Working Papers 3/2014

Balancing confidentiality and usability

Protecting sensitive data in the case of the inward Foreign AffiliaTes Statistics (FATS)

Katri Soininvaara, Teemu Oinonen, Annu Nissinen

Tilastokeskus
Statistikcentralen
Statistics Finland

Inquiries:

Teemu Oinonen
09 17 341

Principal editor: Faiz Alsuhail

Layout: Marita Potila

The publications in the Working Papers series do not represent Statistics Finland’s viewpoints.
The writers are responsible for the contents of the publications.

ISSN 2323–1998
= Working Papers
ISBN 978–952–244–489–9

Abstract

Statistical disclosure control is an issue for statistical institutions, which are responsible both for protecting confidential information collected from statistical units, and for disseminating information to the public. Adhering to the legislation on confidentiality, the statistical institutions face two contradictory problems: one concerning statistical disclosure control and protecting the identities of the statistical units, the other concerning the included level of detail and the usability of the disseminated information for the end users.

As an empirical example of balancing the two problems, the paper reports the results of an experiment conducted with case data on the inward Foreign AffiliaTes Statistics (FATS). The results are analysed to support the further decision making on protecting statistical publications against statistical disclosure, and to help to decide the level of publication detail for the inward FATS. The discussion includes an introduction to statistical disclosure control, the legal environment, and the main stakeholders of the inward FATS, a review of the publication choices found from the experiment, and recommendations based on the analysis of the results.

The current issue with the used statistical disclosure control methods for the inward FATS according to Eurostat is that on average only 10 % of the total cells are safe to publish in the case of the inward FATS table 1G. According to the analysis in this paper, in Finland only 26 % of the cells in the table 1G are active (i.e. there are firms in the cell) and only 34 % of the active cells are safe using the current specifications for the explanatory variables. Consequently, only 554 cells, or 9 % of the total 6 240 cells, are safe to publish in the table 1G.

The findings support Eurostat's claim that the detailed but largely suppressed tables disseminated currently are unlikely to be of much use. The most useful table found in this paper as an alternative configuration for the inward FATS 1G table has a total of 650 cells. Although only a tenth of the size of the original, 92 % of the total cells in the table are active, and 424 cells, or 65 % of the total 650 cells, are safe. As the size of the table gets smaller, the share of safe cells increases greatly. However, usability remains a concern, as the alternative table implies the use of area aggregates instead of country level aggregates. It can also be seen that the absolute amount of information computed as the count of safe cells reduces.

Keywords: Inward Foreign AffiliaTes Statistics, statistical disclosure control, table confidentiality, publication usability, experiment design

Table of contents

1 Introduction

The paper discusses the issue of producing confidential but useful tables in the case of the inward Foreign AffiliaTes Statistics (FATS). In order to ensure that statistical units cannot be identified from the published tables, tabular data need to be protected by statistical disclosure control methods, which prevent possible intruders from disclosing confidential information (Nissinen 2011). However, if protecting the data means that most of the cells in the table cannot be published, there is hardly any sense to publish such tables in the first place (Nissinen 2011). This is the problem of information loss due to the used protection method, currently prominent with the inward FATS.

The paper has three sections. This first section is an introduction, including the descriptions of the inward FATS, the problem, the legal environment, and the definitions of confidentiality and usability concerning the specific needs of the stakeholders for the inward FATS. The second section describes the experiment design and reports the results of the conducted 18–run experiment. The third section summarises the results and gives recommendations on how to balance the problems related to confidentiality and usability in the case of the inward FATS.

The inward FATS are collected from Norway and the European Union member countries to help to determine patterns of internationalisation, as well as to follow the consequences for expanding international business in the European Union. The relevant statistical institutions include Eurostat and the national statistical institutions (NSIs), such as Statistics Finland. The population for the inward FATS are those subsidiaries and branches in the compiling country that are controlled by a foreign entity. The collected data includes the residency of the ultimate controlling institutional unit (*uci*), industry classification (*nace*) and such characteristics as turnover, total purchases, and the number of persons employed. (Eurostat 2012.)

The inward FATS comprises two publications, IFATS Series 1 (1G) and IFATS Series 2 (1G2). Both tables 1G and 1G2 have the same two explanatory variables (uci and nace) but differ in the terms of included details. For Finland the table 1G includes the aggregated total firms in the compiling country (A1), aggregated compiling country enterprises (A2), aggregated foreign controlled enterprises in the compiling country (Z9), area aggregates V1 (EU–27, excluding Finland) and V2 (extra EU–27), C4 (offshore financial centres), and country level aggregates for the 27 EU countries (excluding Finland) and 14 extra–EU countries. The industry classification variable nace includes three industry levels and the sum of the total business economy (from B to N, excluding K and including S95). The table 1G2 includes all the other countries in the world in addition to the ones displayed in the table 1G but only the sum of the total business economy. More detailed descriptions of the tables are available from the FATS compilation manual. (Eurostat 2012.)

As described in the beginning, there are two contradictory issues related to protecting tables. One concerns the protection of the identities of the statistical units, while the other relates to maximizing usability. The problem for the inward FATS tables is that currently information needs to be suppressed in great amounts to prevent disclosure of sensitive information. Additionally, there may be too much detail included in the classification of the explanatory variables for the IFATS Series 1. This results in an ineffective 1G table, where according to Eurostat's analysis around 64 % of its 6 240 cells are zero values (i.e. the cell value is less than 0.5, or there are no firms, so the cell is empty), further 5 % are missing, 20 % need to be suppressed for confidentiality reasons, and thus only 10 % of the cells display safe non-zero values (Eurostat 2013). The aim of this paper is to find a better solution, if there is one.

Combining the issues discussed above, the research problem in this paper is to determine how to produce safe but informative inward FATS tables by using cell suppression for disclosure control, including the changes suggested by Eurostat for the classification of the explanatory variables (Eurostat 2013). The idea of the experiment is to analyse the effects of the different levels of classification for the explanatory variables, the use of different safety rule specifications, and the use of different secondary suppression algorithms [1]. While the goal is to produce recommendations on how to obtain as effective inward FATS publications as possible, the lessons learnt may as well be of use in planning statistical disclosure control for other statistical publications as well.

Table 1.1
Regulations concerning the confidentiality of enterprise statistics

Level	Regulation	Content	Relevant items
Finland	Statistics Act (280/2004)	No legal persons should be harmed with the collected information and the data should be protected at all stages of statistics production.	Section 10
		No statistical unit should be identified directly unless the information is public. Statistical authorities may grant access to confidential data from which statistical unit could be indirectly identified, provided that the data is used for scientific research or statistical surveys concerning social conditions.	Section 11 Section 12 Section 13 Section 18
European Union	Council Regulation (223/2009)	Statistical confidentiality is one of the governing principles for European statistics.	Article 2: e
		Statistical confidentiality concerns such data where single statistical units can be identified by a third party directly or indirectly.	Article 3: 7
		The statistical institutions can disseminate confidential information in such manner that statistical unit cannot be identified.	Article 3: 9 and 10 Article 19 Articles 20-26
European Union	European Statistics Code of Practice	Urges to keep the privacy of data providers.	Principle 5
		Advices to protect the confidentiality of the collected sensitive information.	

Source: The Statistics Act (280/2004), Council Regulation (223/2009), European Statistics Code of Practice (2011).

Disclosure control methods for tabular data can be divided into two main categories: perturbative and non-perturbative methods. Perturbative methods adjust the cell values in the table so that the original values cannot be estimated too accurately. Non-perturbative methods, also called restriction based methods, restrict the data available in the table either by adjusting the structure of the table or by suppressing the cell values. Since the non-perturbative methods are more transparent and

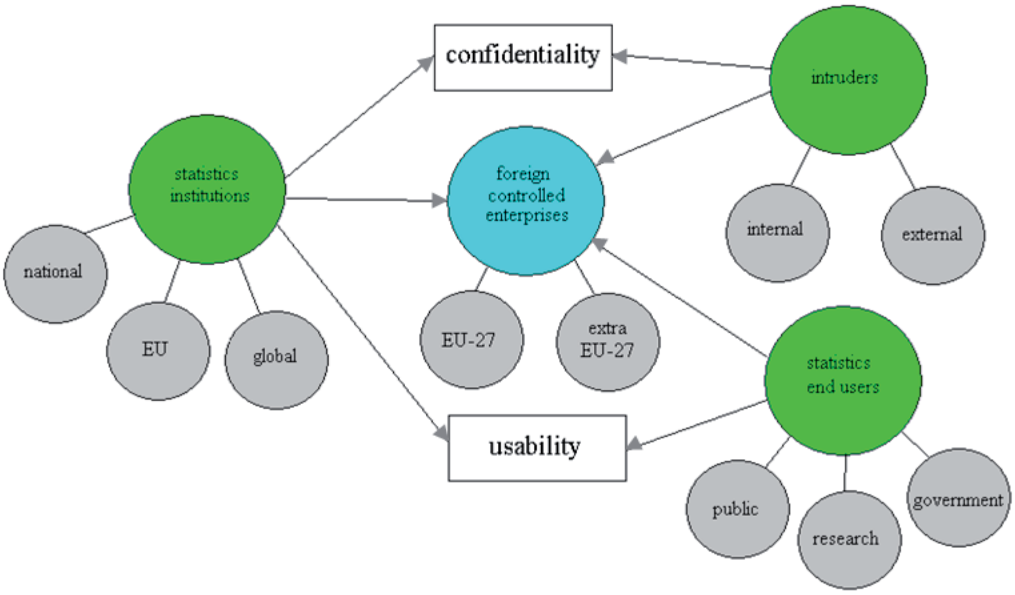
[1] The used parameters, safety rules, or secondary suppression algorithms are not revealed in this paper for publication security reasons. Interested parties with the required security clearance are advised to contact Teemu Oinonen or Annu Nissinen for further information.

easier to understand by the end users, the statistical institutions tend to prefer them in official statistics production. The inward FATS production is not an exception, and the compiling NSIs have settled on such traditional non-perturbative disclosure control methods as cell suppression and modifying the levels of classification for explanatory variables in the table. (Nissinen 2011.)

In Finland the data on industry classification, residency of the controlling institution, size category of turnover and total number of personnel for each enterprise are public information, which means that the cell frequencies in uci x nace tables and some of the response values used in the inward FATS tables are public (The Statistics Act, Section 18). However, most of the other response variables, for example the personnel costs of the enterprises, are classified as confidential information by the Finnish legislation (The Statistics Act, Section 12). The Table 1.1 summarises the key regulations concerning confidentiality of enterprise statistics in Finland.

The two contradictory problems in the statistics production arise from the inconsistent needs for different stakeholders. The key stakeholders are statistical units, which in the case of the inward FATS are foreign controlled enterprises. Confidential information needs to be protected on account of possible intruders, which can be either internal, such as a foreign controlled unit trying to achieve a competitive edge, or external, such as an outsider investor trying to make a profit. Other interested parties include the end users, which comprise the government, the research community, and the public. Finally, there are the statistics compiling and disseminating institutions: the NSIs such as Statistics Finland, the EU level bodies such as Eurostat, and global organisations such as OECD. The Figure 1.1 describes the stakeholders and their interests in the inward FATS statistics. (Eurostat 2012, Nissinen 2011.)

Figure 1.1
Different stakeholders for the inward FATS



Source: Based on Eurostat (2012) and Nissinen (2011).

The Figure 1.1 depicts the two main concerns in the statistics production related to statistical disclosure control: protection of the sensitive data from intruders, and usability of the published data for end users. While both the intruders and the end users are interested in any information they can get from the foreign controlled enterprises, the intruders try to disclose confidential information, whereas the end users are concerned about the usability of the published information. Unfortunately, confidentiality and usability are often contradictory features of statistics. If the level of confidentiality increases, usability decreases, and vice versa. Subsequently, finding a balance between confidentiality and usability is a dilemma, which needs to be solved before any tables containing sensitive data are published. (Eurostat 2012, Nissinen, 2011.)

The need for protection must be evaluated carefully. Assessing the risks for statistical disclosure is an important part of deciding how much protection is needed. For example, it should be taken into consideration that an intruder may have extra-table information, such as knowledge of its own market share in an oligopolistic industry. Depending on the risks, different statistical disclosure control methods are available. Sensitive cells can be protected to different effects by changing the parameters for the cell suppression method. Similarly, reducing the level of detail by aggregating classification levels helps to protect sensitive information by combining the excessively detailed cells together. If the risks have not been evaluated properly and the chosen protection method is not good enough, intruders can disclose confidential information despite the protection. (Nissinen 2011, 2013.)

While increasing the level of protection is easy, retaining usability is not. The cell suppression method results in information loss, as in order to ensure the protection of sensitive cells (i.e. primary suppression) also some non-sensitive information needs to be suppressed (i.e. secondary suppression). The amount of suppressed cells may become so great that the remaining information in the table is too scarce to be used for any purpose. In order to be able to use the tables correctly, the end users need to be informed on how the tables have been protected. However, the used parameter values should not be revealed. (Nissinen 2011, 2013.)

This section has introduced the two contradictory problems related to statistical disclosure control, which are protecting confidential data from intruders, and retaining usability of statistics for end users. As discussed above, the inward FATS include sensitive information, which results in the need to protect the tables build from the inward FATS data. Eurostat has noted that the current problem with the used cell suppression method and the level of detail for the inward FATS table 1G is that only 10 % of the cells are non-zero and can be published (Eurostat 2013). The next section reviews an experiment, which analyses the choices for protection methods available for the statistical institutions compiling inward FATS statistics.

2 Analysis

The idea of the experiment conducted in this paper is to list and compare the different possible inward FATS 1G tables using modified levels of classification, safety rules, and secondary suppression algorithms to protect confidential information. The software used to conduct the experiment is τ-Argus 3.5.0., build 26. The aim is to find an improvement to the current situation, where most of the cells are either zero values (i.e. the cell value is less than 0.5 or the cell is empty) or need to be suppressed for protection.

The experiment is carried out for one of the response variables according to the specifications of the inward FATS table 1G described in the previous section. The inward FATS includes two series (1G and 1G2), so the protection is completed by linking the two tables. Linking the tables ensures that none of the information published in the other table can be used in disclosing confidential information from the other. In order to keep the experiment as simple as possible, only the table 1G is considered, so further discussion on the use of linked tables is outside the scope of this study. For additional review on the use of linked tables, see Nissinen (2011).

Table 2.1
Experiment design

Factors	–	+
A: nace	Std	Mod1
B: nace	Std	Mod2
C: uci	Std	Mod1
D: uci	Std	Mod2
E: safety rule	Std	Mod

The Table 2.1 describes the experiment design, which is based on Box (2006). The standard combination (-) shown describes the current situation, whereas the modifications (+) are suggestions for improvements. The different combinations of the factors are run through two different secondary suppression algorithms. Although implicating a 2^5 or a 32-run experiment, the experiment here includes only 18 runs for each algorithm. This is because the factors A and B refer to the same standard specification of the explanatory variable nace, making the standard combinations of the factors A and B redundant. The factors differ in their modifications, which are stated as *Mod1* and *Mod2*. Similarly, the factors C and D refer to the same standard specification for the explanatory variable uci but have different modifications. This section presents the results for the first algorithm, while the corresponding tables and figures for the second algorithm can be found from the appendices for comparison.

The factors from A to D represent changes in the levels of classification for the explanatory variables. Changing these factors affects the level of detail included in the nace x uci table, and consequently adjusts the size of the table. The standard combination refers to the current classifications, which includes three hierarchical industry levels for *nace* and country level detail for uci. As discussed in the previous section, Eurostat has suggested that the current classifications may be too detailed, so the function of the experiment factors from A to D is to see what happens to active and confidential cell shares when the levels of classification are modified.

In the first modification to nace (factor A), the third level is dropped from examination. The second modification to nace (factor B) drops both the second and the third levels, which corresponds to the current Eurostat proposal (Eurostat 2013). In the first modification to uci (factor C) the country level is dropped, so that only the large area aggregates V1 (EU-27, excluding Finland) and V2 (extra EU-27) and their total aggregate Z9 (all foreign controlled enterprises in Finland), and the aggregate A2 (all Finnish controlled enterprises in Finland) are retained. In the second modification to uci (factor D), continental aggregates are introduced. Additionally, the second uci modification includes the aggregates A2 and Z9.

The factor E refers to the used safety rule. Changing the safety rule affects the amount of sensitive cells, and thus changes the level of confidentiality. The reason for modification to the safety rule in the experiment is to test if it is possible to reduce the share of confidential cells by slacking the used safety rule. A similar reasoning lays behind the use of two different secondary suppression algorithms. The first algorithm is the one used currently to protect the inward FATS in Finland. The second is an alternative algorithm offered by the used software.

The experiment results in four types of cells: active, safe, confidential, and empty. The total cells in the table consist of active and empty cells. Active cells or non-empty cells have firm activity between the particular nace x uci combination. Active cells are either safe or confidential. Safe refers to the cells that are published. Confidential cells consist of primary confidential cells plus the secondary suppressions determined by the used safety rule and the secondary suppression algorithm.

Empty cells refer to such cells, where there is no firm activity between the particular nace x uci combination. It should be noted that Eurostat refers to these as zero cells, which include both empty cells (marked as zeros in the inward FATS tables) and additionally such cells, where the cell value is less than 0.5. Thus in the experiment there are a few cells, which are counted as active, where values equal zero (i.e. there is firm activity in the cell but the value of the activity is zero), whereas in the FATS manual these are calculated as zero cells (i.e. there might or might not be any activity in the cell). The difference stems from the used software, which produces empty cells rather than zeros, and the Eurostat practise, described in more detail in the FATS compilation manual (Eurostat 2012, p. 89). As the count of cells valued zero is negligible in the experiment, the difference does not affect the conclusions.

The Table 2.2 summarises the results of the experiment for the first algorithm. The results for the second algorithm can be found from the Appendix I. It can be seen that the first run, where no modifications to the current situation are made, shows that 66 % of the active cells are suppressed. While this is the worst case in terms of confidential cell share, the unmodified table also includes the most detailed information, its active cell count totalling to 1 636, which includes 554 safe cells. The safe cell count is maximized at 597 with the run number two, where the only factor changed to the current situation is the safety rule factor E. In the modified specification, the safety rule has been slacked, and consequently the confidential cell share reduces to 64 %. From the Appendix I it can be seen that the second algorithm is more efficient, suppressing only 54 % and 53 % of the active cells in the corresponding cases.

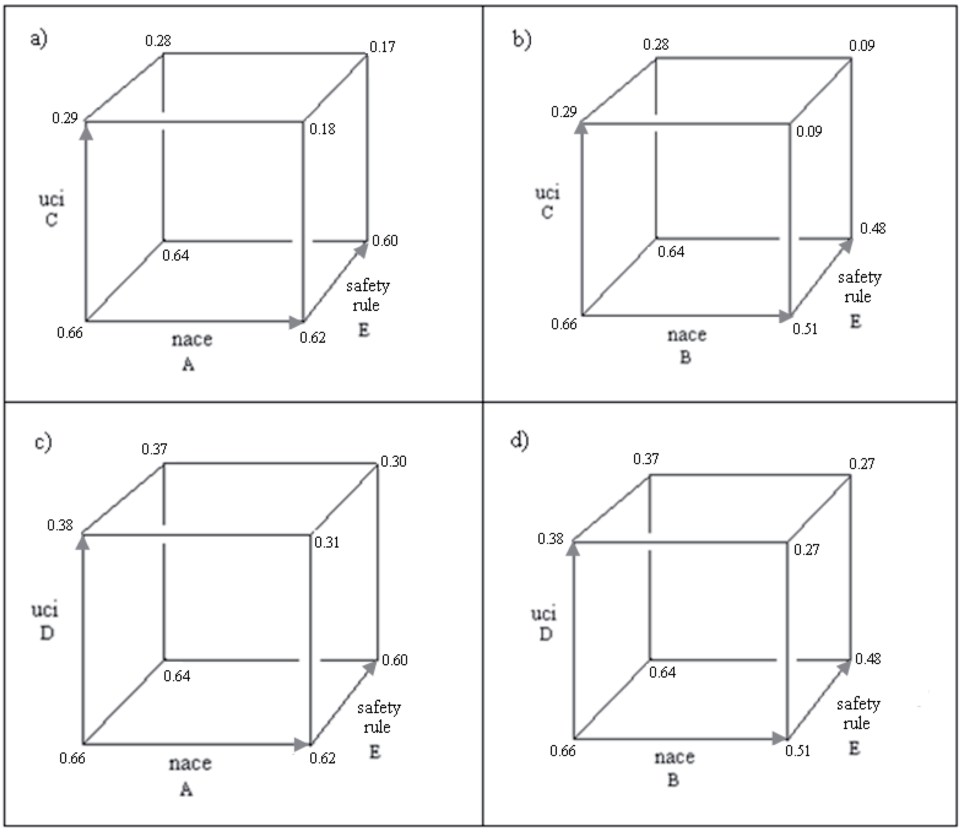
Table 2.2
The 18 runs and the results of the table redesign experiment

Run number	A nace	B nace	C uci	D uci	E safety rule	Total	Total active	Total safe	Safe/ Active	Confiden-tial/Active
1	–	–	–	–	–	6,240	1,636	554	0.34	0.66
2	–	–	–	–	+	6,240	1,636	597	0.36	0.64
3	+	–	–	–	–	3,840	1,148	438	0.38	0.62
4	+	–	–	–	+	3,840	1,148	460	0.40	0.60
5	–	+	–	–	–	864	370	183	0.49	0.51
6	–	+	–	–	+	864	370	191	0.52	0.48
7	–	–	+	–	–	650	599	424	0.71	0.29
8	–	–	+	–	+	650	599	430	0.72	0.28
9	–	–	–	+	–	1,040	656	406	0.62	0.38
10	–	–	–	+	+	1,040	656	414	0.63	0.37
11	+	–	+	–	–	400	381	313	0.82	0.18
12	+	–	+	–	+	400	381	315	0.83	0.17
13	+	–	–	+	–	640	432	298	0.69	0.31
14	+	–	–	+	+	640	432	301	0.70	0.30
15	–	+	+	–	–	90	89	81	0.91	0.09
16	–	+	+	–	+	90	89	81	0.91	0.09
17	–	+	–	+	–	144	112	82	0.73	0.27
18	–	+	–	+	+	144	112	82	0.73	0.27

The best case in terms of confidential cell share is the table resulting from the runs numbered 15 and 16, which are otherwise the same but have different modifications to the safety rule factor E. They result in identical tables, indicating that changing the chosen safety rule does not affect the results in this case. In these runs, both nace and uci are modified. Following the Eurostat suggestion, only the first level of *nace* is retained. Additionally, only the large area aggregates are retained from the variable *uci*. The result is a table with only 89 active cells, which is the smallest amount of safe cells from the alternatives presented here. However, the share of the confidential cells is minimized, as only 9 % of the cells need to be suppressed. Comparing with the Appendix I, it can be seen that in this case the second algorithm results in the same table.

The Figure 2.1, based on Box (2006), is a graphical representation of the results for confidential cell share shown in the Table 2.2. The Appendix II presents a similar figure for the second algorithm. Each cube represents interactions between different combinations of the classification factors from A to D, and the safety rule factor E. Panel a) is a display of the factors A (*nace* with two levels), C (uci with large aggregates) and E (safety rule). Panel b) is a similar display for the factors B (nace with one level), C and E, panel c) is a display for the factors A, D (uci with continental aggregates) and E, and panel d) displays the factors B, D and E. The current situation, i.e. the run number one in the Table 2.2, is located in the left hand corner of each cube in the Figure 2.2. Changing nace means moving to the right and modifying uci means moving up. As a result, the upper right hand corners represent situations, where both variables are modified. Additionally, changing the safety rule means moving to the back panels of the cubes.

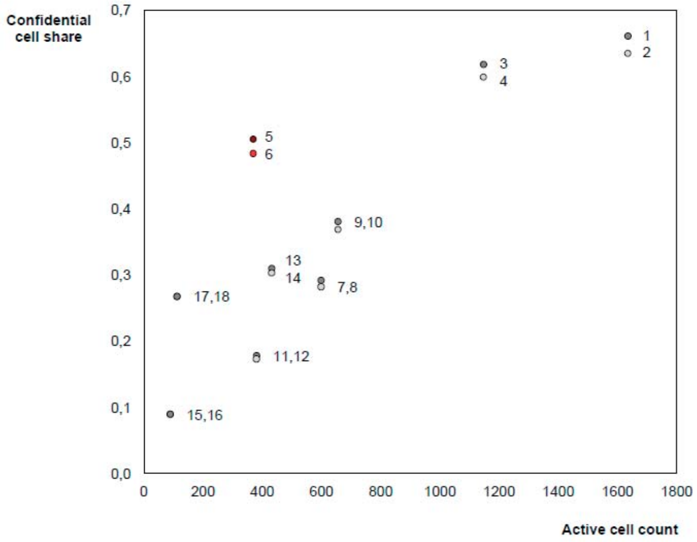
Figure 2.1
Cube display of the experiment results



It can be seen from the Figure 2.1 that changing the safety rule affects the share of the suppressed cells in a minimal way or not at all. The most dramatic reductions in the confidential cell share come from modifying both nace and uci at the same time. While changing nace from three levels to two does not affect the confidential cell share in a significant manner, as seen from the panels a) and c), modifying nace further to contain only one level reduces the confidential cell share by 15 percentage points, as shown by the panels b) and d). Modifying uci changes the confidential cell share even more drastically, and using any combination of the two explanatory variable modifications results in both smaller tables and lower confidential cell shares as seen by looking at the upper right hand corners of each cube. Comparing with the Appendix II, it can be seen that the most results are five to ten percentage points lower with the second algorithm, meaning that the second algorithm is more efficient and results overall in fewer suppressions.

An upward sloping relationship between the number of active cells and the share of confidential cells can be gathered from the Figure 2.2. The total number of active cells in the x-axis describes the size of the table, which also reflects the included level of detail, implying that a larger amount of active cells means a more detailed table. The y-axis describes the share of the cells needed to be suppressed as primary or secondary suppressions. The Eurostat suggestion (retaining the first nace level only, standard uci) has been marked in red. The standard use for the safety rule is marked in darker gray and its modification in lighter gray. A similar display for the second algorithm can be seen in the Appendix III.

Figure 2.2
The table redesign affects the share of confidential cells



	nace	uci	safety rule
1	Std	Std	Std
2	Std	Std	Mod
3	Mod1	Std	Std
4	Mod1	Std	Mod
5	Mod2	Std	Std
6	Mod2	Std	Mod
7	Std	Mod1	Std
8	Std	Mod1	Mod
9	Std	Mod2	Std
10	Std	Mod2	Mod
11	Mod1	Mod1	Std
12	Mod1	Mod1	Mod
13	Mod1	Mod2	Std
14	Mod1	Mod2	Mod
15	Mod2	Mod1	Std
16	Mod2	Mod1	Mod
17	Mod2	Mod2	Std
18	Mod2	Mod2	Mod

While the relationship itself is obvious, as there must be more all types of cells when the size of the table increases, the purpose of the Figure 2.2 is to show the relationships between the alternatives presented in the experiment. Specifically, it should be noted that the count of the active cells reduces rapidly with most of the modifications, while the confidential cell share decreases at a more moderate rate. This indicates that while changing the categories helps to lower the confidential cell share, it is not as straightforward to deduct that reducing the size of the table increases usability.

The Table 2.3 displays nine table alternatives derived from the experiment results. As the chosen safety rule modification did not affect the results in a significant manner, the tables are shown only for the standard safety rule. It can be seen that the original situation results in a large 130 x 48 table, which only has 26 % of its cells filled with active cells. Only 9 % of the total cells are safe. The most drastic modification to the classifications leads to a table with dimensions 18 x 5, which only has 90 cells in total. However, a total of 99 % of these cells are active, in fact, only one of the cells is left empty. This highly aggregated table means that almost all, 90 % of the total cells in this table can be published. The Appendix IV displays a similar table for the second algorithm.

Table 2.3
Summary of the experiment with the standard safety rule

Table number	Run[2]	nace levels	uci	Total	Active	Safe	Active/ Total	Safe/ Total	Table size	Size compared to original
1	5	1	countries	864	370	183	0.43	0.21	18 x 48	0.14
2	3	2	countries	3,840	1,148	438	0.30	0.11	80 x 48	0.62
3	1	3	countries	6,240	1,636	554	0.26	0.09	130 x 48	1.00
4	15	1	large areas	90	89	81	0.99	0.90	18 x 5	0.01
5	11	2	large areas	400	381	313	0.95	0.78	80 x 5	0.06
6	7	3	large areas	650	599	424	0.92	0.65	130 x 5	0.10
7	17	1	continents	144	112	82	0.78	0.57	18 x 8	0.02
8	13	2	continents	640	432	298	0.68	0.47	80 x 8	0.10
9	9	3	continents	1,040	656	406	0.63	0.39	130 x 8	0.17

[2] Cf. Table 2.2.

Removing the third nace level results in the table two, which has 80 x 48 cells. It is sized 62 % of the original table. However, only a slightly larger percentage for safe cells out of the total cells is gained, as it increases to 11 % from the original 9 %. Keeping only the first nace level, as suggested by Eurostat, the result is the table one, which has 18 x 48 cells. While the size of the table compared to the original shrinks to just 14 % of the size of the original table, the total safe cell share increases only to 21 % from the original 9 %. Additionally, keeping an eye on the absolute counts of the safe cells, it can be seen that the number of safe cells reduces to about a third from the original, from 554 to 183. Compared to the table five, where two of the nace levels are retained and uci is modified so that only the large aggregates remain, 313 safe cells are available, and the total safe cell share rises to 78 %. However, the size of the table compared to the original shrinks even further, to 6 %. The Appendix IV shows the summary of the results for the second suppression algorithm.

The Table 2.2 showed that large size is associated with a relatively large absolute number of active cells. However, large size also means that there is more need for protection, resulting in higher confidential cell shares for the largest tables. The Figure 2.2 depicted the increasing relationship between the total active cells and the confidential cell share. However, it was also noted that the rate in which the confidential cell share reduces is lower than the rate in which the size of the table shrinks. This observation means that increasing the usability is not as simple as reducing the size of the table by reclassification of the explanatory variables. This was also seen in the manner in which more safe cells were available to certain combinations of the classification modifications than others in the summary Table 2.3. The Table 2.4 describes the results of an analysis, which aims to take these factors into consideration in choosing the most useful alternative from the choices considered in the experiment.

Each table has been given weights based on three criteria. The first criterion states the share of confidential cells. It reflects the need to minimize the information loss due to protection, which implies that the smaller the confidential cell share, the better the usability. The lowest share gains 30 points and the highest three are given zero points. The second criterion is the safe cell count. It stands for the need to maximize usability, and reflects the absolute amount of information gained from the table. The largest number gets 30 points and the smallest three are given zero points. The third criterion is the size of the table measured by the total cell count. It reflects the need for included detail, implying that more detail is preferred to

Table 2.4
Analysis of the alternative tables

Table number	Confidential cell share	Safe cell count	Size	Weight 1	Weight 2	Weight 3	Total	Rank
1	0.51	183	864	0	0	15	15	x
2	0.62	438	3,840	0	25	25	50	x
3	0.66	554	6,240	0	30	30	60	x
4	0.09	81	90	30	0	0	30	x
5	0.18	313	400	25	10	0	35	x
6	0.29	424	650	15	20	10	45	1
7	0.27	82	144	20	0	0	20	x
8	0.31	298	640	10	5	5	20	3
9	0.38	406	1040	5	15	20	40	2

less detail. The largest table gets 30 points and the three smallest tables are given zero points. The total points are calculated as the sum of the three weight columns. Finally, if a table got zero points with any criteria, it is given a rank x, otherwise it is ranked based on the weighted points.

According to the analysis, the usability is maximised with the tables six, eight and nine. Tables six and nine retain the original three nace levels, whereas the table eight drops the third *nace* level. Table six includes aggregating the uci so that only large area aggregates remain. Tables eight and nine aggregate uci so that the countries are grouped under continental aggregates. The Eurostat (2013) suggestion, table one, gets an x-rank, as it is given zero points in the first and second criteria (confidential cell share and safe cell count). Moreover, it does not rank highly in the final criterion either, and results in the lowest total point score in the analysis.

The Table 2.4 shows that the table six is ranked the highest given the three criteria after the x-ranked alternatives are dropped from the analysis. Thus the most useful table found in this paper as an alternative format for the inward FATS 1G table has a total of 650 cells. 92 % of the total cells are active, and 424 of the cells are safe. Although the size is just 10 % of the original and the reduction in size means also a reduction in the absolute count of safe cells, this is a great improvement from the current situation looking at the confidential cell share, which increases from 9 % to 65 %. However, usability remains a concern, as the alternative table implies large area aggregates, which replace the more detailed country level information. Nonetheless, as long as the need for the level of protection stays the same, this is the most usable version of the 1G table alternatives presented in this experiment given the three utilised criteria.

3 Conclusions

The objective of statistical institutions is both to protect the confidentiality of statistical units and to disseminate data for public use. These contradictory goals manifest as problems with balancing confidentiality and usability. The research problem in this paper set in the first section has been to determine how to produce safe but informative inward FATS tables by using cell suppression for disclosure control, including the changes suggested by Eurostat for the reclassification of the explanatory variables (Eurostat 2013). The second section presented an experiment, where the effects of the different levels of classification for the explanatory variables, the use of different safety rule specifications, and the use of different suppression algorithms were analysed in depth.

In the first section of the paper, the legislation on statistical confidentiality was reviewed and the demands for confidentiality and usability were analysed through a stakeholder analysis. For the inward FATS in Finland, the current legislation means that the tables need to be protected in such manner that no statistical unit or their confidential attributes can be disclosed from the published statistics. The main stakeholders in the case of the inward FATS are the enterprise informants, the possible internal or external intruders, the end users, and the compiling statistical institutions. The main issue for the inward FATS publications is that due to the need to protect the identities of the statistical units, a large part of the tables needs to be suppressed. Additionally, Eurostat has suggested that the currently used classifications may be too detailed for the needs of the end users (Eurostat 2013).

The experiment on the case data was reviewed in the second section. It was shown that the current 1G table is not very useful to the end users as only 9 % of the total cells are active and safe (cf. table three in the Tables 2.3 and 2.4). However, aggregating the classifications too much would result in highly aggregated table, of which most could be published, but none would likely describe any useful information (cf. table four in the Tables 2.3 and 2.4). The analysis shown in the Table 2.4 illustrated that the most usable alternative ranks the highest with three criteria: confidential cell share, the absolute amount of safe cells, and the size of the table.

Eurostat has suggested aggregating the activity breakdown variable nace to contain its first level only (cf. table one in the Tables 2.3 and 2.4, Eurostat 2013). The findings from this paper indicate that the size of such table reduces to 14 % of the original, and the absolute amount of safe cells reduces drastically, while the confidential cell share remains relatively high at 51 %. Thus, the findings of this experiment do not support the Eurostat (2013) proposal for the table 1G. According to the analysis in this paper, the most usable table has a total of 650 cells (cf. table six in the Tables 2.3 and 2.4). With this table, there are three levels for nace and large area aggregates for uci. 92 % of the total cells are active, and 65 % are safe. While the results were generally similar for the other suppression algorithm, it was noted that especially for the large tables the second algorithm was more efficient, resulting in five to ten percentage points lower secondary suppression rates. For the most usable table, the total safe cell share increases to 69 %.

The size of the most usable table suggested in this paper is only 10 % of the original 1G table, which indicates that a lot of detail is lost in aggregating uci from the country level to the area aggregates. Thus, a concern remains that such aggregation might not be useful to those end users, who have been using the currently available detailed nace information on the country level on the EU member countries and the fourteen most important partner countries. On the other hand, it should be noted that the researchers would still have access to the 1G2 table, where all of the world's countries are presented with the sum of the total business economy. As the alternative 1G table presents the aggregated geographical totals (EU and the rest of the world) for the detailed business activity breakdown, the two inward FATS tables would intuitively complement each other to as much detail as possible given the current demand for protection.

Although this paper has shown that using aggregation of the explanatory variables can result in great reductions in the confidential cell share in the case of the inward FATS, there remains other concerns related to the usability of the tables. Alternative ways to publish the inward FATS data could be considered in order to be able to publish more detailed information. One option could be to use the perturbative statistical disclosure control methods (cf. Nissinen 2011). However, as perturbative methods are not currently used in Statistics Finland or in the inward FATS production, more research should be conducted to examine the suitability of such methods in the case of the inward FATS and similar statistics.

This paper has discussed the task of balancing confidentiality and usability in the case of the inward FATS data. If the required level of confidentiality is given by the current legislation, there is not much room to manoeuvre the level of usability either. By reducing the size of the table and the amount of detail, fewer sensitive cells remain in the table. However, the demand for detail depends on the uses of the statistics. There may also exist conflicting needs, where different types of end users need different types of information. Therefore the appropriate level of detail is difficult to adjust from the point of view of the NSIs, as acknowledging the needs for specific users is most of the time outside the scope of their capabilities.

Choosing the most efficient variable uci for aggregation, the current situation of the inward FATS series 1G can be improved at least to some extent. While it is definitely more useful to have non-suppressed cells rather than suppressed ones, aggregation in itself is a method of reducing information. Consequently, the question remains whether the amount of information increases to any usable degree, even if the confidential cell share reduces. The paper shows that the task of balancing confidentiality and usability by aggregation and secondary suppression algorithms is nearly impossible.

References

Box, G.E.P. (2006). Improving almost anything: Ideas and essays, John Wiley & Sons, Inc.: Hoboken, NJ.

Council Regulation (EC) No. 223/2009 of 11 March 2009 on European statistics and repealing regulation, Official Journal of the European Union, 31.3.2009, L87, pp. 164-173. Available from: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:0173:en:PDF. Accessed on 1.8.2013.

Eurostat (2012). Foreign AffiliaTes Statistics (FATS) recommendations manual, 3rd ed., Eurostat Methodologies and Working papers. Available from http://epp.eurostat.ec.europa.eu/portal/page/portal/product_details/publication?p_product_code=KS-RA-12-016. Accessed on 1.8.2013.

Eurostat (2013). FATS in FRIBS, draft document for JOINT FATS WG meeting. 10th June 2013, Luxembourg. Internal documentation.

European Statistics Code of Practice for the National and Community Statistical Authorities (2011). Adopted by the European Statistical System Committee on 28th September 2011. Available from: http://epp.eurostat.ec.europa.eu/portal/page/portal/product_details/publication?p_product_code=KS-32-11-955. Accessed on 1.8.2013.

Nissinen, A. (2011). Taulukkoaineistojen tilastolliset tietosuojamenetelmät (trans. Statistical disclosure control methods for tabular data), Master's thesis, University of Helsinki: Helsinki, Finland.

Nissinen, A. (2013). Ohjeet taulukkomuotoisten yritystietojen suojaamiseksi (trans. Instructions to protect tabular data including enterprise information), Statistics Finland. Presentation on 11th April 2013. Internal documentation.

The Statistics Act (280/2004). Available in English through: http://tilastokeskus.fi/meta/lait/2013-09-02_tilastolaki_en.pdf. Accessed on 19.9.2013.

Appendices

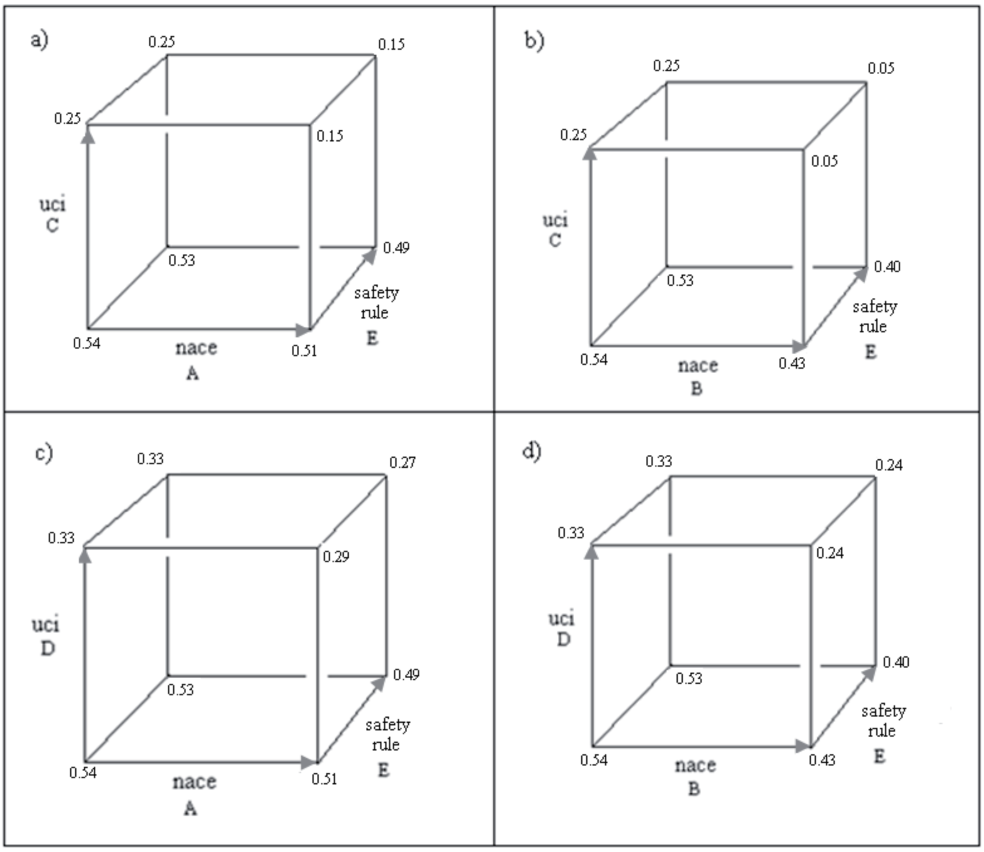
Appendix I.

Experiment with the second algorithm

Run number	A nace	B nace	C uci	D uci	E safety rule	Total	Total active	Total safe	Safe/ Active	Confidential/ Active
1	−	−	−	−	−	6,240	1,636	749	0.46	0.54
2	−	−	−	−	+	6,240	1,636	775	0.47	0.53
3	+	−	−	−	−	3,840	1,148	560	0.49	0.51
4	+	−	−	−	+	3,840	1,148	589	0.51	0.49
5	−	+	−	−	−	864	370	212	0.57	0.43
6	−	+	−	−	+	864	370	222	0.60	0.40
7	−	−	+	−	−	650	599	448	0.75	0.25
8	−	−	+	−	+	650	599	451	0.75	0.25
9	−	−	−	+	−	1,040	656	438	0.67	0.33
10	−	−	−	+	+	1,040	656	442	0.67	0.33
11	+	−	+	−	−	400	381	321	0.84	0.16
12	+	−	+	−	+	400	381	322	0.85	0.15
13	+	−	−	+	−	640	432	308	0.71	0.29
14	+	−	−	+	+	640	432	314	0.73	0.27
15	−	+	+	−	−	90	89	81	0.91	0.09
16	−	+	+	−	+	90	89	81	0.91	0.09
17	−	+	−	+	−	144	112	85	0.76	0.24
18	−	+	−	+	+	144	112	85	0.76	0.24

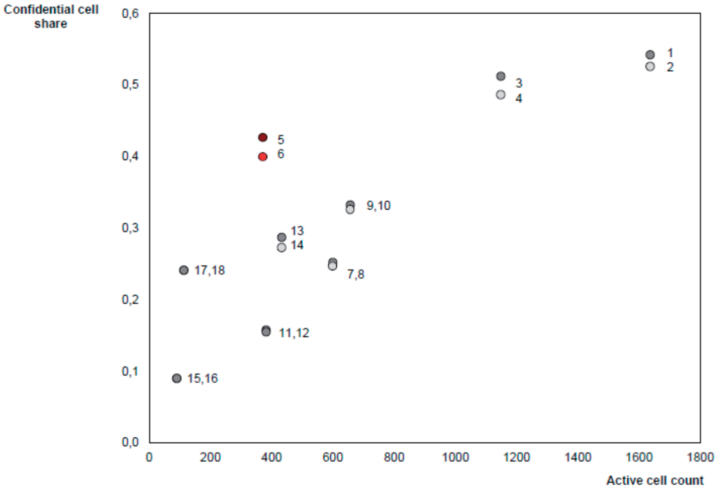
Appendix II.

Cube display for the second algorithm



Appendix III.

Scatter plot for the second algorithm



	nace	uci	safety rule
1	Std	Std	Std
2	Std	Std	Mod
3	Mod1	Std	Std
4	Mod1	Std	Mod
5	Mod2	Std	Std
6	Mod2	Std	Mod
7	Std	Mod1	Std
8	Std	Mod1	Mod
9	Std	Mod2	Std
10	Std	Mod2	Mod
11	Mod1	Mod1	Std
12	Mod1	Mod1	Mod
13	Mod1	Mod2	Std
14	Mod1	Mod2	Mod
15	Mod2	Mod1	Std
16	Mod2	Mod1	Mod
17	Mod2	Mod2	Std
18	Mod2	Mod2	Mod

Appendix IV.

The table choices with the second algorithm

Table number	Run[3]	nace level	uci	Total	Active	Safe	Active/ Total	Safe/ Total	Table size	Size compared to original
1	5	1	countries	864	370	212	0.43	0.25	18 x 48	0.14
2	3	2	countries	3,840	1,148	560	0.30	0.15	80 x 48	0.62
3	1	3	countries	6,240	1,636	749	0.26	0.12	130 x 48	1.00
4	15	1	large areas	90	89	81	0.99	0.90	18 x 5	0.01
5	11	2	large areas	400	381	321	0.95	0.80	80 x 5	0.06
6	7	3	large areas	650	599	448	0.92	0.69	130 x 5	0.10
7	17	1	continents	144	112	85	0.78	0.59	18 x 8	0.02
8	13	2	continents	640	432	308	0.68	0.48	80 x 8	0.10
9	9	3	continents	1,040	656	438	0.63	0.42	130 x 8	0.17

[3] Cf. Appendix 1.

Tilastokeskus
Statistikcentralen
Statistics Finland