Claudia Bauzer Medeiros
Max Egenhofer
Elisa Bertino (Eds.)

LNCS 3633

Advances in Spatial and Temporal Databases

9th International Symposium, SSTD 2005
Angra dos Reis, Brazil, August 2005
Proceedings

Springer

Claudia Bauzer Medeiros Max Egenhofer
Elisa Bertino (Eds.)

Advances in Spatial and Temporal Databases

9th International Symposium, SSTD 2005
Angra dos Reis, Brazil, August 22-24, 2005
Proceedings

Springer

Volume Editors

Claudia Bauzer Medeiros
University of Campinas, Institute of Computing
CP 6176, 13084-971 Campinas, Brazil
E-mail: cmbm@ic.unicamp.br

Max Egenhofer
University of Maine
National Center for Geographic Information and Analysis
348 Boardman Hall, Orono, ME 04469-5711, USA
E-mail: max@spatial.maine.edu

Elisa Bertino
Purdue University, Department of Computer Science
West Lafayette, IN, USA
E-mail: bertino@cerias.purdue.edu

Library of Congress Control Number: 2005929997

CR Subject Classification (1998): H.2, H.3, H.4, I.2.4

ISSN 0302-9743
ISBN-10 3-540-28127-4 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-28127-6 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11535331 06/3142 5 4 3 2 1 0

Lecture Notes in Computer Science 3633

Preface

It is our great pleasure to introduce the papers of the proceedings of the 9th International Symposium on Spatial and Temporal Databases – SSTD 2005. This year's symposium continues the tradition of being the premier forum for the presentation of research results and experience reports on leading edge issues of spatial and temporal database systems, including data models, systems, applications and theory. The mission of the symposium is to share innovative solutions that fulfill the needs of novel applications and heterogeneous environments and identify new directions for future research and development. SSTD 2005 gives researchers and practitioners a unique opportunity to share their perspectives with others interested in the various aspects of database systems for managing spatial and temporal data and for supporting their applications.

A total of 77 papers were submitted this year from several countries. After a thorough review process, the program committee accepted 24 papers covering a variety of topics, including indexing techniques and query processing, mobile environments and moving objects, and spatial and temporal data streams. We are very pleased with the variety of the symposium's topics, and we are proud of the resulting strong program.

Many people contributed to the success of the SSTD 2005 program. First of all, we would like to thank the authors for providing the content of the program, and all the members of the program committee and the additional reviewers, for their detailed comments. Philippe Rigaux was of help in adding functions to his program MyReview, which was used in the reviewing process. We would also like to express our gratitude to Gilberto Câmara, the general chair of SSTD 2005, for his constant guidance and advice on many organizational aspects of the symposium and for his work on the local arrangements. Finally, we would like to thank our sponsors (notably INPE – the Brazilian National Institute for Space Research) who have enabled us to hold a successful meeting. We are also grateful for the support of the Brazilian Computer Society (SBC).

We hope that you find this program to be both beneficial and enjoyable and that the symposium provides you with the opportunity to meet other researchers and practitioners from institutions around the world. Enjoy!!

August 2005

Claudia Bauzer Medeiros,
Max Egenhofer,
Elisa Bertino

Organization

SSTD 2005 was organized by the Department of Image Processing of INPE – the National Institute for Space Research (Instituto Nacional de Pesquisas Espaciais), Brazil.

Executive Committee

General Conference Chair:	Gilberto Câmara (Department of Image Processing, INPE, Brazil)
Program Chairs:	Elisa Bertino (Department of Computer Sciences, Purdue University, USA)
	Max Egenhofer (NCGIA, University of Maine, USA)
	Claudia Bauzer Medeiros (Institute of Computing, University of Campinas, Brazil)

Program Committee

Amr El Abbadi	UC Santa Barbara, USA
Walid G. Aref	Purdue University, USA
Alessandro Artale	University of Bolzano, Italy
Alberto Belussi	University of Verona, Italy
Michela Bertolotto	University College Dublin, Ireland
Gilberto Camara	National Institute for Space Research, Brazil
Marco Casanova	Dept Informatics, PUC-Rio, Brazil
Barbara Catania	University of Genoa, Italy
Christophe Claramunt	Naval Academy Research Institute, France
Matt Duckham	University of Melbourne, Australia
Fred Fonseca	Penn State University, USA
Fosca Giannotti	CNR, Italy
Ralf Hartmut Güting	University of Hagen, Germany
Kathleen Hornsby	University of Maine, USA
Christian S. Jensen	Aalborg University, Denmark
Christopher Jones	Cardiff University, UK
Daniel Keim	University of Constance, Germany
Eamonn Keogh	University of California, Riverside, USA
George Kollios	Boston University, USA
Bart Kuijpers	University of Limburg, Belgium
Mario Nascimento	University of Alberta, Canada
Raymond Ng	University of British Columbia, Canada

Silvia Nittel	NCGIA, University of Maine, USA
Beng Chin Ooi	National University of Singapore, Singapore
Peter van Oosterom	Delft Univ. of Technology, The Netherlands
Dimitris Papadias	Hong Kong Univ. of Science and Technology, Hong Kong
Jignesh Patel	University of Michigan, USA
Sunil Prabhakar	Purdue University, USA
Philippe Rigaux	University Paris IX, France
Andrea Rodríguez-Tastets	University of Concepción, Chile
Ana Carolina Salgado	University of Pernambuco, Brazil
George Samaras	University of Cyprus, Cyprus
Peter Scheuermann	Northwestern University, USA
Markus Schneider	University of Florida, USA
Bernhard Seeger	University of Marburg, Germany
Cyrus Shahabi	University of Southern California, USA
Shashi Shekhar	University of Minnesota, Minneapolis, USA
Rick Snodgrass	University of Arizona, USA
Stefano Spaccapietra	EPFL Lausanne, Switzerland
Paolo Terenziani	Università del Piemonte Orientale, Italy
Yannis Theodoridis	University of Piraeus, Greece
Nectaria Tryfona	Computer Technology Institute, Greece
Michalis Vazirgiannis	Athens Univ. of Economics and Business, Greece
Agnes Voisard	Fraunhofer ISST and FU Berlin, Germany
Ouri Wolfson	University of Illinois at Chicago, USA

Additional Referees

Nagender Bandi
Benjamin Bustos
Hu Cao
Reynold Cheng
Alminas Civilis
Stephen Cole
Carlo Combi
Stephane Coulondre
Pier Luigi Dragotti
Ying Feng
Robson N. Fidalgo
Elias Frentzos
Marios Hadjieleftheriou
Christoph Heinz
Xuegang Huang
Mohammad R. Kolahdouzan
Feifei Li
Xiang Lian
Dan Lin
Juhong Liu
Andrei Lopatenko
Anna Maddalena
Florian Mansmann
Ahmed Metwally
Mohamed Mokbel
Kyriakos Mouratidis
Guillaume Noel
Andrea Nucita
Nikos Pelekis
Paola Podesta
Gianna Reggio
Mehdi Sharifzadeh
Sarvjeet Singh
Joern Schneidewind
Chengyu Sun
Valeria C. Times
Goce Trajcevski
Yicheng Tu
Tao Wan
David Yang
Kiyoung Yang
Huabei Yin
Man Lung Yiu
Yuni Xia
Xiaopeng Xiong
Bo Xu
Mingwu Zhang
Hartmut Ziegler

Sponsoring Institutions

Sponsor – National Institute of Space Research – INPE – Brazil
Support – Brazilian Computer Society (SBC)

Lecture Notes in Computer Science

For information about Vols. 1–3511

please contact your bookseller or Springer

Vol. 3633: C. Bauzer Medeiros, M. Egenhofer, E. Bertino (Eds.), Advances in Spatial and Temporal Databases. XIII, 433 pages. 2005.

Vol. 3632: R. Nieuwenhuis (Ed.), Automated Deduction - CADE-20. XIII, 459 pages. 2005. (Subseries LNAI).

Vol. 3626: B. Ganter, G. Stumme, R. Wille (Eds.), Formal Concept Analysis. X, 349 pages. 2005. (Subseries LNAI).

Vol. 3621: V. Shoup (Ed.), Advances in Cryptology – CRYPTO 2005. XI, 568 pages. 2005.

Vol. 3619: X. Lu, W. Zhao (Eds.), Networking and Mobile Computing. XXIV, 1299 pages. 2005.

Vol. 3615: B. Ludäscher, L. Raschid (Eds.), Data Integration in the Life Sciences. XII, 344 pages. 2005. (Subseries LNBI).

Vol. 3608: F. Dehne, A. López-Ortiz, J.-R. Sack (Eds.), Algorithms and Data Structures. XIV, 446 pages. 2005.

Vol. 3607: J.-D. Zucker, L. Saitta (Eds.), Abstraction, Reformulation and Approximation. XII, 376 pages. 2005. (Subseries LNAI).

Vol. 3598: H. Murakami, H. Nakashima, H. Tokuda, M. Yasumura, Ubiquitous Computing Systems. XIII, 275 pages. 2005.

Vol. 3597: S. Shimojo, S. Ichii, T.W. Ling, K.-H. Song (Eds.), Web and Communication Technologies and Internet-Related Social Issues - HSI 2005. XIX, 368 pages. 2005.

Vol. 3596: F. Dau, M.-L. Mugnier, G. Stumme (Eds.), Conceptual Structures: Common Semantics for Sharing Knowledge. XI, 467 pages. 2005. (Subseries LNAI).

Vol. 3594: J.C. Setubal, S. Verjovski-Almeida (Eds.), Advances in Bioinformatics and Computational Biology. XIV, 258 pages. 2005. (Subseries LNBI).

Vol. 3587: P. Perner, A. Imiya (Eds.), Machine Learning and Data Mining in Pattern Recognition. XVII, 695 pages. 2005. (Subseries LNAI).

Vol. 3586: A.P. Black (Ed.), ECOOP 2005 - Object-Oriented Programming. XVII, 631 pages. 2005.

Vol. 3584: X. Li, S. Wang, Z.Y. Dong (Eds.), Advanced Data Mining and Applications. XIX, 835 pages. 2005. (Subseries LNAI).

Vol. 3583: R.W. H. Lau, Q. Li, R. Cheung, W. Liu (Eds.), Advances in Web-Based Learning – ICWL 2005. XIV, 420 pages. 2005.

Vol. 3582: J. Fitzgerald, I.J. Hayes, A. Tarlecki (Eds.), FM 2005: Formal Methods. XIV, 558 pages. 2005.

Vol. 3581: S. Miksch, J. Hunter, E. Keravnou (Eds.), Artificial Intelligence in Medicine. XVII, 547 pages. 2005. (Subseries LNAI).

Vol. 3580: L. Caires, G.F. Italiano, L. Monteiro, C. Palamidessi, M. Yung (Eds.), Automata, Languages and Programming. XXV, 1477 pages. 2005.

Vol. 3579: D. Lowe, M. Gaedke (Eds.), Web Engineering. XXII, 633 pages. 2005.

Vol. 3578: M. Gallagher, J. Hogan, F. Maire (Eds.), Intelligent Data Engineering and Automated Learning - IDEAL 2005. XVI, 599 pages. 2005.

Vol. 3577: R. Falcone, S. Barber, J. Sabater-Mir, M.P. Singh (Eds.), Trusting Agents for Trusting Electronic Societies. VIII, 235 pages. 2005. (Subseries LNAI).

Vol. 3576: K. Etessami, S.K. Rajamani (Eds.), Computer Aided Verification. XV, 564 pages. 2005.

Vol. 3575: S. Wermter, G. Palm, M. Elshaw (Eds.), Biomimetic Neural Learning for Intelligent Robots. IX, 383 pages. 2005. (Subseries LNAI).

Vol. 3574: C. Boyd, J.M. González Nieto (Eds.), Information Security and Privacy. XIII, 586 pages. 2005.

Vol. 3573: S. Etalle (Ed.), Logic Based Program Synthesis and Transformation. VIII, 279 pages. 2005.

Vol. 3572: C. De Felice, A. Restivo (Eds.), Developments in Language Theory. XI, 409 pages. 2005.

Vol. 3571: L. Godo (Ed.), Symbolic and Quantitative Approaches to Reasoning with Uncertainty. XVI, 1028 pages. 2005. (Subseries LNAI).

Vol. 3570: A. S. Patrick, M. Yung (Eds.), Financial Cryptography and Data Security. XII, 376 pages. 2005.

Vol. 3569: F. Bacchus, T. Walsh (Eds.), Theory and Applications of Satisfiability Testing. XII, 492 pages. 2005.

Vol. 3568: W.-K. Leow, M.S. Lew, T.-S. Chua, W.-Y. Ma, L. Chaisorn, E.M. Bakker (Eds.), Image and Video Retrieval. XVII, 672 pages. 2005.

Vol. 3567: M. Jackson, D. Nelson, S. Stirk (Eds.), Database: Enterprise, Skills and Innovation. XII, 185 pages. 2005.

Vol. 3566: J.-P. Banâtre, P. Fradet, J.-L. Giavitto, O. Michel (Eds.), Unconventional Programming Paradigms. XI, 367 pages. 2005.

Vol. 3565: G.E. Christensen, M. Sonka (Eds.), Information Processing in Medical Imaging. XXI, 777 pages. 2005.

Vol. 3564: N. Eisinger, J. Małuszyński (Eds.), Reasoning Web. IX, 319 pages. 2005.

Vol. 3562: J. Mira, J.R. Álvarez (Eds.), Artificial Intelligence and Knowledge Engineering Applications: A Bioinspired Approach, Part II. XXIV, 636 pages. 2005.

Vol. 3561: J. Mira, J.R. Álvarez (Eds.), Mechanisms, Symbols, and Models Underlying Cognition, Part I. XXIV, 532 pages. 2005.

Vol. 3560: V.K. Prasanna, S. Iyengar, P.G. Spirakis, M. Welsh (Eds.), Distributed Computing in Sensor Systems. XV, 423 pages. 2005.

Vol. 3559: P. Auer, R. Meir (Eds.), Learning Theory. XI, 692 pages. 2005. (Subseries LNAI).

Vol. 3558: V. Torra, Y. Narukawa, S. Miyamoto (Eds.), Modeling Decisions for Artificial Intelligence. XII, 470 pages. 2005. (Subseries LNAI).

Vol. 3557: H. Gilbert, H. Handschuh (Eds.), Fast Software Encryption. XI, 443 pages. 2005.

Vol. 3556: H. Baumeister, M. Marchesi, M. Holcombe (Eds.), Extreme Programming and Agile Processes in Software Engineering. XIV, 332 pages. 2005.

Vol. 3555: T. Vardanega, A.J. Wellings (Eds.), Reliable Software Technology – Ada-Europe 2005. XV, 273 pages. 2005.

Vol. 3554: A. Dey, B. Kokinov, D. Leake, R. Turner (Eds.), Modeling and Using Context. XIV, 572 pages. 2005. (Subseries LNAI).

Vol. 3553: T.D. Hämäläinen, A.D. Pimentel, J. Takala, S. Vassiliadis (Eds.), Embedded Computer Systems: Architectures, Modeling, and Simulation. XV, 476 pages. 2005.

Vol. 3552: H. de Meer, N. Bhatti (Eds.), Quality of Service – IWQoS 2005. XVIII, 400 pages. 2005.

Vol. 3551: T. Härder, W. Lehner (Eds.), Data Management in a Connected World. XIX, 371 pages. 2005.

Vol. 3548: K. Julisch, C. Kruegel (Eds.), Intrusion and Malware Detection and Vulnerability Assessment. X, 241 pages. 2005.

Vol. 3547: F. Bomarius, S. Komi-Sirviö (Eds.), Product Focused Software Process Improvement. XIII, 588 pages. 2005.

Vol. 3546: T. Kanade, A. Jain, N.K. Ratha (Eds.), Audio- and Video-Based Biometric Person Authentication. XX, 1134 pages. 2005.

Vol. 3544: T. Higashino (Ed.), Principles of Distributed Systems. XII, 460 pages. 2005.

Vol. 3543: L. Kutvonen, N. Alonistioti (Eds.), Distributed Applications and Interoperable Systems. XI, 235 pages. 2005.

Vol. 3542: H.H. Hoos, D.G. Mitchell (Eds.), Theory and Applications of Satisfiability Testing. XIII, 393 pages. 2005.

Vol. 3541: N.C. Oza, R. Polikar, J. Kittler, F. Roli (Eds.), Multiple Classifier Systems. XII, 430 pages. 2005.

Vol. 3540: H. Kalviainen, J. Parkkinen, A. Kaarna (Eds.), Image Analysis. XXII, 1270 pages. 2005.

Vol. 3539: K. Morik, J.-F. Boulicaut, A. Siebes (Eds.), Local Pattern Detection. XI, 233 pages. 2005. (Subseries LNAI).

Vol. 3538: L. Ardissono, P. Brna, A. Mitrovic (Eds.), User Modeling 2005. XVI, 533 pages. 2005. (Subseries LNAI).

Vol. 3537: A. Apostolico, M. Crochemore, K. Park (Eds.), Combinatorial Pattern Matching. XI, 444 pages. 2005.

Vol. 3536: G. Ciardo, P. Darondeau (Eds.), Applications and Theory of Petri Nets 2005. XI, 470 pages. 2005.

Vol. 3535: M. Steffen, G. Zavattaro (Eds.), Formal Methods for Open Object-Based Distributed Systems. X, 323 pages. 2005.

Vol. 3534: S. Spaccapietra, E. Zimányi (Eds.), Journal on Data Semantics III. XI, 213 pages. 2005.

Vol. 3533: M. Ali, F. Esposito (Eds.), Innovations in Applied Artificial Intelligence. XX, 858 pages. 2005. (Subseries LNAI).

Vol. 3532: A. Gómez-Pérez, J. Euzenat (Eds.), The Semantic Web: Research and Applications. XV, 728 pages. 2005.

Vol. 3531: J. Ioannidis, A. Keromytis, M. Yung (Eds.), Applied Cryptography and Network Security. XI, 530 pages. 2005.

Vol. 3530: A. Prinz, R. Reed, J. Reed (Eds.), SDL 2005: Model Driven. XI, 361 pages. 2005.

Vol. 3528: P.S. Szczepaniak, J. Kacprzyk, A. Niewiadomski (Eds.), Advances in Web Intelligence. XVII, 513 pages. 2005. (Subseries LNAI).

Vol. 3527: R. Morrison, F. Oquendo (Eds.), Software Architecture. XII, 263 pages. 2005.

Vol. 3526: S. B. Cooper, B. Löwe, L. Torenvliet (Eds.), New Computational Paradigms. XVII, 574 pages. 2005.

Vol. 3525: A.E. Abdallah, C.B. Jones, J.W. Sanders (Eds.), Communicating Sequential Processes. XIV, 321 pages. 2005.

Vol. 3524: R. Barták, M. Milano (Eds.), Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems. XI, 320 pages. 2005.

Vol. 3523: J.S. Marques, N. Pérez de la Blanca, P. Pina (Eds.), Pattern Recognition and Image Analysis, Part II. XXVI, 733 pages. 2005.

Vol. 3522: J.S. Marques, N. Pérez de la Blanca, P. Pina (Eds.), Pattern Recognition and Image Analysis, Part I. XXVI, 703 pages. 2005.

Vol. 3521: N. Megiddo, Y. Xu, B. Zhu (Eds.), Algorithmic Applications in Management. XIII, 484 pages. 2005.

Vol. 3520: O. Pastor, J. Falcão e Cunha (Eds.), Advanced Information Systems Engineering. XVI, 584 pages. 2005.

Vol. 3519: H. Li, P. J. Olver, G. Sommer (Eds.), Computer Algebra and Geometric Algebra with Applications. IX, 449 pages. 2005.

Vol. 3518: T.B. Ho, D. Cheung, H. Liu (Eds.), Advances in Knowledge Discovery and Data Mining. XXI, 864 pages. 2005. (Subseries LNAI).

Vol. 3517: H.S. Baird, D.P. Lopresti (Eds.), Human Interactive Proofs. IX, 143 pages. 2005.

Vol. 3516: V.S. Sunderam, G.D.v. Albada, P.M.A. Sloot, J.J. Dongarra (Eds.), Computational Science – ICCS 2005, Part III. LXIII, 1143 pages. 2005.

Vol. 3515: V.S. Sunderam, G.D.v. Albada, P.M.A. Sloot, J.J. Dongarra (Eds.), Computational Science – ICCS 2005, Part II. LXIII, 1101 pages. 2005.

Vol. 3514: V.S. Sunderam, G.D.v. Albada, P.M.A. Sloot, J.J. Dongarra (Eds.), Computational Science – ICCS 2005, Part I. LXIII, 1089 pages. 2005.

Vol. 3513: A. Montoyo, R. Muñoz, E. Métais (Eds.), Natural Language Processing and Information Systems. XII, 408 pages. 2005.

Vol. 3512: J. Cabestany, A. Prieto, F. Sandoval (Eds.), Computational Intelligence and Bioinspired Systems. XXV, 1260 pages. 2005.

Table of Contents

Advanced Query Processing II

Indexing Schemes and Structures

Novel Applications and Real Systems

Moving Objects and Mobile Environments

Advanced Query Processing III

Selectivity Estimation of High Dimensional Window Queries via Clustering

Christian Böhm, Hans-Peter Kriegel, Peer Kröger, and Petra Linhart

Institute for Computer Science, University of Munich, Germany
{boehm, kriegel, kroegerp, linhart}@dbs.ifi.lmu.de

Abstract. Query optimization is an important functionality of modern database systems and often based on estimating the selectivity of queries before actually executing them. Well-known techniques for estimating the result set size of a query are sampling and histogram-based solutions. Sampling-based approaches heavily depend on the size of the drawn sample which causes a trade-off between the quality of the estimation and the time in which the estimation can be executed for large data sets. Histogram-based techniques eliminate this problem but are limited to low-dimensional data sets. They either assume that all attributes are independent which is rarely true for real-world data or else get very inefficient for high-dimensional data. In this paper we present the first multivariate parametric method for estimating the selectivity of window queries for large and high-dimensional data sets. We use clustering to compress the data by generating a precise model of the data using multivariate Gaussian distributions. Additionally, we show efficient techniques to evaluate a window query against the Gaussian distributions we generated. Our experimental evaluation shows that this approach is significantly more efficient for multidimensional data than all previous approaches.

1 Introduction

The storage and management of vectors of a multidimensional feature space has become an important basic functionality of a database system. Advanced applications such as multimedia [1], CAD [2], molecular biology [3], etc. require efficient and effective methods for content based similarity search and data mining. Such methods are typically based on feature vectors of moderate or high dimensionality. Although a vast number of index structures [4,5] and access methods [6] for vector data has been proposed, database management systems do not yet support the storage and retrieval of vector data in the same way as relational data from applications such as accounting and billing. In order to give full support to advanced applications the database system needs efficient and effective techniques for query optimization. One of the most important challenges in query optimization is the estimation of the selectivity of a query predicate. While a number of techniques to model the data distribution and thus to estimate the selectivity are known for one- and low-dimensional data spaces, this is still an unsolved problem for data spaces of medium to high dimensionality.

C. Bauzer Medeiros et al. (Eds.): SSTD 2005, LNCS 3633, pp. 1–18, 2005.

Three different paradigms of data modelling for selectivity estimation in general can be distinguished: Histograms, sampling, and parametric techniques. Of those three, only sampling can be directly applied without modification in higher dimensional data spaces. Many different sampling methods have been proposed. They share the common idea to evaluate the predicate on top of a small subset of the actual database objects and to extrapolate the observed selectivity. The well-known techniques differ in the way how the sample is drawn as well as in the determination of the suitable size of the sample. The general drawback of sampling techniques is that the accuracy of the result is strictly limited by the sample rate. To get an accurate estimation of the selectivity, a large sample (>10%) of the database is required. To evaluate the query on top of the large sample is not much cheaper than to evaluate it on the original data set which limits its usefulness for query optimization.

Histogram techniques, the most prevalent paradigm to model the data distribution in the one-dimensional case, have a different problem. This concept is very difficult to be carried over to the multidimensional case, even for low or moderate dimensional data. One way to adapt one-dimensional histograms to multidimensional data is to describe the distribution of the individual attributes of the vectors independently by usual histograms. These histograms are sometimes called marginal distributions. In this case, the selectivity of multidimensional queries can be determined easily provided that the attributes are statistically independent, i.e. neither correlated nor clustered. Real-world data sets, however, rarely fulfill this condition. Another approach is to partition the data space by a multidimensional grid and to assign a histogram bin to each grid cell. This approach may be possible for two- and three-dimensional spaces. However, for higher dimensional data this method becomes inefficient and ineffective since the number of grid cells is exponential in the dimensionality. Techniques of dimensionality reduction such as Fourier transformation, wavelets, principal component analysis or space-filling curves (Z-ordering, Hilbert) may reduce this problem to some extent. The possible problem reduction, however, is limited by the intrinsic dimensionality of the data set.

The idea of parametric techniques is to describe the data distribution by curves (functions) which have been fitted into the data set. In most cases Gaussian functions (normal distributions) are used. Instead of using one single Gaussian, a set of multivariate Gaussians can be fitted into the data set which makes the technique more accurate. Each Gaussian is then described by three parameters (mean, variance and the relative weight of the Gaussian in the ensemble). This approach can be transferred into the multidimensional case by two techniques. Like described above for histograms, the marginal distribution of each attribute can be modelled independently by a set of Gaussians. The multidimensional query selectivity can be estimated by combining the marginal distributions. This approach leads to similar problems like marginal histograms.

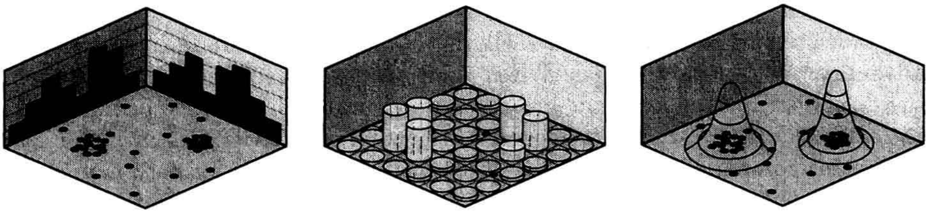
Therefore, our solution is different. Our technique directly models the multidimensional data distribution by a set of multivariate Gaussian functions. There are two options to use the Gaussian primitives: The Gaussians can either be used with a matrix containing both variances and covariances or with a vector of the multivariate variances only. As we will discuss later, both approaches have their advantages and disadvantages. When using Gaussians with covariance matrix, the data distribution can be described more accurately by a single primitive. On the other side, more storage is needed for the covariance matrices ($O(d^2)$ for each Gaussian) compared to the variance vector approach ($O(d)$ for each Gaussian). Moreover, the processing cost for reading the parameters and for the determination of the estimated selectivity is much higher when covariance matrices are used. Let us note that, unlike the approaches using marginal distributions, our Gaussian technique with no covariance matrix does not rely on the attribute independence assumption. This technique assumes attribute independence for each individual Gaussian primitive only, but places no constraints to the overall data distribution. We will discuss this issue in more detail in Section 4, an experimental validation is given in Section 5.

To obtain a collection of Gaussians distributions we apply a clustering algorithm. Clustering is the task of grouping vectors into different subsets (the clusters) such that the intra-cluster similarity is maximized and the inter-cluster similarity is minimized. That means points belonging to the same cluster are close together whereas points of different clusters are far away from each other. Many different algorithms have been proposed such as k-means [7], single-link [8], density-based clustering [9,10] and many others. Most of these algorithms use a point as a representative of each cluster. In contrast, the EM clustering algorithm (expectation maximization) [11] uses a multivariate Gaussian function as a cluster representative. We will discuss the suitability of different variants of the EM algorithm for our problem of getting a good approximation of the actual data distribution.

To summarize our contribution, we propose in this paper a new cost model for estimating the selectivity of multidimensional queries on top of vector data of medium to high dimensionality. The data distribution is represented by a set of multivariate Gaussian functions that have been determined using the EM clustering algorithm. We develop two methods for estimating the selectivity of window queries and range queries using the multivariate Gaussians. We demonstrate experimentally the superiority of our approach over competitive cost models based on histograms and sampling. The remainder of our paper is organized as follows: In Section 2 we discuss related work on selectivity estimation and point out our contribution. Section 3 and 4 describes in detail our proposed methods to find a representation of the data distribution by an ensemble of multivariate Gaussian functions using EM clustering and to estimate the selectivity on top of this model. Section 5 contains the experimental evaluation, and section 6 concludes our paper.

2 Related Work

In this chapter, we review current approaches for selectivity estimation and discuss their potentials.



(a) One dimensional histograms (b) Multi dimensional histograms (c) Selectivity estimation via clustering

Fig. 1. Visualization of different concepts for selectivity estimation

2.1 Review

Recent work on selectivity estimation can be categorized into three classes, namely histogram-based methods, sampling-based methods, and parametric methods. In the following, we review and discuss the most important representatives of each class briefly.

Histogram-based Methods. The most widespread approach for selectivity estimation in practice is the use of histograms. In general, the data space is partitioned into buckets, and the frequency of points inside each bucket is computed. We can distinguish between one-dimensional and multi-dimensional histograms.

Selectivity estimation using one-dimensional histograms is based on the assumption that the attributes of the data set are independent, i.e. there is no correlation between different dimensions of the feature space. For each dimension, a histogram is built and the selectivity of a window query q is estimated in each dimension separately. The selectivity of q in the full-dimensional space is evaluated by multiplying the selectivity estimations for each attribute. Equi-width histograms [12] compute buckets of fixed size and variable point frequency, whereas equi-depth histograms [13] compute buckets of variable size and fixed point frequency.

With growing dimensionality of the feature space, the recombination of one-dimensional buckets becomes costly. Thus, in recent years, multi-dimensional histograms have been investigated. Multi-dimensional equi-depth histograms [14] partition the feature space into multi-dimensional buckets with variable size and fixed point frequency. In [14] an algorithm to construct multi-dimensional equi-depth histograms is presented that iteratively partitions the data space along each attribute into a fixed number of buckets, where the order of attributes is fixed. The selectivity of a window query q is estimated analogously to one-dimensional histograms taking the buckets into account that intersect with q. The algorithm MHIST [15] partitions the data space along the single attributes in a similar way, but decides in each step which attribute is partitioned rather than processing the attributes in a fixed order.

STHoles [16] is a recent approach that proposes hierarchically organized multi-dimensional histograms. A histogram may contain another histogram com-