STATISTICAL DISCLOSURE CONTROL IN CATALONIA AND THE CRISES GROUP

Josep Domingo
Departament d'Enginyeria Química, Estadística i I.O.
ETSE
Universitat Rovira i Virgili

Statistical offices release two kinds of data through their statistical databases: tabular data and microdata sets (individual respondent records). In both cases, data dissemination should be performed in a way that does not lead to disclosure of individual information but preserves the informational content as much as possible. This is known as the statistical disclosure control (SDC) problem. While there is a long experience in table dissemination, microdata dissemination is a much more recent activity.

Data security, including statistical disclosure control, is the research topic of the CRISES group, based at the Universitat Rovira i Virgili, Tarragona, Catalonia (http://www.etse.urv.es/recerca/crises). In English CRISES stands for «CRyptography and Inference Surveillance in Electronic Systems»; in Catalan it stands for «CRiptografia i Secret EStadístic» (Cryptography and Statistical Data Confidentiality).

Although the activity of CRISES includes other data security topics beyond SDC (cryptography, secure e-commerce, etc.), the next sections focus on the contributions made by CRISES and other related Catalan research groups to the field of statistical disclosure control. Section 1 below describes Catalan activity on SDC during the period 1993-1999. Section 2 reports on U.S. government funded project OTTILIE (1999-2000) for assessment of microdata protection methods. Finally, Section 3 is an update on current research being carried by Catalan researchers under European 5th FP projects CASC and AMRADS.

Catalan activity in SDC (1993-1999): Probably the first Catalan paper published on SDC was by an IDESCAT researcher back in J. Turmo, 1993. Based on that paper and subsequent work, IDESCAT released a sample of the 1991 population census of Catalonia (IDESCAT, 1995). Sampling was the procedure used to control the risk of an individual being reidentified (Garín and Ripoll, 1999). In the period 1996-1999, CRISES was the only research group in Catalonia doing work on SDC, supported by CICYT project TIC95-0903-C02-02 and by two IDESCAT research contracts. Work done in this period encompasses both tabular data and microdata.

SDC applied to tabular data has the goal of avoiding exact disclosure of an individual attribute. For contingency or magnitude tables, this means that cells to which a small number of individuals contribute should be specially protected. Cell suppression is the most common approach to tabular data protection. In Domingo-Ferrer and Mateo-Sanz (1996), showed that cell suppression can fail to meet its security goals when the protec-

ted tables contain one or more quantitative factors. In Domingo-Ferrer and Mateo-Sanz (1999), a computationally efficient method for tabular data protection based on resampling proposed.

Regarding SDC for microdata, the CRISES group concentrated on microaggregation, which is a family of methods for «masking» (i.e. protecting) an original microdata set and turning it into a publishable protected microdata set. The rationale behind microaggregation is that confidentiality rules in use allow publication of microdata sets if the data records correspond to groups of k or more individuals, where no individual dominates (i.e. contributes too much to) the group and k is a threshold value. Strict application of such confidentiality rules leads to replacing individual values with averages computed on small aggregates (microaggregates) prior to publication. This is the basic principle of microaggregation. The microaggregation problem consists of finding the k-partition (partition of the set of records into groups of size at least k) that minimizes the within group sum-of-squares: this can be seen as minimizing the information loss caused by the protection process, because the more homogeneous is a group, the less information loss is caused when replacing the group values by their average. In Mateo-Sanz and Domingo-Ferrer (1999), introduced data-oriented microaggregation, where the size of group can vary depending on data, with the only constraint that it be $\geqslant k$ (prior proposals required that all groups be of size k). Multivariate microaggregation was described in Mateo-Sanz and Domingo-Ferrer (1998). Optimal solutions to the microaggregation problem and heuristic solutions were the topic of Domingo-Ferrer and Mateo-Sanz (2002).

Another topic that was investigated at that time was delegation of statistical data. Delegation seeks to allow a data user to obtain exact statistical results without disclosing to him the exact value of data (see Domingo-Ferrer and Sánchez del Castillo, 1997 and the patent Domingo-Ferrer and Sánchez del Castillo, 1998).

A specially visible action by CRISES in this period was the organization of the *Statistical Data Protection'98* conference sponsored by Eurostat (Lisbon, Portugal, March 1998). See Domingo-Ferrer and Mateo-Sanz, 1998, Domingo-Ferrer (1999).

The OTTILIE project: In 1999 a one-year project called OTTILIE (*Optimizing the Tradeoff beTween Information Loss and dIsclosurE risk for microdata*) was awarded to the CRISES group and to a researcher at IIIA-CSIC (Institut d'Investigació en Intel·ligència Artificial) by the U. S. Bureau of the Census.

OTTILIE was a twelve-month project that sought to demonstrate a methodology to compare the existing methods for masking (SDC-protecting) microdata. In SDC, there is a tension between the information loss in by application of a particular masking method to an original dataset and the disclosure risk associated to the released masked dataset (probability that masked data may lead to disclosure of individual information). OTTILIE defined procedures to measure information loss and disclosure risk; then these measures were combined to construct an overall score for a masking method. Experi-

ments with a wide range of microdata masking methods were carried out and eventually a ranking of methods by score was produced (Domingo-Ferrer, Mateo-Sanz and Torra, 2001, Domingo-Ferrer and Torra, 2001). For continuous microdata, rank swapping and microaggregation were singled out as particularly well-performing masking methods (a method was defined to perform well if it scored low, i.e. if it could achieve low information loss and disclosure risk at the same time). For categorical microdata, none of the tried methods clearly outperformed the rest.

The European projects CASC and AMRADS: On January 1, 2001, the European projects CASC («Computational Aspects of Statistical Confidentiality», IST-2000-25069) and AMRADS («Accompanying Measure for Research and Development in Statistics», IST-2000-26125) were launched. Both projects span over three years and have connections with SDC.

The CASC project aims at perfecting the ARGUS SDC software (Argus, 1999) by enriching it with more methods for protecting tabular data and microdata. Four Catalan contractors belong to the CASC consortium: IDESCAT, UPC, IIIA-CSIC and URV (the CRISES group in the latter case). Their roles are as follows:

- IDESCAT will essentially take a user role and will provide testing for methods being developed.
- UPC is committed to work on tabular data protection, by exploring new heuristics for the secondary cell suppression problem based on network. See Castro (2001).
- IIIA-CSIC concentrates on microaggregation for categorical variables (Domingo-Ferrer and Torra, 2001) and also on new record linkage methods for empirical assessment of disclosure risk (Torra, 2000).
- CRISES-URV is working on microdata protection, specifically:
  - Implementing a library with all known microaggregation methods, including those proposed by Domingo-Ferrer and Mateo-Sanz (2002), in view of inclusion in the Argus software.
  - Characterizing the complexity of optimal microaggregation. An achievement has been to prove NP-hardness for optimal microaggregation (Oganian and Domingo-Ferrer, 2001).
  - Exploring microdata protection through synthetic data generation. Work is in progress to compare the best masking methods identified in Domingo-Ferrer and Torra, 2001) with new approaches based on synthetic data.
  - Providing mechanisms for multilevel access to microdata. Water-marking techniques are exploited to provide multilevel access to masked data: the higher the clearance of a user, the more masking she can remove (unprivileged users just see masked data, while top privileged users can retrieve the original data, see Domingo-Ferrer, Mateo-Sanz and Sebé, 2001).

Finally, the AMRADS projects intends to disseminate best practices in official statistics. The CRISES group is a subcontractor for this project in charge of organizing the AMRADS Workshop on Statistical Data Confidentiality (Luxembourg, December 13-14, 2001).

References

1. Hundepool, A. and Willenborg, L. (1999). «ARGUS: Software from the SDC project», *Joint ECE/Eurostat Work Session on Statistical Confidentiality*, Thessaloniki, Greece. http://www.unece.org/stats/docments/1999.03.confidentiality.htm
2. Castro, J. (2001). «Using modeling languages for the complementary suppression problem through network models», *2nd Joint ECE/Eurostat Work Session on Statistical Confidentiality*, Skopje, Macedonia. http://www.unece.org/stats/documents/2001/03/confidentiality/24.e.pdf
3. Domingo-Ferrer, J. and Mateo-Sanz, J. M. (1996). «On the security of cell suppression in contingency tables with quantitative factors», in *Proceedings of the 3rd International Seminar on Statistical Confidentiality*, Ljubljana: Eurostat-Statistical Office of the Republic of Slovenia, 208-217.
4. Domingo-Ferrer, J. and Sánchez del Castillo, R. X. (1997). «An implementable scheme for secure delegation of statistical data», a *Information Security-ICICS'97* (Lecture Notes in Computer Science 1334), eds. Y. Han, T. Okamoto and S. Qing, Berlin: Springer-Verlag, 445-451.
5. Mateo-Sanz, J. M. and Domingo-Ferrer, J. (1998). «A comparative study of microaggregation methods», *Qüestiió*, 22, 511-526.
6. Domingo-Ferrer, J. and Sánchez del Castillo, R. X. (2000). «A method for secure delegation of statistical data», Spanish patent P9800608, filed in the *Spanish Boletín Oficial de la Propiedad Industrial*, 16, 4559.
7. Domingo-Ferrer, J. and Mateo-Sanz, J. M. (1998). «Current directions in statistical data protection», *Research in Official Statistics*, 2, 105-112.
8. Domingo-Ferrer, J. and Mateo-Sanz, J. M. (1999). «On resampling for statistical confidentiality in contingency tables», *Computers & Mathematics with Applications*, 38, 13-32.
9. Domingo-Ferrer, J. (ed.) (1999). *Statistical Data Protection*, Luxembourg: Office for Official Publications of the European Communities.
10. Domingo-Ferrer, J., Mateo-Sanz, J. M. and Torra, V. (2001). «Comparing SDC methods for microdata on the basis of information loss and disclosure risk», *Proceedings of ETK-NTTS'2001*, 2, Luxembourg: Office for Official Publications of the European Communities.
11. Domingo-Ferrer, J. and Torra, V. «A quantitative comparison of disclosure control methods for microdata», in *Confidentiality, Disclosure and Data Access*. North-Holland, 113-135 (to appear).

12. Domingo-Ferrer, J., Mateo-Sanz, J. M. and Sebé, F. (2001). «Watermarking for multilevel access to statistical databases», in *IEEE International Conference on Information Technology: Coding and Computing-ITCC'2001*, Piscataway NJ: IEEE Computer Society, 243-247.
13. Domingo-Ferrer, J. and Torra, V. «Aggregation techniques for statistical confidentiality», in *Aggregation Operators: New Trends and Applications*, (eds. R. Mesiar and T. Calvo), Physica Verlag (to appear).
14. Domingo-Ferrer, J. and Mateo-Sanz, J. M. (2002). «Practical data-oriented microaggregation for statistical disclosure control», *IEEE Transactions on Knowledge and Data Engineering* (to appear). Preprint available at http://www.computer.org/tkde/STATISTICAL DATABASE.htm.
15. IDESCAT-Statistics Catalonia. (1996). *Sample of 1991 Population Census of Catalonia.*
16. Garín, A. and Ripoll, E. (1999). «Performance of μ-Argus in disclosure control of uniqueness in populations», *Joint ECE/Eurostat Work Session on Statistical Confidentiality*, Thessaloniki, Greece. http://www.unece.org/stats/docments/1999.03.confidentiality.htm
17. Mateo-Sanz, J. M. and Domingo-Ferrer, J. (1999). «A method for data- oriented multivariate microaggregation», in *Statistical Data Protection*, ed. J. Domingo-Ferrer, Luxembourg: Office for Official Publications of the European Communities, 89-99.
18. Oganian, A. and Domingo-Ferrer, J. «On the complexity of optimal microaggregation for statistical disclosure control», *UNECE Statistical Journal* (to appear).
19. Torra, V. (2000). «Towards the re-identification of individuals in data files with non-common variables», in *European Conference on Artificial Intelligence*, Amsterdam: IOS Press.
20. Turmo, J. (1993). «Pertorbacions aleatòries amb compensació: una tècnica per a la protecció d'informació estadística confidencial», *Qüestiió*, 17, 413-435.