STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN COMMUNITIES

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

Joint ECE/Eurostat Work Session on Statistical Data Confidentiality
(Skopje, The former Yugoslav Republic of Macedonia, 14-16 March 2001)

Working Paper No. 14
English only

Topic II: Impact of new technological developments in software, communications and computing on SDC

DISCLOSURE CONTROL FOR DEMOGRAPHIC STATISTICS

REDEFINED GUIDELINES AND DEVELOPMENT OF METHODS AT STATISTICS FINLAND

Contributed paper

Submitted by Statistics Finland[1]

I. INTRODUCTION

1. The majority of statistics produced by Statistics Finland are created by employing different government registers. For example, the Population Census has been done in Finland since 1990 using about 30 different registers. Joint use of registers is facilitated by various standardised personal, building and area codes. Finland's register-based census data can also be regarded as particularly flexible geographic information material, because statistical units can be located at the accuracy of one building. This is possible by means of map coordinates defining the location of each building.

2. Finland's Statistics Act and Personal Data Act define the main framework for the processing of sensitive statistical material. In addition to these, Statistics Finland has prepared redefined guidelines for data security and data protection.

3. The data obtained by Statistics Finland for statistical purposes are mainly confidential (Statistics Act, Section 17). The Statistics Act (Section 13) obliges securing of data protection and application of good statistical practices in data processing. Such methods are to be used in data processing that render it impossible to identify the subject of the data from the resulting statistical data. Likewise, unintentional disclosure of confidential data must be prevented in every possible way.

4. Development of data processing and lower prices of necessary equipment and programmes have increased the use and requirements of statistical data. At the same time, the goal has been to produce a wider variety of statistics faster than before and data protection has become an even greater challenge than ever. Demand for small area statistics and geographic information has grown. The definition of the target area, address or building on the map has increased the risk of identifying the statistical units involved. Electronic dissemination of statistics, even as multi-dimensional statistical databases, has made it easier for users to identify statistical units if they so want. The Internet as a distribution channel has given all citizens easy access to a growing number of statistics. Citizens' right to information has increased but private persons' right to data protection may have become impaired at the same time.

5. It is clear that old directions concerning data protection have in recent years become insufficient for many statistical agencies. Statistics Finland created two data protection work groups during 2000 to prepare redefined guidelines for protection of business data as well as of personal data.

[1] Prepared by Marja Tammilehto-Luode.

6. This paper presents the proposals set forth by the personal data protection workgroup for redefined data protection guidelines. In addition, a new data protection method under development is presented. This method is chiefly based on distribution of detailed geographic information.

II. STATISTICS FINLAND'S REDEFINED DATA PROTECTION GUIDELINES

II.1 Revision of data protection guidelines

7. The implementation of the statistical data protection procedure is guided by three central principles:

(i) Legislation;
(ii) Ethics of statistics;
(iii) Reliability of the statistical agency.

8. The most important act in Finland controlling statistics production and related data collection, processing and release is the Statistics Act from 1994. A second significant act, the Personal Data Act, took effect in Finland in 1999, replacing the earlier Personal Data File Act. The Personal Data Act regulates, in particular, all processing of personal data, including protection of such data. The Statistics Act stipulates that no person's privacy, or business or professional secret should be endangered during the processing, storage or destruction of the data collected for statistical purposes.

9. The professional ethics of statistics define the values and operating principles relating to the field of statistics in more detail than legislation. The professional ethical norms of statistics are described in Statistics Finland's guide on professional ethics (1993), which is based, for example, on the International Statistical Institute ISI's guide on professional ethics (1985), on the recommendation of the Council of Europe concerning the protection of personal data used for statistical purposes (1997), and on the fundamental principles of official statistics approved by the ECE (1992). According to the norms regarding data release, it should not be possible to use statistics published or released to customers to find out information concerning individual statistical units.

10. The reliability of the statistical authority is based on the principles presented above, that is, on compliance with the Statistics Act and ethical norms. Reliance on Statistics Finland is a prerequisite for the success of Statistics Finland's data collection and similarly for the production of reliable and relevant statistical data. Even a slight blemish in Statistics Finland's reputation in this respect may lead to significant non-response to inquiries and endanger its right to use registers.

11. The above principles and their application are redefined in Statistics Finland's internal guidelines, which are mainly intended for statistics production supporting conventional paper publication. According to an inquiry sent to various units in 2000, data protection is in practice implemented in different ways in Statistics Finland's different units. The implementation has been dependent on the operational environment in which the statistics are compiled. Such environmental factors are institutional and also technical matters relating to the collection, nature and use of the data. According to the inquiry, almost all statistical units employed some agreed limits that define the smallest figures that can be published. However, these limits vary from one unit to another and with respect to different statistical materials. Protection was often solved for each specific case.

12. Statistics Finland set up two work groups during 2000 to prepare a revision and/or redefinition of the data protection guidelines for business data applied to statistical tables and also for statistical materials based on personal data. The aim was to chart the problems, to make proposals for improving the situation and to introduce revised confidentiality regulations.

13. Both workgroups drew up revised data protection guidelines listing the key principles and some more detailed directives. The objective of the data protection group concerned with population statistics was to collect all central principles and clarify and combine previous guidelines. The purpose was also to revise the guidelines by further specifying their application. Special attention was paid to new kinds of

dissemination channels and file forms. The guidelines were not very detailed but the workgroup recommended their further specification and documentation with respect to different areas of statistics.

14. The workgroup dealing with data protection of personal data paid particular attention to the growing risk of indirect identification when statistics are distributed in electronic format and/or as multi-dimensional databases. Increasing use of small area statistics has also heightened the data protection risk. More frequent use of geographic information management does not as such increase the data protection risk, but along with it, demands for more detailed statistics have grown. The work group is of the opinion that tabular data are close to microdata especially when a multi-dimensional table is formed with detailed classifications, or when the area unit of the table is small and there are many variables, such as in some grid materials. Then it has to be considered separately for each case if the material has to be examined as individual level material, in which case its release is subject to permission.

15. The data protection group recommended that person-based materials be divided into three groups according to the sensitivity of the variables used:

i) The least sensitive materials which contain only demographic basic variables, such as age, sex, family size, age or sex distribution;
ii) Somewhat sensitive materials which contain other than the least sensitive variables or sensitive variables;
iii) Sensitive materials which contain sensitive data as defined in the Personal Data Act (Sections 11 to 12) and describe or are meant to describe:

– race or ethnic origin,
– a person's social, political or religious convictions or trade union membership,
– a criminal act, punishment or other sanction for an offence,
– health, illness or disability of a person, or treatment or comparable measures to which he/she has been subjected,
– a person's sexual orientation or behaviour,
– a person's need for social welfare, or the social welfare services, support or benefits received.

16. The data protection measures are primarily dependent on the group to which the variables of the statistical material belong. The risk of disclosing individual data is thus mainly dependent on the sensitivity of the variables and the number of cases in different table cells. For example, the size of the target population, the number of the variables and the size of the area and the accuracy of its location also have an effect on the risk of disclosure.

II.2 Implementation of new guidelines

17. The personal data protection work group submitted its final report containing the revised guidelines for the protection of tabular personal data on 31 December 2000. The work group recommended approval of the guidelines by Statistics Finland's Management Group, after which they would be distributed to all units. The statistical directors of the units processing personal data would then be responsible for the implementation of the guidelines in accordance with the recommendations. The guidelines still need to be specified with regard to individual statistics. The work group advised that all data protection measures of person-based table materials be written down with explanations for each area of statistics and/or for each unit of responsibility.

18. The operational units should see to it that defining the risk of disclosing personal data would become an essential part of the statistics production process. Means for defining the disclosure risk and methods for implementing data protection should be developed in each unit together with the Information Technology Services and the Statistical R&D unit. The objective is to implement the procedures supporting these recommendations programmatically (automatically and/or interactively) in connection with the production process of each statistics by the end of 2001.

19. The work group also proposed some changes to information service agreements relating to possible data encryption measures, user right limits and data description. It was also put forward that Statistics Finland's Administrative and Legal Services unit would find out whether it would be possible

to add to Finland's Statistics Act a principle stating that no intentional attempts should be made to identify personal data in statistical material by means of other figures in a table, for example (cf. Sweden's Act concerning Official Statistics, Section 10, SFS 1992:889).

III. DEVELOPMENT OF DATA PROTECTION OF SMALL AREA STATISTICS

III.1 Development project

20. Demand for small area statistics has increased continuously during the last five years. Data users require ever more versatile information on ever smaller areas. Increase in the use of Geographic Information Systems and the larger use of other data processing methods in a wide range of applications sets new demands on statistics. To improve the quality of statistical materials and to develop the service, Statistics Finland started a joint project with the Department of Statistics of the University of Jyväskylä. One of the aims was to clarify the data protection problem in geographic information materials and to present possible solutions.

21. A problem for data protection of small area statistics is mainly statistical areas on which there are only a few observations. Statistics Finland does not usually release data on such areas. The method of "missing data" used at the moment produces incomplete material and thus weakens the quality of statistics. The data protection quality of the used methods, that is, identifiability of individual cases, cannot either be estimated explicitly (only probabilities under certain conditions).

22. The research project aims to develop methods that will both improve data protection and the quality of materials released to customers. The starting point is that data protection of the material is not based on the selection of a single statistical data protection method, but it is a comprehensive procedure where several factors have to be considered and a number of decisions be made. These decisions are founded both on the requirements set by customers on the data material and consideration given to data protection of individuals.

III.2 Data protection method of geographic information material

23. The local restricted imputation method (LRI) (Markkula 1999) developed in a research project between the University of Jyväskylä and Statistics Finland is intended especially for data protection of areally aggregated geographic information. The usual data protection methods are not directly applicable to the spatial material or they change it so that the spatial relations are distorted (they disturb the data with respect to the information included in the geographic information). The method suggested thus takes into account not only small area units but also the spatial relations of the data. The method hence gives particular attention to the special characteristics of geographic information material required by the applications. Such characteristics include the modifiability of the units, natural hierarchy and spatial relations of the data.

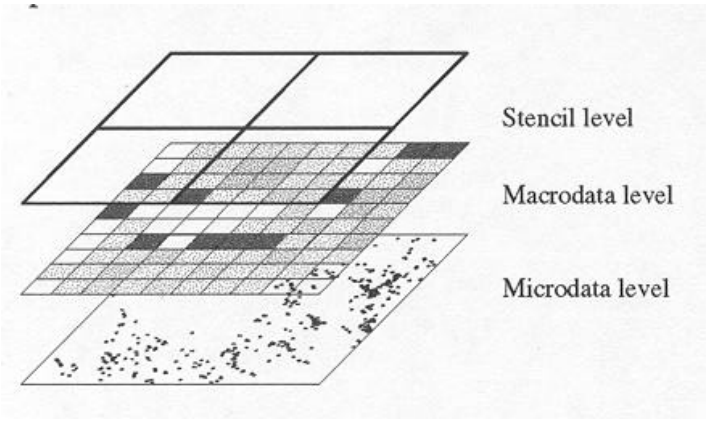
24. The LRI method changes the original material as little as possible. The effect of the method on the statistical distribution of the variables of the material is small. The purpose is to maintain the accuracy level of the material. The material produced as a result of data protection is locally consistent. Following from the local consistency, the data are accurate in the larger area unit level. The spatial relations are mostly retained.

25. The LRI is implemented in three stages: 1) the definition of area division, spatial configuration and the selection of the so-called stencil areas, 2) the definition of the risk areas, and 3) the processing of the risk areas by means of spatial imputation.

26. Defining the area division of the material is as such an independent data protection method. The size and number of the areas have a direct impact on the number of the risk areas, that is, the number of the areas requiring data protection. The size of the areas can be defined optimally for data protection. The areas falling below a certain threshold value are defined as risk areas. The threshold value can be the smallest permitted number of observations in the area. The number of risk areas produced by different area divisions can be estimated from the sample by applying the Poisson-Gamma model. The stencil

areas are hierarchically higher level areas where the distribution of the variables corresponds to the original material on that aggregation level. The stencil definition determines the locality for the following operations.

Figure 1: The spatial configuration of the data



27. After fixing the area division, there remains a group of risk areas not fulfilling the data protection conditions. In a conventional method these risk areas would be suppressed, which would decrease the information included in the material. In the LRI method new values are calculated for all defined risk areas. The calculation is made locally, for each stencil area at a time. The method can be called spatial imputation as the method also takes account of the spatial information of the material.

28. Many different methods can be used in imputation. Two simple basic methods are imputation by mean and imputation by random permutation.

29. Local imputation by mean is a method where the detected cell frequency of the risk areas is replaced by imputed frequency based on the mean of risk areas included in each stencil area. Local imputation by mean does not change the mean of the whole material but it reduces the variance. Local imputation by random permutation is a method where the imputed frequency is produced by dividing the detected frequencies belonging to the stencil group by probabilities e.g. according to the rectangular distribution into risk areas. Local imputation by random permutation has no effect on the mean and variance of the target population.

30. Imputation by mean corresponds to the microaggregation method applied to non-spatial material and the principles of imputation by random permutation are close to the data swapping method of non-spatial materials.

III.3 Implementation of the new method

31. In the development of the LRI method a test material was used which comprised an area of 20 km x 20 km divided into grid squares of 1 km x 1 km. In this area, 201 of 400 grid squares of 1 km^2 were inhabited by altogether 2,790 people. A grid square of 5 km x 5 km was chosen as the stencil area. Table 1 compares the results after imputation by mean and imputation by random permutation to a situation if the risk areas in the material had simply been suppressed. In this example the selected threshold value for the risk area is five inhabitants.

Table 1: Means and variances produced by different methods in the material (Markkula 2000)

	True	LRI-mean	LRI-permutation	Suppression
Mean	13.88	13.88	13.88	21.31
Variance	1759.63	1759.28	1759.63	2791.97
N	201	201	201	121

32. The results indicate that suppression distorts the results strongly, while the LRI retains the material fairly well as original. The means remain unchanged and even in random permutation the variance of the material is the same as before. Imputation by mean should decrease the variance somewhat, but in this case it stays almost as it was.

33. For further testing of the data protection methods and particularly for their implementation, the University of Jyväskylä built an SAS programme that runs the LRI interactively. The programme operates chiefly in three stages and the user can give it parameters. The method allows selection of different areas, both grid square and polygon-shaped, and of various risk limits and different imputation methods.

34. Statistics Finland has continued the testing of the LRI method. We intend to examine the behaviour of different sized and shaped statistical areas in the implementation of this method. The aim is also to prepare guidelines for the use of the method and make specified recommendations on the risk areas and area division. The method will be developed into a user-friendly product so that its implementation would be as easy as possible in Statistics Finland's units. For the time being, testing of the method and reporting on its suitability are still uncompleted.

IV. SUMMARY

35. The user environment and distribution channels of statistics have changed and are at the moment changing at high speed. New IT applications not only facilitate but also diversify the use of statistics. Finland's register-based statistical system, where statistical units can be located at the accuracy of one building, gives flexible means for producing small area statistics of various kinds. However, data protection may in the end be an obstacle to dissemination of data.

36. Finland's legislation regarding the processing of statistics was revised in the 1990s. However, various specifying directions relate mostly to statistics production based on conventional paper distribution. In order to renew dispersed and somewhat outdated guidelines, two workgroups were established by Statistics Finland during 2000: one for data protection of business statistics and the other for data protection of person-based materials. The purpose of both workgroups was to revise the guidelines concerning tabular, aggregated statistical material to correspond better to the needs rising from changing demand and distribution. The workgroups produced renewed guidelines for data protection during 2000. The implementation of these recommendations will start in the course of 2001.

37. Lack of suitable data protection methods has been partly the reason for that data protection measures at Statistics Finland have in practice often been made case specifically. This applies particularly to data protection relating to geographic information materials. Therefore, Statistics Finland developed jointly with the Department of Statistics of the University of Jyväskylä methods that would be suitable for data protection of geographic information, if necessary. The Local Restricted Imputation method (LRI) is a method that takes account of the special features of geographic information. By means of this method, the effects of data protection on statistical distributions would be as small as possible. Data protection is made locally so that the data will always be accurate in a hierarchically higher area level. This fairly promising method is at the moment being tested with Statistics Finland's total material. The method is probably best suited for research materials based on geographic information.

References

The Council of Europe's recommendation on personal data (1997). Recommendation No. R (97) 18 and Explanatory Memorandum Of The Committee Of Ministers To Member States Concerning The Protection Of Personal Data Collected And Processed For Statistical Purposes *(Adopted by the Committee of Ministers on 30 September 1997 at the 602nd meeting of the Ministers' Deputies).* www.coe.fr/dataprotection/rec/r(97)18e.htm.

The European Parliament and Council directive on the protection of individuals with regard to the processing of personal data 95/46/EC. "EU's data protection directive".

Personal Data Act (523/1999).

Henkilötietolain soveltamisohje (2000). Tilastokeskus TK-00-578-00. 15.5.2000. (Application directive on the Personal Data Act (2000).)

Henkilötietosuojatyöryhmä (2000). Henkilötietosuojatyöryhmän loppuraportti. 20.12.2000. Tilastokeskuksen Muistio. (Personal data protection workgroup (2000). Final report of the personal data protection workgroup.)

Hänninen, Minna (1997). Tilastolliset tietosuojamenetelmät ja niiden käyttö (Statistical data protection methods and their use). Tilastokeskuksen katsauksia 1997/3, Oy Edita Ab, ISBN 951-727-295-2.

ISI's professional ethical guide (1985). ISI - INTERNATIONAL STATISTICAL INSTITUTE . DECLARATION ON PROFESSIONAL ETHICS, Adopted: August 1985. http://www.cbs.nl/isi/ethics.htm.

Markkula, Jouni (2000). Disclosure Control Problem of Georeferenced Data. The Yearbook of the Finnish Statistical Society. 1999-2000. Helsinki 2000. ISSN 0355-5941.
Markkula, Jouni (1999). Statistical Disclosure Control of Small Area Statistics Using Local Restricted Imputation. ISI 99. 52nd Session. Bulletin of the International Statistical Institute. Contributed Papers. Book 2. pp. 267-268.

Ohjeet taulukkomuotoisen yritystiedon suojaamiselle (2000). Tilastokeskus 00-888-00, 21.6.2000. (Directions on protection of tabular business data (2000).)

Tilastokeskuksen ammattieettinen opas (1993). Toimi oikein tilastoalalla. Käsikirjoja 30. Tilastokeskus. Helsinki 1993. (Statistics Finland's guide on professional ethics (1993). Correct practice in the field of statistics. Handbooks 30.)

Statistics Act (62/1994) (734/1995 and 1030/1998).

Tilastolain (62/1994) soveltamisohje Tilastokeskuksessa (2000). TK-00-579-00. 15.5.2000. Tilastokeskus. (Application directive on the Statistics Act (62/1994) at Statistics Finland (2000).)

Tietojen suojaus Tilastokeskuksessa (1996). Tietosuojaohje. TK-00-1603-96. 17.9.1996. (Data protection at Statistics Finland (1996). Data protection guidelines.)

Fundamental principles of official statistics approved by the UN's Statistical Division (1994). E/CN.3./1994/18, E/1994/29.