Josep Domingo-Ferrer
Vicenç Torra (Eds.)

LNCS 3050

Privacy in Statistical Databases

CASC Project Final Conference, PSD 2004
Barcelona, Catalonia, Spain, June 2004
Proceedings

Springer

Josep Domingo-Ferrer Vicenç Torra (Eds.)

Privacy in Statistical Databases

CASC Project Final Conference, PSD 2004
Barcelona, Catalonia, Spain, June 9-11, 2004
Proceedings

Springer

Volume Editors

Josep Domingo-Ferrer
Universitat Rovira i Virgili
Department of Computer Engineering and Mathematics
Av. Països Catalans 26, 43007 Tarragona, Catalonia, Spain
E-mail: jdomingo@etse.urv.es

Vicenç Torra
Institut d'Investigació en Intel·ligència Artificial
Campus de Bellaterra, 08193 Bellaterra, Catalonia, Spain
E-mail: vtorra@iiia.csic.es

Library of Congress Control Number: 2004106913

CR Subject Classification (1998): H.2.8, G.3, K.4.1, I.2.4

ISSN 0302-9743
ISBN 3-540-22118-2 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag is a part of Springer Science+Business Media

springeronline.com

Typesetting: Camera-ready by author, data conversion by Olgun Computergrafik
Printed on acid-free paper SPIN: 11008736 06/3142 5 4 3 2 1 0

Lecture Notes in Computer Science 3050

Lecture Notes in Computer Science

For information about Vols. 1–2980

please contact your bookseller or Springer-Verlag

Vol. 3092: J. Eckstein, H. Baumeister (Eds.), Extreme Programming and Agile Processes in Software Engineering. XVI, 358 pages. 2004.

Vol. 3091: V. van Oostrom (Ed.), Rewriting Techniques and Applications. X, 313 pages. 2004.

Vol. 3084: A. Persson, J. Stirna (Eds.), Advanced Information Systems Engineering. XIV, 596 pages. 2004.

Vol. 3083: W. Emmerich, A.L. Wolf (Eds.), Component Deployment. X, 249 pages. 2004.

Vol. 3076: D. Buell (Ed.), Algorithmic Number Theory. XI, 451 pages. 2004.

Vol. 3074: B. Kuijpers, P. Revesz (Eds.), Constraint Databases and Applications. XII, 181 pages. 2004.

Vol. 3070: L. Rutkowski, J. Siekmann, R. Tadeusiewicz, L.A. Zadeh (Eds.), Artificial Intelligence and Soft Computing - ICAISC 2004. XXV, 1208 pages. 2004. (Subseries LNAI).

Vol. 3066: S. Tsumoto, R. Słowiński, J. Komorowski, J.W. Grzymala-Busse (Eds.), Rough Sets and Current Trends in Computing. XX, 853 pages. 2004. (Subseries LNAI).

Vol. 3065: A. Lomuscio, D. Nute (Eds.), Deontic Logic. X, 275 pages. 2004. (Subseries LNAI).

Vol. 3064: D. Bienstock, G. Nemhauser (Eds.), Integer Programming and Combinatorial Optimization. XI, 445 pages. 2004.

Vol. 3063: A. Llamosí, A. Strohmeier (Eds.), Reliable Software Technologies - Ada-Europe 2004. XIII, 333 pages. 2004.

Vol. 3062: J.L. Pfaltz, M. Nagl, B. Böhlen (Eds.), Applications of Graph Transformations with Industrial Relevance. XV, 500 pages. 2004.

Vol. 3060: A.Y. Tawfik, S.D. Goodwin (Eds.), Advances in Artificial Intelligence. XIII, 582 pages. 2004. (Subseries LNAI).

Vol. 3059: C.C. Ribeiro, S.L. Martins (Eds.), Experimental and Efficient Algorithms. X, 586 pages. 2004.

Vol. 3058: N. Sebe, M.S. Lew, T.S. Huang (Eds.), Computer Vision in Human-Computer Interaction. X, 233 pages. 2004.

Vol. 3056: H. Dai, R. Srikant, C. Zhang (Eds.), Advances in Knowledge Discovery and Data Mining. XIX, 713 pages. 2004. (Subseries LNAI).

Vol. 3054: I. Crnkovic, J.A. Stafford, H.W. Schmidt, K. Wallnau (Eds.), Component-Based Software Engineering. XI, 311 pages. 2004.

Vol. 3053: C. Bussler, J. Davies, D. Fensel, R. Studer (Eds.), The Semantic Web: Research and Applications. XIII, 490 pages. 2004.

Vol. 3052: W. Zimmermann, B. Thalheim (Eds.), Abstract State Machines 2004. Advances in Theory and Practice. XII, 235 pages. 2004.

Vol. 3051: R. Berghammer, B. Möller, G. Struth (Eds.), Relational and Kleene-Algebraic Methods in Computer Science. X, 279 pages. 2004.

Vol. 3050: J. Domingo-Ferrer, V. Torra (Eds.), Privacy in Statistical Databases. IX, 367 pages. 2004.

Vol. 3047: F. Oquendo, B. Warboys, R. Morrison (Eds.), Software Architecture. X, 279 pages. 2004.

Vol. 3046: A. Laganà, M.L. Gavrilova, V. Kumar, Y. Mun, C.K. Tan, O. Gervasi (Eds.), Computational Science and Its Applications - ICCSA 2004. LIII, 1016 pages. 2004.

Vol. 3045: A. Laganà, M.L. Gavrilova, V. Kumar, Y. Mun, C.K. Tan, O. Gervasi (Eds.), Computational Science and Its Applications – ICCSA 2004. LIII, 1040 pages. 2004.

Vol. 3044: A. Laganà, M.L. Gavrilova, V. Kumar, Y. Mun, C.K. Tan, O. Gervasi (Eds.), Computational Science and Its Applications – ICCSA 2004. LIII, 1140 pages. 2004.

Vol. 3043: A. Laganà, M.L. Gavrilova, V. Kumar, Y. Mun, C.J.K. Tan, O. Gervasi (Eds.), Computational Science and Its Applications – ICCSA 2004. LIII, 1180 pages. 2004.

Vol. 3042: N. Mitrou, K. Kontovasilis, G.N. Rouskas, I. Iliadis, L. Merakos (Eds.), NETWORKING 2004, Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications. XXXIII, 1519 pages. 2004.

Vol. 3039: M. Bubak, G.D.v. Albada, P.M. Sloot, J. Dongarra (Eds.), Computational Science - 2004. LXVI, 1271 pages. 2004.

Vol. 3038: M. Bubak, G.D.v. Albada, P.M. Sloot, J. Dongarra (Eds.), Computational Science - ICCS 2004. LXVI, 1311 pages. 2004.

Vol. 3037: M. Bubak, G.D.v. Albada, P.M. Sloot, J. Dongarra (Eds.), Computational Science - ICCS 2004. LXVI, 745 pages. 2004.

Vol. 3036: M. Bubak, G.D.v. Albada, P.M. Sloot, J. Dongarra (Eds.), Computational Science - ICCS 2004. LXVI, 713 pages. 2004.

Vol. 3035: M.A. Wimmer (Ed.), Knowledge Management in Electronic Government. XII, 326 pages. 2004. (Subseries LNAI).

Vol. 3034: J. Favela, E. Menasalvas, E. Chávez (Eds.), Advances in Web Intelligence. XIII, 227 pages. 2004. (Subseries LNAI).

Vol. 3033: M. Li, X.-H. Sun, Q. Deng, J. Ni (Eds.), Grid and Cooperative Computing. XXXVIII, 1076 pages. 2004.

Vol. 3032: M. Li, X.-H. Sun, Q. Deng, J. Ni (Eds.), Grid and Cooperative Computing. XXXVII, 1112 pages. 2004.

Vol. 3031: A. Butz, A. Krüger, P. Olivier (Eds.), Smart Graphics. X, 165 pages. 2004.

Vol. 3030: P. Giorgini, B. Henderson-Sellers, M. Winikoff (Eds.), Agent-Oriented Information Systems. XIV, 207 pages. 2004. (Subseries LNAI).

Vol. 3029: B. Orchard, C. Yang, M. Ali (Eds.), Innovations in Applied Artificial Intelligence. XXI, 1272 pages. 2004. (Subseries LNAI).

Vol. 3028: D. Neuenschwander, Probabilistic and Statistical Methods in Cryptology. X, 158 pages. 2004.

Vol. 3027: C. Cachin, J. Camenisch (Eds.), Advances in Cryptology - EUROCRYPT 2004. XI, 628 pages. 2004.

Vol. 3026: C. Ramamoorthy, R. Lee, K.W. Lee (Eds.), Software Engineering Research and Applications. XV, 377 pages. 2004.

Vol. 3025: G.A. Vouros, T. Panayiotopoulos (Eds.), Methods and Applications of Artificial Intelligence. XV, 546 pages. 2004. (Subseries LNAI).

Vol. 3024: T. Pajdla, J. Matas (Eds.), Computer Vision - ECCV 2004. XXVIII, 621 pages. 2004.

Vol. 3023: T. Pajdla, J. Matas (Eds.), Computer Vision - ECCV 2004. XXVIII, 611 pages. 2004.

Vol. 3022: T. Pajdla, J. Matas (Eds.), Computer Vision - ECCV 2004. XXVIII, 621 pages. 2004.

Vol. 3021: T. Pajdla, J. Matas (Eds.), Computer Vision - ECCV 2004. XXVIII, 633 pages. 2004.

Vol. 3019: R. Wyrzykowski, J. Dongarra, M. Paprzycki, J. Wasniewski (Eds.), Parallel Processing and Applied Mathematics. XIX, 1174 pages. 2004.

Vol. 3016: C. Lengauer, D. Batory, C. Consel, M. Odersky (Eds.), Domain-Specific Program Generation. XII, 325 pages. 2004.

Vol. 3015: C. Barakat, I. Pratt (Eds.), Passive and Active Network Measurement. XI, 300 pages. 2004.

Vol. 3014: F. van der Linden (Ed.), Software Product-Family Engineering. IX, 486 pages. 2004.

Vol. 3012: K. Kurumatani, S.-H. Chen, A. Ohuchi (Eds.), Multi-Agnets for Mass User Support. X, 217 pages. 2004. (Subseries LNAI).

Vol. 3011: J.-C. Régin, M. Rueher (Eds.), Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems. XI, 415 pages. 2004.

Vol. 3010: K.R. Apt, F. Fages, F. Rossi, P. Szeredi, J. Váncza (Eds.), Recent Advances in Constraints. VIII, 285 pages. 2004. (Subseries LNAI).

Vol. 3009: F. Bomarius, H. Iida (Eds.), Product Focused Software Process Improvement. XIV, 584 pages. 2004.

Vol. 3008: S. Heuel, Uncertain Projective Geometry. XVII, 205 pages. 2004.

Vol. 3007: J.X. Yu, X. Lin, H. Lu, Y. Zhang (Eds.), Advanced Web Technologies and Applications. XXII, 936 pages. 2004.

Vol. 3006: M. Matsui, R. Zuccherato (Eds.), Selected Areas in Cryptography. XI, 361 pages. 2004.

Vol. 3005: G.R. Raidl, S. Cagnoni, J. Branke, D.W. Corne, R. Drechsler, Y. Jin, C.G. Johnson, P. Machado, E. Marchiori, F. Rothlauf, G.D. Smith, G. Squillero (Eds.), Applications of Evolutionary Computing. XVII, 562 pages. 2004.

Vol. 3004: J. Gottlieb, G.R. Raidl (Eds.), Evolutionary Computation in Combinatorial Optimization. X, 241 pages. 2004.

Vol. 3003: M. Keijzer, U.-M. O'Reilly, S.M. Lucas, E. Costa, T. Soule (Eds.), Genetic Programming. XI, 410 pages. 2004.

Vol. 3002: D.L. Hicks (Ed.), Metainformatics. X, 213 pages. 2004.

Vol. 3001: A. Ferscha, F. Mattern (Eds.), Pervasive Computing. XVII, 358 pages. 2004.

Vol. 2999: E.A. Boiten, J. Derrick, G. Smith (Eds.), Integrated Formal Methods. XI, 541 pages. 2004.

Vol. 2998: Y. Kameyama, P.J. Stuckey (Eds.), Functional and Logic Programming. X, 307 pages. 2004.

Vol. 2997: S. McDonald, J. Tait (Eds.), Advances in Information Retrieval. XIII, 427 pages. 2004.

Vol. 2996: V. Diekert, M. Habib (Eds.), STACS 2004. XVI, 658 pages. 2004.

Vol. 2995: C. Jensen, S. Poslad, T. Dimitrakos (Eds.), Trust Management. XIII, 377 pages. 2004.

Vol. 2994: E. Rahm (Ed.), Data Integration in the Life Sciences. X, 221 pages. 2004. (Subseries LNBI).

Vol. 2993: R. Alur, G.J. Pappas (Eds.), Hybrid Systems: Computation and Control. XII, 674 pages. 2004.

Vol. 2992: E. Bertino, S. Christodoulakis, D. Plexousakis, V. Christophides, M. Koubarakis, K. Böhm, E. Ferrari (Eds.), Advances in Database Technology - EDBT 2004. XVIII, 877 pages. 2004.

Vol. 2991: R. Alt, A. Frommer, R.B. Kearfott, W. Luther (Eds.), Numerical Software with Result Verification. X, 315 pages. 2004.

Vol. 2990: J. Leite, A. Omicini, L. Sterling, P. Torroni (Eds.), Declarative Agent Languages and Techniques. XII, 281 pages. 2004. (Subseries LNAI).

Vol. 2989: S. Graf, L. Mounier (Eds.), Model Checking Software. X, 309 pages. 2004.

Vol. 2988: K. Jensen, A. Podelski (Eds.), Tools and Algorithms for the Construction and Analysis of Systems. XIV, 608 pages. 2004.

Vol. 2987: I. Walukiewicz (Ed.), Foundations of Software Science and Computation Structures. XIII, 529 pages. 2004.

Vol. 2986: D. Schmidt (Ed.), Programming Languages and Systems. XII, 417 pages. 2004.

Vol. 2985: E. Duesterwald (Ed.), Compiler Construction. X, 313 pages. 2004.

Vol. 2984: M. Wermelinger, T. Margaria-Steffen (Eds.), Fundamental Approaches to Software Engineering. XII, 389 pages. 2004.

Vol. 2983: S. Istrail, M.S. Waterman, A. Clark (Eds.), Computational Methods for SNPs and Haplotype Inference. IX, 153 pages. 2004. (Subseries LNBI).

Vol. 2982: N. Wakamiya, M. Solarski, J. Sterbenz (Eds.), Active Networks. XI, 308 pages. 2004.

Vol. 2981: C. Müller-Schloer, T. Ungerer, B. Bauer (Eds.), Organic and Pervasive Computing – ARCS 2004. XI, 339 pages. 2004.

Preface

Privacy in statistical databases is about finding tradeoffs to the tension between the increasing societal and economical demand for accurate information and the legal and ethical obligation to protect the privacy of individuals and enterprises, which are the source of the statistical data. Statistical agencies cannot expect to collect accurate information from individual or corporate respondents unless these feel the privacy of their responses is guaranteed; also, recent surveys of Web users show that a majority of these are unwilling to provide data to a Web site unless they know that privacy protection measures are in place.

"Privacy in Statistical Databases 2004" (PSD 2004) was the final conference of the CASC project ("Computational Aspects of Statistical Confidentiality", IST-2000-25069). PSD 2004 is in the style of the following conferences: "Statistical Data Protection", held in Lisbon in 1998 and with proceedings published by the Office of Official Publications of the EC, and also the AMRADS project SDC Workshop, held in Luxemburg in 2001 and with proceedings published by Springer-Verlag, as LNCS Vol. 2316.

The Program Committee accepted 29 papers out of 44 submissions from 15 different countries on four continents. Each submitted paper received at least two reviews. These proceedings contain the revised versions of the accepted papers. These papers cover the foundations and methods of tabular data protection, masking methods for the protection of individual data (microdata), synthetic data generation, disclosure risk analysis, and software/case studies.

Many people deserve our gratitude. The conference and these proceedings would not have existed without the Organization Chair, Enric Ripoll, and the Organizing Committee (Jordi Castellà, Antoni Martínez, Francesc Sebé and Julià Urrutia). In evaluating the papers submitted we received the help of the Program Committee and four external reviewers (Jörg Höhne, Silvia Polettini, Yosef Rinott and Giovanni Seri).

We also thank all the authors of submitted papers and apologize for possible omissions.

March 2004

Josep Domingo-Ferrer
Vicenç Torra

Privacy in Statistical Databases – PSD 2004

Program Committee

Rakesh Agrawal (IBM Almaden, USA)
David Brown (ONS, UK)
Jordi Castro (Polytechnical University of Catalonia)
Lawrence Cox (Nat. Center for Health Statistics, USA)
Ramesh Dandekar (EIA, USA)
Stephen Fienberg (Carnegie Mellon University, USA)
Luisa Franconi (ISTAT, Italy)
Sarah Giessing (Destatis, Germany)
Anco Hundepool (Statistics Netherlands, The Netherlands)
Julia Lane (Urban Institute, USA)
Josep M. Mateo-Sanz (Universitat Rovira i Virgili, Catalonia)
William E. Winkler (Census Bureau, USA)
Laura Zayatz (Census Bureau, USA)

Program Chairs

Josep Domingo-Ferrer (Universitat Rovira i Virgili, Catalonia)
Vicenç Torra (IIIA-CSIC, Catalonia)

Organization Committee

Jordi Castellà Roca (Universitat Rovira i Virgili, Catalonia)
Antoni Martínez-Ballesté (Universitat Rovira i Virgili, Catalonia)
Francesc Sebé (Universitat Rovira i Virgili, Catalonia)
Julià Urrutia (IDESCAT, Catalonia)

Organization Chair

Enric Ripoll (IDESCAT, Catalonia)

Table of Contents

Foundations of Tabular Protection

Methods for Tabular Protection

Masking for Microdata Protection

Risk in Microdata Protection

Synthetic Data

Software and Case Studies

Survey on Methods for Tabular Data Protection in ARGUS

Sarah Giessing

Federal Statistical Office of Germany, 65180 Wiesbaden
Sarah.giessing@statistik-bund.de

Abstract. The paper introduces into the methodology for disclosure limitation offered by the software package τ-ARGUS. Those methods have been applied to the data sets of a library of close-to-real-life test instances. The paper presents results of the tests, comparing the performance of the methods with respect to key issues such as practical applicability, information loss, and disclosure risk. Based on these results, the paper points out which of the alternative methods offered by the package is likely to perform best in a given situation.

1 Introduction

Data collected within government statistical systems is usually provided as to fulfil requirements of many users differing widely in the particular interest they take in the data. Data are published at several levels of detail in large tables, based on elaborate hierarchical classification schemes. In many cases, cells of these tables contain information on single, or very few respondents. In the case of establishment data, given the meta information provided along with the cell values (typically: industry, geography, size classes), those respondents could be easily identifiable. Therefore, measures for protection of those data have to be put in place. The choice is between suppressing part of the information (cell suppression), or perturbing the data.

The software τ-ARGUS [13], as emerging from the European project CASC (= Computational Aspects of Statistical Confidentiality) [12], offers methods to identify sensitive cells, a choice of algorithms to select secondary suppressions, programs to compute interval bounds for suppressed cells (audit), and to generate synthetic values to replace suppressed original ones in a publication. Section 2 will introduce into the methods offered (or foreseen to be offered) by the package.

These methods have been applied to data sets of a of a library of close-to-real-life test instances. Section 3 will present empirical results, comparing the performance of the methods with respect to key issues concerning practical applicability, information loss, and disclosure risk.

As a conclusion from the test results, section 4 will provide some guidelines for users, recommending specific methods to apply in certain situations.

2 Methodological Background

τ-ARGUS offers a variety of options for a disseminator to formulate protection requirements which will be discussed in section 2.1. When cell suppression is used as

J. Domingo-Ferrer and V. Torra (Eds.): PSD 2004, LNCS 3050, pp. 1–13, 2004.

disclosure limitation technique, in a first step sensitive cells will be suppressed (*primary suppressions*). In a second step, other cells (so called '*secondary*' or 'complementary' *suppressions*) must be suppressed along with these so called 'primary suppressions' in order to prevent the possibility that users of the published table would be able to recalculate primary suppressions. The problem of finding an optimum set of suppressions is known as the 'secondary cell suppression problem'. τ-ARGUS offers a choice of algorithms to select secondary suppressions as outlined in section 2.2.

By solving a set of equations implied by the additive structure of a statistical table, and some additional constraints on cell values (such as non-negativity) it is possible to obtain upper and lower bounds for the suppressed entries of a table. The package offers to derive the bounds of these so called 'feasibility intervals' (sec. 2.3). Based on ideas of [5], a method for controlled tabular adjustment (CTA) has been implemented to supply users with synthetic values located within those intervals which could be used to replace suppressed original values in a publication (sec. 2.4).

2.1 Formulation of Protection Requirements

τ-ARGUS offers various options to formulate protection requirements. The software can be used to prevent exact disclosure of respondent data only, or to also avoid inferential disclosure to some degree.

When it is enough to prevent exact disclosure of respondent data, users of τ-ARGUS specify the parameter *n* of a minimum frequency rule. In that case, secondary suppressions would be selected in such a way that the width of the feasibility interval for any sensitive cell is non-zero, i.e. the interval does not contain the true cell value only. When it is not enough to prevent exact disclosure, but the risk of approximate disclosure must also be limited, users of τ-ARGUS specify parameters of the p%-rule, or dominance rule. The goal is to find a set of secondary suppressions ensuring that the resulting bounds of the feasibility interval of any sensitive cell cannot be used to deduce bounds on an individual respondent contribution that are too close according to the sensitivity criterion employed. Results of [3,4] can be used to compute the so called 'protection level'. Bounds of the feasibility interval must not be closer than the protection level. Formulas corresponding to the p %- and (n,k)-dominance rule are given in table 3 of the appendix.

It should, however, be mentioned here that some problems have not yet been fully solved in the current version of τ-ARGUS: With some of the secondary cell suppression algorithms offered, it may happen that the algorithm considers a single respondent cell to be properly protected even though there is only one other suppression in the same row/column/... of the table and this suppression is another single respondent cell. This may cause disclosure: the respondent contributing to either of the two single respondent cells will be able to recover the value of the other single respondent. Similar problems of exact disclosure may arise with single respondent cells which are not in the same row/column/... of the table. Another problem still unsolved for any algorithm offered by the package, is the problem of assigning protection levels in such a way that aggregates published implicitly (so called 'multi-cells') in a protected table will always be non-sensitive.

2.2 Algorithms for Secondary Cell Suppression

The goal of secondary cell suppression is to find a valid suppression pattern satisfying the protection requirements of the sensitive cells (see 2.1 above), while minimizing the loss of information associated with the suppressed entries. The 'classical' formulation of the secondary cell suppression problem is a combinatorial optimisation problem, which is computationally extremely hard to solve. τ-ARGUS offers a variety of algorithms to find a valid suppression pattern even for sets of large hierarchical tables linked by linear interrelations. It is up to the user to trade-off quality vs. quantity, that is to decide how much resources (computation time, costs for extra software etc.) he wants to spend in order to improve the quality of the output tables with respect to information loss. The package offers a choice basically between four different approaches:

OPTIMAL Fischetti/Salazar methodology aims at the optimal solution of the cell suppression problem [8]. A feasible solution is offered at an early stage of processing, which is then optimised successively. It is up to the user to stop execution before the optimal solution has been found, and accept the solution reached so far. The user can also choose the objective of optimisation, i.e. choose between different measures of information loss. Note that the method relies on high performance, commercial OR solvers.

MODULAR The *HiTaS* method [7] subdivides hierarchical tables into sets of linked, unstructured tables. The cell suppression problem is solved for each subtable using Fischetti/Salazar methodology [8]. Backtracking of subtables avoids consistency problems when cells belonging to more than one subtable are selected as secondary suppressions.

NETWORK The concept of an algorithm based on *network flow methodology* has been outlined in [1]. Castro's algorithm aims at a heuristic solution of the CSP for 2-dimensional tables. Network flow heuristics are known to be highly efficient. It may thus turn out that the method is able to produce high quality solutions for large tables very quickly. τ-ARGUS offers an implementation applicable to 2-dimensional tables with hierarchical substructure in one dimension. A license for a commercial OR solver will not be required to run the algorithm.

HYPERCUBE The *hypercube algorithm* GHM*ITER* developed by R.D. Repsilber ([see 5,6]) is a fast alternative to the above three OR based methods. This heuristic is able to provide a feasible solution even for extremely large, complex tables without consuming much computer resources. The user, however, has to put up with a certain tendency for over-suppression.

SINGLETON Special application of GHM*ITER*, addressing only the protection of single respondent cells. The method is meant to be used as preprocessing for the OPTIMAL and NETWORK methods for which a solution for the problem with single respondent cells mentioned in sec. 2.1 has not yet been implemented.

With respect to the hypercube and the modular method, both involving backtracking of subtables, it should be noted that such methods are not '*global*'. This causes a certain disclosure risk (see [4] for problems related to non-global methods for secondary cell suppression.).

2.3 Audit: Computing the Feasibility Intervals

By solving a set of equations implied by the additive structure of a statistical table, and some additional constraints on cell values (such as non-negativity) it is possible to obtain upper and lower bounds for the suppressed entries of a table. These bounds are solutions to the following linear programming problem (c.f. [8]):

$$\begin{array}{l} \text{Min } y_i \quad \text{, and} \quad \text{Max } y_i \quad \text{subject to} \\ \sum_{i\in I} m_{ij} y_i = b_j \quad , j \in J \\ lb_i \le y_i \le ub_i \quad , i \in P \cup S \\ y_i = a_i \quad , i \notin P \cup S \end{array}$$

where the additive structure of the table is given by the set of linear equations $\sum_{i\in I} m_{ij} y_i = b_j \quad , j \in J$ (typically $b_j = 0$, and $m_{ij} \in \{-1,0,1\}$). *I*, *P*, and *S* denote the set of all cells, of the sensitive cells, and of the secondary suppressions, respectively, and ub_i, lb_i are constraints on the cell values a_i. τ-ARGUS assumes $ub_i - a_i = a_i - lb_i = q \cdot a_i$. By default, the parameter *q* is set to 1.

2.4 Controlled Tabular Adjustment for Tables with Suppressions

The authors of [5] suggest controlled tabular adjustment (CTA) to compute synthetic values which could be used to replace suppressed original ones in a publication. The idea of CTA is to determine synthetic values that are 'as close as possible'[1] to the original ones for the non-sensitive cells, but at some 'safe' distance for the sensitive cells.

In the following, we consider a variant of this approach: As the idea of CTA is still fairly new, the method not yet established as a standard for tabular data protection, we thought it could be a natural way of familiarizing those who are used to tables protected by cell suppression with the new methodology, if it is presented as 'just to release some additional information' on the suppressed entries. Therefore, while with CTA methods suggested so far (see [5, 6, 2, 11]) all cells are candidates for adjustment, in our variant adjustment is restricted to the suppressed cells of a protected table. Synthetic values are then obtained as solution to the following LP problem:

$$\text{Min} \quad \sum_{i\in P^+} w_i y_i^+ + \sum_{i\in P^-} w_i y_i^- + \sum_{i\in S} w_i (y_i^+ + y_i^-) + W\left(\sum_{i\in P^-} y_i^+ + \sum_{i\in P^+} y_i^- \right)$$

subject to:

$$\begin{array}{l} \sum_{i\in P\cup S} m_{ij} (y_i^+ - y_i^-) = b_j - \sum_{i\notin P\cup S} m_{ij} a_i \quad , j \in J \\ 0 \le y_i^+ \le \max_i - a_i \ , \ i \in P \cup S \\ 0 \le y_i^- \le a_i - \min_i \ , \ i \in P \cup S \\ y_i^+ \ge upl_i \ , \ i \in P^+ \\ y_i^- \ge lpl_i \ , \ i \in P^- , \end{array}$$

[1] Using a L1 distance.

where $\max_i$ and $\min_i$ denote the solutions to the LP problem of section 2.3, w_i are weights obtained by a cost function such as $w_i = \frac{1}{1+a_i}$, W is a very large constant, and upl_i and lpl_i are upper and lower protection levels for the sensitive cells computed according to 2.1. Synthetic values for $i \in P \cup S$ are then defined as $a_i + y_i^+ - y_i^-$.

P^+ and P^-, the sets of sensitive cells which are adjusted to (or beyond) their upper (P^+) or lower (P^-) protection level, are determined in advance of solving the LP problem by a simple heuristic as outlined in [5]: Considering protection level and cell size, cells on the lowest level of the table are allocated to P^+ and P^- in alternating sequence. For allocation of higher level sensitive cells, we consider allocation of corresponding lower level sensitive cells. Unfortunately, our simple heuristic tends to give solutions where both, y_i^+, and y_i^- are positive for a few of the sensitive cells, i.e. where the synthetic cell value will be too close to the true cell value. For discussion of this problem, and a suggestion how to solve it, see [2].

3 Application

The methods described in section 2 have been applied to data sets of a library of close-to-real-life test instances. Section 3.1 explains the test scenario. Section 3.2. compares the performance of the alternative cell suppression algorithms. Results of further processing (audit and CTA) are presented in section 3.3.

3.1 Data Sets

A synthetic datafile has been constructed based on typical real-life structural business data. The algorithm used for generating the synthetic data has been designed as to preserve those properties of typical tabulations of the data relevant for cell suppression, i.e. structures of variables, location of sensitive and zero cells, cell sensitivity, and number of contributions for low frequency cells. The file consists of nearly 3 mio records. It offers three categorical (i.e. explanatory) variables, and a variety of response variables one of which was chosen for the applications[2]. Of the categorical variables, one offers a (7-level) hierarchical structure. For some of the tabulations, only one of the non-hierarchical variables was considered. The depth of the hierarchical variable was varied. In this way six tables were generated, three 2-dimensional and three 3-dimensional ones with size (i.e. number of cells) varying between 460 000 and 150 000 cells. A p%-rule was employed for primary suppression. See table 4 in the appendix for details.

[2] Due to technical problems with the current version of τ-ARGUS the last 5 digits of the variable which had up to 15 digits were dropped in advance of tabulation.

Except for the network flow method which was not yet available for application to hierarchical tables, all the algorithms listed in 2.2 have been applied to the six tables described above. Unfortunately, for the largest table (table 6) only the hypercube method ended properly. A run with the modular method could not be completed, runs with the optimal methods were not even attempted because of the expected exhaustive CPU usage (several days). Using CPLEX 7.5 as OR-solver, runs were carried out on a Windows NT PC, Intel Pentium III processor, 261 MB Ram.

While for the 2-dimensional tables processing times were short enough to be of no concern, for the 3-dimensional tables, application of Linear Programming based methods took considerably more time than a run of the hypercube method. With increasing depth of hierarchical structure, the effectiveness of the modular implementation (compared to 'Optimal') regarding reduction of execution time grows: for the 4-levels table 5, execution time is reduced by 9 hours (from 12h10 to 3h11), while for the 3-levels table 4 the reduction is only 21 minutes (from 1h33 to 1h12). See table 5 (appendix) for details.

As mentioned in 2.2, with method 'Optimal' it is up to the user to stop execution before the optimal solution has been found, and accept the solution reached so far. It should be noted that we actually made use of this option. Thus, not all the suppression patterns generated by this method can be considered truly 'optimal'.

3.2 Results of Secondary Cell Suppression

This section compares performances of the algorithms on the test tables with respect to number and added values of the secondary suppressions. Concerning the LP-based methods (Mod, Opt, Si/Opt), results presented in table 1 below were obtained when using the response variable as cost function.

Table 1. Information Loss due to Secondary Suppression.

Table	Hier. Levels	No Cells	No Suppressions (%)				Added Value of Suppressions (%)			
			Hyp	Mod	Opt	Si/Opt	Hyp	Mod	Opt	Si/Opt
			2-dimensional tables							
1	3	460	6.96	4.35	4.78	7.61	0.18	0.05	0.03	0.05
2	4	1050	10.95	8.29	7.43	11.52	0.98	0.62	0.58	0.71
3	6	8230	14.92	11.48	15.36	17.97	6.78	1.51	1.64	2.06
			3-dimensional tables							
4	3	8280	14.63	10.72	14.96	16.44	6.92	1.41	0.63	1.58
5	4	18900	17.31	15.41	19.19	20.00	12.57	3.55	2.32	4.48
6	6	148140	15.99	-	-	-	23.16	-	-	-

With respect to the number of secondary suppressions, method 'Modular' performed best on all tables except for table 2, where method 'Optimal' suppressed 9 cells less. Except for the 2 smallest tables, 'Optimal' with cell value as cost function performs even worse than the hypercube method.