hypertension lateron.

It was intended originally, to include only the data from the studies on the pharmacokinetics of Urapidil and Lidocaine as purely numerical examples of the application of the multigraph method. However, since the interpretation of cluster analyses makes little sense without a knowledge of the scientific background to the data, it was decided to include a much fuller exposition of these studies. More details of the compartmental models — together with some modifications — are given in my habilitation thesis. These models are the result of the discussions with Dr. R. Haerlin and Dr. V. Steinijans (Byk Gulden Lomberg Chemicals, Konstanz). The complete data can be found in the doctorial dissertations of R. Dworatzek and W. Heitz (both in the Department of Internal Medicine II of the University of Köln).

The idea of developing interactive programs for numerical classification on the Prime computer of the Institute of Medical Documentation and Statistics would not have been realized without the enormous work of J. Kunert and H. Herrmann. The program package is composed of two independent parts. Part 1 is used to find clusters (it is described in Chapter 3). With the second part, the hypothesis of homogeneity of a data set, i.e., the hypothesis of randomness of clusters, can be tested using finite or asymptotic test statistics from the theory of random multigraphs (see Chapter 5). The algorithms for the finite tests are based on our doctorial dissertation. At the present, H. Herrmann, J. Kunert and I are re-writing these programs at the Department of Thoracic and Cardiovascular Surgery of the University of Düsseldorf, where I have been employed since summer 1986, and at the Institute of Medical Documentation and Statistics — and at home, of course — so that they will be user-friendly and will run on Prime computers and personal computers.

I want to thank all those persons who encouraged and supported me in writing this book. H. Herrmann showed me how to use the TEX scientific text system and wrote most of the macros for the preparation of the manuscript. Professor J. Wierman kindly assisted my first attempts in translating my thesis in Posnań, and Professor A.D. Barbour from the Institute of Applied Mathematics of the University of Zürich immediately agreed to read the final manuscript and made suggestions to improve it.

Finally, I must thank the editor of this series of monographs, Dr. D.P.F. Möller, and the Vieweg Verlag for their patience and encouragement. I also thank Byk Gulden Lomberg Chemicals, Konstanz, and Prime Computers, Wiesbaden, for their financial support.

Neuss, Spring 1988 Erhard Godehardt

TABLE OF CONTENTS