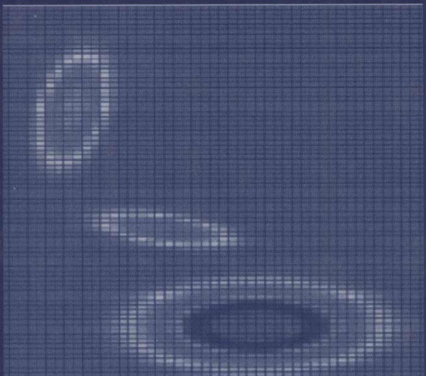
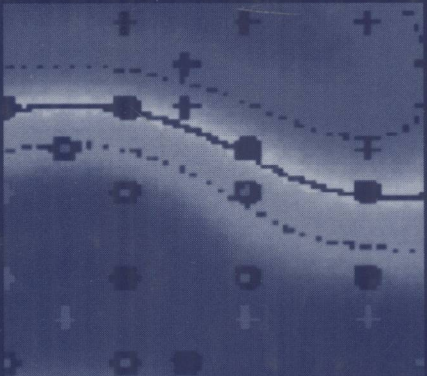
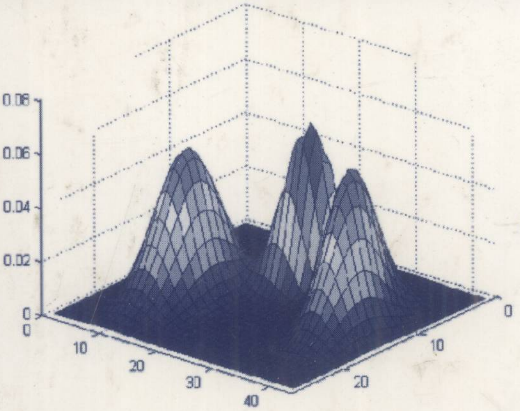
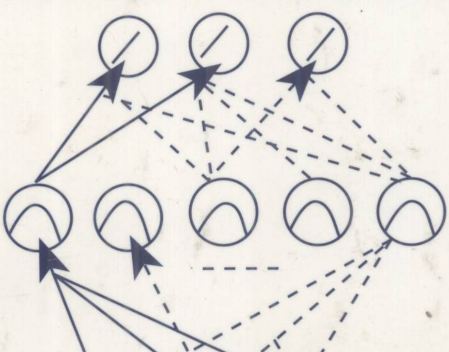
Statistical Data Mining and Knowledge Discovery









Edited by
Hamparsum Bozdogan

CHAPMAN & HALL/CRC

Statistical Data Mining and Knowledge Discovery

Edited by
Hamparsum Bozdogan

CHAPMAN & HALL/CRC

A CRC Press Company
Boca Raton London New York Washington, D.C.

Library of Congress Cataloging-in-Publication Data

Statistical data mining and knowledge discovery / edited by Hamparsum Bozdogan.
p. cm.
Includes bibliographical references and index.
ISBN 1-58488-344-8
1. Data mining—Statistical methods. 2. Computer algorithms. 3. Knowledge acquisition (Expert systems) I. Bozdogan, H. (Hamparsum), 1945-

QA76.9.D343S685 2003
006.3′07′27—dc21 2003051462

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

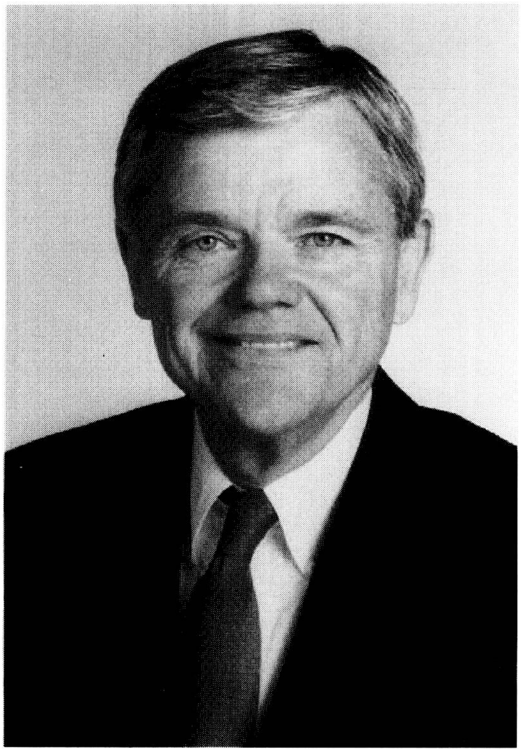
Direct all inquiries to CRC Press LLC, 2000 N.W. Corporate Blvd., Boca Raton, Florida 33431.

Visit the CRC Press Web site at www.crcpress.com

International Standard Book Number 1-58488-344-8
Library of Congress Card Number 2003051462
Printed in the United States of America 1 2 3 4 5 6 7 8 9 0
Printed on acid-free paper

Statistical Data Mining and Knowledge Discovery

DEDICATION



Dean of the College of Business Administration
1977 – 2001
The University of Tennessee
Knoxville, TN

DEDICATED TO C. WARREN NEEL FOR HIS OUTSTANDING SERVICE FOR OVER TWO DECADES TO THE COLLEGE OF BUSINESS ADMINISTRATION AND THE UNIVERSITY OF TENNESSEE

FOREWORD

Even before there was a name for it I had been interested in pattern recognition. My first recollection was meeting people who built model aircraft during WW II. The models were used to train observers to recognize differences in the shape of an enemy plane compared to the planes of allies. Mastering the subtleties of different shapes became an important skill of many servicemen.

The second time I was aware of the power of pattern recognition was while working for International Paper. The best executives, so I became aware, were those who could take large diverse data files and form a conclusion about an appropriate strategy for the company. That single capability stood out as a defining trait of the better executives.

While attending graduate schools the thought of differing data processing capabilities of executives again crossed my mind. I broached the subject with several industrial psychologists challenging them to develop an instrument to assess that dimension of executive behavior.

As the years went by I continued to bump in to the executive trait of pattern recognition. It was particularly evident as an organization began a strategic planning process. Some were bound to the process of planning while others could vision an outcome long before data foretold an outcome.

Finally, in the mid-1980s I began to read of chaos theory and focused my thoughts to the math of patterns and concept of fractals. By the 1990s with the emergence of large data files, discussions with faculty interested in neuromechanisms, and the growing awareness of biological models supplanting mechanical models as paradigms, I became more convinced that pattern recognition was to be an extremely important field of study. But, I needed someone to champion the effort at Tennessee.

The idea of hosting an international conference was conceived through e-mail communications during May of 1999 between Professor Ham Bozdogan and myself while he was a Visiting Senior Scientist at Tilburg University in Tilburg, The Netherlands.

When he returned to the University of Tennessee, I encouraged Professor Bozdogan to host a meeting in Data Mining and Knowledge Discovery.

The rest is history.

The book is a wonderful testimony to Professor Bozdogan's creative energy in a new and exciting field. It is indeed an honor to have the work of some of the best minds of the field in this book. And, although undeserved, it is flattering to have this volume dedicated to me.

I know of few events in life that can match the excitement associated with an emerging field of study. Data mining and the growing importance of pattern recognition offers society an invaluable tool to capture the full measure of the information age.

C. Warren Neel
Former Dean
College of Business Administration
The University of Tennessee
Knoxville, TN 37996

EDITOR'S PREFACE

"The value of data is no longer in how much of it you have. In the new regime, the value is in how quickly and how effectively can the data be reduced, explored, manipulated and managed."

Usama Fayyad
President & CEO of digiMine, Inc.

C. Warren Neel International Conference on Statistical Data Mining and Knowledge Discovery took place during June 22-25, 2002 at the Marriott Hotel in Knoxville, Tennessee. It was a privilege to host and chair this prestigious conference.

The idea of hosting a conference of this magnitude was born through e-mail correspondence between me and Dr. Warren Neel, then the Dean of the College of Business Administration (CBA) of the University of Tennessee (UT) during May of 1999. Dean Neel in several e-mail messages asked me to investigate Europe's use of data mining while I was on a research visit as a Senior Scientist at Tilburg University in Tilburg, The Netherlands, during May-June of 1999.

After my inquiries both in The Netherlands and other European countries, I reported to Dean Neel on what Europe was doing in the areas of Data Mining, Knowledge Discovery and E-Business. Then, I proposed to him that we should hold an international conference on this area. In his e-mail dated Tuesday, May 18, 1999, Dean Neel replied:

"Ham,

It sounds very promising to me. Let's talk about the necessary steps when you return. Have a good stay.

Warren"

When I returned to UT, Dean Neel encouraged me to organize a Data Mining conference here at UT, and he told me that he would support such an activity. Shortly after this, he was appointed as the Finance Commissioner by the Governor of the State of Tennessee and left UT for Nashville. Because of his vision and genuine support, it was appropriate for me to name the Conference after Dean Warren Neel.

The primary focus of this important conference was to bring national and international experts and practitioners together in this hot cutting-edge research area to share and disseminate new research and developments covering the wide spectrum of areas such as: market segmentation, customer choice behavior, customer profiling, fraud detection and credit scoring, information complexity and Bayesian modeling, econometric and statistical data mining, prediction, and policy-making, manufacturing, improving information quality in loan approval, web mining between eCustomer care and web controlling, data mining in hyperspectral imaging, direct investments in financial assets, textual data mining, neural networks and airport safety logic, evaluating polygraph data, nuclear power plant load forecasting, implementation of data mining in organizations, mammographic computer-aided detection, genomics, proteomics, and many more areas with emphasis to real-world applications.

About 100 researchers from 15 different countries around the world participated in the conference. There were 70 paper presentations including the following conference keynote lectures:

- The Role of Bayesian and Frequentist Multivariate Modeling In Statistical Data Mining, **S. James Press**, University of California

- Intelligent Statistical Data Mining with Information Complexity and Genetic Algorithms, **Hamparsum Bozdogan**, The University of Tennessee

- Econometric and Statistical Data Mining, Prediction, and Policy-Making, **Arnold Zellner,** University of Chicago

- Visual Data Mining, **Edward J. Wegman**, George Mason University

- Top 10 Data Mining Mistakes, **John Elder**, Elder Research, Inc.,

- Data Mining Evolved: Challenges, Applications, and Future Trends, **Usama Fayyad**, digiMine, Inc.,

- Large Contingency Tables: Strategies for Analysis and Inference, **Stephen Fienberg**, Carnegie-Mellon University

- The Evolution of e-Business Intelligence in the ERP World, **Naeem Hashmi**, Information Frameworks.

Statistical data mining is the process of selecting and exploring large amounts of complex information and data using modern statistical techniques and new generation computer algorithms to discover hidden patterns in the data.

This book contains a collection of selected representative papers of the thematic areas covered during the conference, including some of the keynote lectures. It is with regret that Usama Fayyad and Naeem Hashmi could not make a contribution to this volume as the keynote speakers. I am grateful and extend my sincere thanks to all the contributors to this volume, the Session Organizers and Chairs. All the submitted papers were reviewed and put in the format of Chapman & Hall/CRC book style. As the editor, I am ultimately responsible for any inadvertent errors or omissions.

As the host and the chair of the C. Warren Neel International Conference on Statistical Data Mining and Knowledge Discovery, I am indebted to many sponsors of the conference. They include:

Department of Statistics, College of Business Administration, SAS Institute, UT Science Alliance, Office of Research Administration, Oak Ridge National Laboratory (ORNL), Center of Information Technology & Research (CITR), Statistical Innovations, Inc., and Classification Society of North America (CSNA).

I express my sincere thanks for their generous support, which made this conference a reality.

I would like to acknowledge with my sincere thanks the countless hours of help I received in the beginning from Gina Keeling, and later from Derek Norton, my graduate student, who took the job of coordinating the conference. His computer talents were a tremendous help in maintaining the conference web site, updating, and handling other chores of the organizational matters.

I would like to acknowledge the assistants and the wonderful help of Dr. Halima Bensmail, my Conference Co-Chair, and our Ph.D. students: Yongjae Kwon, Junghun Nam, Kim Cooper, Derek Norton, and Michael Lanning. They drove the conference vans, picked up the participants from the airport, and escorted them to different restaurants during the lunch breaks, and so forth. I am grateful to all of them. I am grateful to Karen Poland, Administrative Services Assistant, and Becky Russell, Principal Secretary of the Department of Statistics, in the handling and for helping me with the administrative details on the financial aspects of the Conference.

Avideo Company provided us with a superb audio-visual capability along with the technical assistants of the Computer Resources Group of the College. I am grateful to Dr. Clifton Woods, III, Vice Provost of the University of Tennessee for loaning us five state-of-the-art laptops used throughout the Conference.

I am grateful to Bob Stern, Senior Editor of Chapman & Hall/CRC, for his understanding, patience, and support throughout the final preparation of the book and who extended the deadline several times before delivery of the final version of this book. Dr. June Kim and my Ph.D. student, Dr. Zhenqiu Liu, have been a tremendous help in fixing and compiling the LaTex format of the manuscripts. I acknowledge their contributions with my sincere thanks. I extend my thanks to my colleagues in the Department of Statistics and anyone whom I left out for creating a congenial atmosphere in which to work.

Finally, it is with great honor and pleasure that I dedicate this book to Dean C. Warren Neel who was behind the initial inception and the delivery of this Conference. Without his vision and encouragement for almost four years ago now, we would not be able to make a mark with the contributions of eminent scholars in this new and emerging field of Statistical Data Mining and Knowledge Discovery.

As we know, the field of Statistics is undergoing a fundamental transformation and it is in an evolutionary stage. Its continued health depends very much on its participation in cross-disciplinary research and scholarly activity in many fields. Today, in the information age we live in, with increasingly sophisticated technology for gathering and storing data, many organizations and businesses collect massive amounts of data at accelerated rates and in ever-increasing detail. Massive data sets pose a great challenge to many cross-disciplinary fields in business, including modern statistics. Such data sets have large number of dimensions and often have huge numbers of observations. They are categorical, discrete, quantitative, and often are mixed data types. This high dimensionality and different data types and structures have now outstripped the capability of traditional statistical methods, data analysis, and graphical and data visualization tools.

It is my hope that the reader will find many interesting ideas and challenges in these invaluable contributed papers covering diverse areas of Statistical Data Mining and Knowledge Discovery and that these contributions will stimulate further research in this new cutting-edge field.

Hamparsum Bozdogan
Toby and Brenda McKenzie Professor in Business,
Information Complexity, and Model Selection
Department of Statistics
The University of Tennessee
Knoxville, TN 37996

EDITOR'S BIO

Dr. Hamparsum ("Ham") Bozdogan is Toby and Brenda McKenzie Professor in Business, Information Complexity and in Model Selection in the Department of Statistics, College of Business Administration at the University of Tennessee (UT), Knoxville.

Dr. Bozdogan received his B.S. degree in mathematics, 1970 from the University of Wisconsin-Madison, and both his M.S. and Ph.D. degrees in mathematics, 1978 and 1981, respectively, from the University of Illinois at Chicago majoring in probability and statistics (multivariate statistical analysis and model selection) with a full-minor in operations research. He joined the faculty of UT in the Fall of 1990. Prior to coming to UT he was on the faculty of the University of Virginia in the Department of Mathematics, and was a Visiting Associate Professor and Research Fellow at the prestigious "Akaike's Institute," The Institute of Statistical Mathematics in Tokyo, Japan, during 1988.

Ham is a nationally and internationally recognized expert in the area of informational statistical modeling and model selection. In particular, on the celebrated Akaike's (1971) Information Criterion (AIC), he has extended its range of applications broadly, and has identified and repaired its lack of consistency with a new criterion of his own, which is now being used in many statistical software packages including SAS. Dr. Bozdogan is the developer of a new model selection and validation criterion called ICOMP (ICOMP for 'information complexity'). His new criterion for model selection cleverly seeks, through information theoretic ideas, to find a balance among badness of fit, lack of parsimony, and profusion of complexity. This measures the "statistical chaos" in the model for a given complex data structure. From this basic work, he has undertaken the technical and computational implementation of the criterion to many areas of applications. These include: choosing the number of component clusters in mixture-model cluster analysis, determining the number for factors in frequentist and Bayesian factor analysis, dynamic econometric modeling of food consumption and demand in the U.S. and The Netherlands, detecting influential observations in vector autoregressive models, to mention a few. His results elucidate many current inferential problems in statistics in linear and nonlinear multivariate models and ill-posed problems. His informational modeling techniques are currently being used by many doctoral students at UT, in U.S., and around the world.

Dr. Bozdogan is the recipient of many distinguished teaching and research awards at UT such as:

- The Bank of America Faculty Leadership Medal Award of the College of Business Administration (CBA), April 2001.

- The Hoechst Roussel Teaching and Research Award of the College of Business Administration (CBA), April, 1997.

- Won world research competition award in applied econometric modeling among 28 worldwide participating teams to forecast U.S. and Dutch food consumption during September 1996.

- Chancellor's Award for Research and Creative Achievement, the University of Tennessee, Knoxville (UTK), April 7, 1993. This award is given each year to 10 UTK faculty who have recently made significant contributions in their field of study.

His work has been published in many diverse and leading journals. He is the editor of six books:

1. *Multivariate Statistical Modeling and Data Analysis.* Editor with A. K. Gupta, D. Reidel Publishing Company, Dordrecht, Holland, 1987.

2. *Theory & Methodology of Time Series Analysis,*Volume 1, Proceedings of First U.S./Japan Conference on The Frontiers of Statistical Modeling: An Informational Approach. Editor, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994.

3. *Multivariate Statistical Modeling*, Volume 2, Proceedings of First U.S./Japan Conference on The Frontiers of Statistical Modeling: An Informational Approach. Editor, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994.

4. *Engineering & Scientific Applications of Informational Modeling*, Volume 3, Proceedings of First U.S./Japan Conference on The Frontiers of Statistical Modeling:

An Informational Approach. Editor, Kluwer Academic Publishers, The Netherlands, 1994.

5. *Measurement and Multivariate Analysis*, Co-editor with S. Nishisato, Y. Baba, and K. Kanefuji, Springer-Verlag, Tokyo, Japan, March 2002.

He is the author of forthcoming modern textbooks:

- *Statistical Modeling and Model Evaluation: A New Informational Approach*,
- *Informational Complexity and Multivariate Modeling*

using an open architecture easy-to-use command driven computational environment MATLAB.

Dr. Bozdogan is a member of many professional societies and serves as the referee to many prestigious statistical journals. His current research innovations are in developing intelligent hybrid models between any complex modeling problem, genetic algorithms (GAs) and his information complexity *ICOMP* criterion as the fitness function. Coupled with this, his current research is focused on a long-standing problem of model selection under misspecification, and combining robustness and misspecification within *ICOMP* criteria. He is developing new techniques which are robust and misspecification resistant. This is important because this new approach provides researchers and practitioners with knowledge of how to guard against the misspecification of the model as we actually fit and evaluate these models, and guard against spurious observations. These new developments and results are very important in many areas of applied and basic research (e.g., in business, engineering, social and behavioral, and medical sciences), which is currently ignored.

LIST OF PRESENTERS

S. James Press
Department of Statistics, University of California, Riverside, CA 92521-0138, USA
jpress@ucrac1.ucr.edu

Hamparsum Bozdogan
Department of Statistics, University of Tennessee, Knoxville, TN 37996-0532, USA
bozdogan@utk.edu

Arnold Zellner
Graduate School of Business, University of Chicago, Chicago, IL 60637, USA
fazellne@gsb.uchicago.edu

Edward J. Wegman
Center for Computational Statistics, George Mason University, Fairfax, VA 22030-4444, USA
ewegman@gmu.edu

Adrian Dobra
ISDS, Duke University, Durham, NC 27708, USA
adobra@stat.duke.edu

Elena A. Erosheva
Department of Statistics, University of Washington, Seattle, WA 98195, USA
elena@stat.washington.edu

Stephen E. Fienberg
Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA 15213, USA
fienberg@stat.cmu.edu

Aleksandra B. Slavkovic
Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA 15213, USA
sesa@stat.cmu.edu

Hyunjoong Kim
Department of Statistics, University of Tennessee, Knoxville, TN 37996-0532, USA
hjkim@utk.edu

Lynette Hunt
Department of Statistics, University of Waikato, Hamilton, New Zealand
lah@stats.waikato.ac.nz

Zhenqiu Liu
Community Health Research Group, University of Tennessee, Knoxville, TN 37996-0532, USA
zliu@utk.edu

Andrei V. Gribok
Department of Electrical and Computer Engineering, University of Tennessee,
Knoxville, TN 37996-0532, USA
agribok@utk.edu

Aleksey M. Urmanov
Physical Sciences Center, Sun Microsystems, Inc., San Diego, CA 92121, USA
aleksey.urmanov@sun.com

Christopher M. Hill
Industrial Engineering & Management Systems, University of Central Florida,
Orlando, FL 32816-2450, USA
christopher-hill@us.army.mil

Masahiro Mizuta
Center for Information & Multimedia Studies, Hokkaido University, Sapporo, Japan
mizuta@cims.hokudai.ac.jp

Mutasem Hiassat
Automatic Control & Systems Engineering, University of Sheffield, Sheffield, UK
matt@hiassat.freeserve.co.uk

Belle R. Upadhyaya and Baofu Lu
Department of Nuclear Engineering, University of Tennessee,
Knoxville, TN 37996-0532, USA
bupadhya@utk.edu and *blu@utk.edu*

Francois Boussu
Ecole Nationale Superieure des Arts et Industries Textiles, Roubaix, 59056, France
francois.boussu@ensait.fr

Jean-Jacques Denimal
Laboratory of Probability and Statistics, University of Sciences and Technologies of Lille, 59655, France
Jean-Jacques.Denimal@univ-lille1.fr

Friedrich Leisch
Institut für Statistik, Vienna University of Technology,
Vienna, A-1040 Austria
Friedrich.Leisch@ci.tuwien.ac.at

Hairong Qi
Department of Electrical and Computer Engineering, University of Tennessee,
Knoxville, TN 37996-0532, USA
hqi@utk.edu

Sami Al-Harbi
School of Information Systems, University of East Anglia, Norwich NR4 7TJ, UK
shh@sys.vea.ac.uk

J. Michael Lanning
Department of Statistics, University of Tennessee, Knoxville, TN 37996-0532, USA
jlanning@utk.edu

Jay Magidson
Statistical Innovations, Belmont, MA 02478, USA
jay@statisticalinnovations.com

M. Ishaq Bhatti
Department of Operations Management & Business Statistics
Sultan Qaboos University, Muscat, Oman
abish@squ.edu.om

Amar Gupta
Productivity From Information Technology (PROFIT) Initiative MIT
Sloan School of Management, MIT, Cambridge, MA 02139, USA
agupta@mit.edu

Dong Xu
Protein Informatics Group, Life Sciences Division,
Oak Ridge National Laboratory, Oak Ridge, TN 37831-6480, USA
xud@ornl.gov

Andreas Geyer-Schulz and Andreas Neumann
Information Services and Electronic Markets,University of Karlsruhe
(TH), 76128 Karlsruhe, Germany
andreas.neumann@em.uni-karlsruhe.de

Michael Berry
Department of Computer Science, University of Tennessee, Knoxville, TN 37996-0532, USA
berry@cs.utk.edu

Julien Blanchard
Ecole Polytechnique de l'université de rue Lu Nantes, 44000 France
julien.blanchard@polytech.univ-nantes.fr