It's confidential

How to Avoid the Brickwall of Confidentiality when Linking Micro Data

European Conference on Quality in Official Statistics (Q2014)

Jon Mortensen (jmo@dst.dk)
Head of section, External Economy, Statistics Denmark[1]

Søren Burman (sbu@dst.dk)
Head of section, External Economy, Statistics Denmark[1]

Abstract

Confidentiality is essential to quality in official statistics. By ensuring that published statistics do not disclose any individual or enterprise, confidentiality spurs trust, which is crucial in assuring delivery of reliable information from respondents. Confidentiality is thus fundamental to all statistical agencies. It cannot be compromised. However, confidentiality also creates tensions between an enthusiastic and growing user demand for micro data and producers of statistics concerned with the protection of identifiable information.

One way to meet both user demands for micro data and confidentiality concerns is to produce standardized statistics based on linking of micro data. Eurostat's Trade by Enterprise Characteristics (TEC) statistics does precisely that and is generally considered to be an important tool for a better understanding of the global economy. This tension was evident in two recent Eurostat pilot studies, which aimed at testing new datasets for TEC and explore the potential for including also International Trade in Services in the framework. The article discusses the studies' results focusing on confidentiality issues and provides guidelines for how to tabulate data based on linked micro data without hitting the brickwall of confidentiality.

[1] Any views or opinions presented in this paper are solely those of the authors and do not necessarily represent those of Statistics Denmark.

1. Introduction

Written by two producers of statistics this paper is "a voice from the floor" in respond to the growing user demand for linking of micro data and the tension this creates in regard to confidentiality. Confidentiality is essential to quality in official statistics. By ensuring that published statistics do not reveal information on any individual or enterprise, confidentiality spurs trust, which is crucial in assuring delivery of reliable information from respondents. Confidentiality is thus fundamental to all statistical agencies. It cannot be compromised. Conversely, confidentiality is often experienced by users of statistics as a brickwall separating them from essential information. One way to meet both user demands for more detailed data and confidentiality concerns is to produce standardized statistics based on linking of micro data. Eurostat's Trade by Enterprise Characteristics (TEC) statistics does precisely that and is generally considered an important tool for better understanding of the global economy. But as was evident in two recent Eurostat pilot studies, which aimed at testing new datasets for TEC and explore the potential for including also International Trade in Services in the framework (called S-TEC) in addition to trade in goods, the tension between what users want and what producers can provide remains.

The paper discusses the studies' results with a focus on confidentiality issues. We will show that an ambition to tabulate disaggregated, micro-linked data along two or more dimensions simultaneously necessitates a high share of confidential cells. We argue that this significantly reduces the relevance of the datasets for the users, while the compilation process becomes disproportionately taxing for the producer. The level of granularity simply becomes too high. Instead we suggest decreasing the level of granularity by only disaggregating tables along one dimension at the time. This would reduce the proportion of confidential cells and ease the compilation process. Of course a high level of granularity was one of the ambitions of the pilot studies in order to provide policy makers and analysts at international organizations (such as the EU-commission and the OECD) with more detailed micro-linked data. We argue, however, that a more appropriate avenue for pursuing this ambition is to give trusted international organizations access to the micro data itself in anonymized form. A service which domestically already is the norm in many countries, but stops at the national border.

2. The Brickwall of Confidentiality

National statistical institutions (NSIs) collect information about individual persons and enterprises. This information is confidential. NSIs will therefore typically suppress the value of cells with confidentiality problems in published tables (or de-identification of accessible micro data, see below),

thus sometimes erecting a brickwall between users and the information they seek. This creates tension between the two basic responsibilities all NSIs have in common, to provide access to high-quality statistics and at the same time maintain confidentiality. These two responsibilities are not immediately compatible. Confidentiality is a responsibility NSIs hold to the providers of the data. NSIs must assure data providers (individuals, households, organizations and enterprises) that participation is not harmful. High-quality data is a responsibility to the public. NSIs must produce useful statistics for a wide array of data users – policy makers, researchers, analysts etc. In this way, NSIs serve two masters – each with conflicting interests and concerns[2].

The challenge for the NSI is to resolve this inherent tension. One way to do this would be to prioritize one master above the other. Essentially, NSIs are funded to produce statistics, and the level of funding is determined by policy makers, who are also major users of statistics. So why don't NSIs just make users happy and produce statistics with little regard to confidentiality? We see three fundamental motivations for NSIs to priority confidentiality:

1. Legal constraints
2. Ethical concerns
3. Quality considerations

Confidentiality is often required by laws and regulations, but NSIs also impose confidentiality for other than legal reasons. Ethical concerns are central in many NSIs approach to confidentiality. Also, confidentiality spurs trust, which is crucial in assuring delivery of reliable information from respondents. In this way confidentiality is essential to producing high-quality statistics. So when users of statistics hit the brickwall of confidentiality they should keep in mind that the wall is there to ensure that the data they *can* access is of high-quality.

3. Implementing confidentiality

Erecting a brickwall of confidentiality is a time-consuming, cumbersome process often with a high degree of manual work, which few statistical compilers enjoy. It is thus not something NSIs do unless necessary. Probably the most popular method for implementing confidentiality in tables is to suppress (hide) risky cells by replacing the value with a predefined symbol. Finding the risky cells is usually quite straight forward, using predefined rules such as the dominance criteria, where a cell is considered risky if two enterprises in the cell have equal to or more than 85 pct. of the trade in the

[2] [1] Duncan, George T., Elliot, Mark, Salazar-González, Juan-José (2011), "Statistical Confidentiality – Principles and practive", Springer

given cell. However, in order to secure that a risky cell is confidential, one must not be able to deduct its value as a residual, by subtracting the non-risky cells from the total. Thus, a *secondary suppression* of cells is necessary to ensure sufficient confidentiality for the risky cells. Secondary confidentiality procedures are not straight forward, and the secondary suppression is the real burden of confidentiality from a compilers point of view, since there is no standardized method for doing this. This is best shown with a simple example. Imagine a secondary confidentiality procedure that is automatized to choose the smallest value to suppress, and will check horizontal lines in numerical order (line 1 before line 2, etc.) before vertical columns (column A before column B, etc.). In the case below, the cell C4 is found to be risky (marked with gray), and therefore the procedure chooses to suppress cell A3 (marked with red), since this is the lowest value (B3 and D3 both have the value of 4).

	A	B	C	D	Total
1	2 (S)	9	8 (S)	10	29
2	7 (S)	2	1 (S)	8	18
3	3 (S)	4	4 (C)	4	15
Total	12	15	13	22	62

(C) is primary confidentiality and (S) is secondary confidentiality

After ensuring that all cells in the horizontal lines are properly suppressed, the procedure continues to check the columns starting from column A, and therefore chooses to suppress cell A1 and C1 (marked with green). The secondary confidentiality procedure suppresses a combined value of 13 in order to suppress cell C4. If the procedure had started with column C instead of column A, then the suppressed cells would have been the blue cells, which would result in a total suppressed value from secondary confidentiality of 11. However, if one looks at the table, it is obvious that an even better approach would be to suppress B2, C2 and B3. This shows that it is very hard to ensure an optimal secondary confidentiality procedure, and solving the issue manually becomes very difficult with more complex tables. A higher number of cells and dimensions only multiply this: large detailed and multidimensional tables are *ceteris paribus* associated with high-numbers of risky cells and a time-consuming confidentiality implementation process. The result is often tables with low data utility that are time-consuming for the producer to compile.

4. Access to micro data

The data utility of many standardized statistical tables – either due to many suppressed cells or to simplified tables to avoid suppression – is too low for many advanced users of statistics. They pre-

fer access to the underlying micro data. Generally, NSIs accept this as a legitimate need and facilitate such access under certain strict conditions. Statistics Denmark, for example, grants access to micro data through remote access to a server placed at Statistics Denmark. The server is separated from the production network and only contains anonymized micro data. Access is limited to researchers and analysts from pre-approved research environments (e.g. in universities, sector research institutes, government agencies, consultancies firms, NGOs, etc.). Like in most (if not all) other countries, this service stops at the border. Foreign and supranational research environments need not apply.

Internationally, work is underway to change this. An OECD expert group aiming to encourage and facilitate cross-border access to official micro data has identified a set of recommendations for achieving this. According to the expert group none of their recommendations are "in themselves radical. Each can be adopted with relatively small changes in current practice. No new legislation, nor substantial new infrastructure, nor new technology is called for[3]." Instead they point to the lack of cross-border trust and encourage NSIs to widen "inner circles of trust" to include cross-border arrangements. This would support trans-national analysis and policy making and potentially provide a deeper understanding of cross-border phenomena, e.g. "globalization".

5. The two pilot studies

Eurostat's Trade by Enterprise Characteristics (TEC) statistics provide data on international trade at the enterprise level by linking International Trade in Goods Statistics (ITGS) with the Statistical Business Register (SBR). This is an important development since the statistics provides answers to the question: "What kind of enterprises are behind international trade flows?" (in terms of size, business activity and ownership). Another important aspect of TEC is the fact that new statistics can be made without increasing the burden of the data suppliers, by linking two existing sources. TEC started out as a series of voluntary pilot studies for the reference years 2005 and 2006 carried out by 19 EU member states. Since reference year 2009, compilation of TEC has been mandatory for all EU members, which provide data in tabulated form once every year (no later than reference year plus 18 months).

[3] [2] OECD (2014), "Export group for international collaboration on microdata acces", Executive summary, STD/CSSP(2014)9

In 2013, Eurostat invited member states to participate in a pilot study aimed at expanding the existing TEC framework by introducing additional tables with new indicators. The year before, Eurostat had launched a pilot study to expand the TEC framework (under the name S-TEC) from focusing solely in trade in goods to also include trade in services, an increasingly important aspect of international trade. Below we discuss these two pilot studies[4] with a focus on challenges presented by confidentiality issues.

5.1 Experience from the TEC exercise

In the TEC pilot study, three new tables where proposed. Two were three dimensional and one was two dimensional. The impact from confidentiality on all three tables was high (see table 1 and 2).

Table 1: The proposed TEC tables

Table	Dimensions	Number of cells	Pct. of confidential cells	Impact from confidentiality
1	Type of trader, Activity, Geography	351	23	High
2	Export intensity, Activity, Geography	819 (701)	25	High
3	Activity (increased detail), Geography	684	30	High

It is clear that for the proposed TEC-tables 1 and 2, the combination of three dimensions and a relatively high level of disaggregation of business activity (into 39 categories) means that they are very sensitive to confidentiality procedures. Table 3 is equally sensitive to confidentiality (if not more so) although only two-dimensional. However, the level of disaggregation of the business activity dimension is near-complete with 228 categories (see table 2).

Table 2: Detail level of the dimensions for the TEC tables

Dimension	Detail level	Hereof totals and subtotals
Type of trader	5	2
Export intensity	7	1
Activity	39 (228)	Not measured
Geography	3	1

The main findings from this pilot studies was that the level of multidimensional disaggregation in the proposed tables decreases the relevance of the datasets for the users, while the compilation process becomes disproportionately taxing for the producer. The level of granularity simply becomes too high. Instead we suggest decreasing the level of granularity by only disaggregating tables along

[4] [3] Løve, F., Nielsen, M., Mortensen, J., Burman, S (2013), "Linking micro data on external trade and business statistics", MEETS Grant agreement number 20721.2013.002 – 2013.158

one dimension at the time. This would ensure a more appropriate relation between data utility and the level of resources needed in the production process.

5.2 Experience from the S-TEC exercise

In the S-TEC pilot project six tables were proposed. Of these six tables half is three dimensional tables and the other half is two dimensional tables, ranging from 80 cells to 1080 cells for each flow. As it can be seen in table 3 below, the three dimensional tables (table 1, 2 and 3) is highly affected by the confidentiality procedures, having a large share of their trade suppressed, compared with the two dimensional tables, with the exception of table 4, which also is severely affected by the confidentiality procedures. Table 4 needed additional effort due to complex "inter-table" relations that made it possible to derive suppressed cells in table 4 by using totals from table 1, 2 and 3.

Table 3: The proposed S-TEC tables

Table	Dimensions	Number of cells	Avg. pct. of suppressd cells	Impact from confidentiality
1	Size, Activity, Geography	270	25	Severe
2	Servicetype, Activity, Geography	1080	6	High
3	Ownership, Activity, Geography	324	36	Medium
4	Activity (increased detail), Geography	333	27	Severe
5	Servicetype, Size	100	12	Low
6	Servicetype, Ownership (decreased detail)	80	13	Low

In table 4 below the detail level of the different dimensions is shown, and even though the detail level of the dimensions are relatively aggregated (at least compared to the corresponding TEC tables), the impact of an extra dimension is surprisingly high. E.g. if the geographical dimension is removed from the tables (only having three values (EU, non-EU and total)), the impact from suppressed cells would decline rapidly for the tables, and if one of the larger dimension, such as size, servicetype, ownership or activity was removed, practically no trade would be suppressed.

Table 4: Detail level of the dimensions for the S-TEC tables

Dimension	Detail level	Hereof totals and subtotals
Size	5	1
Servicetype	20	4
Ownership	6 (4)	2 (1)
Activity	18 (111)	1 (22)
Geography	3	1

The main findings in the S-TEC pilot project is that it is possible to compile advanced S-TEC tables, but that the two dimensional tables have the best value for money in the sense that they can be

produced relatively easy, with little burden on compilers, and will still be very valuable for users, since this information cannot be found elsewhere. It is possible to make the three dimensional tables, but the benefit of another dimension is rapidly decreasing, both due to the number of suppressed cells, but also the absolute amount of trade that are being suppressed. Another important finding is to make sure that "inter-table" relations are taken account when designing tables. Having complex relations between tables, results in an unnecessary complexity when performing secondary confidentiality procedures, resulting in less optimal secondary confidentiality procedures and therefore more suppressed cells.

6. Concluding remarks

The ambition to create more information without collecting additional data by combining existing sources is a win-win situation for both users and suppliers of statistics. The TEC and S-TEC pilot projects aim at shedding light on the nature of the enterprises engaged in international trade, but most of the proposed tables are too detailed (at least for a small country like Denmark), resulting in a high level of suppressed cells in order to ensure confidentiality. Many of the proposed tables are three dimensional, drastically increasing the detail level. Adding more dimensions to the tables increases the effort needed to apply confidentiality, since each new dimension added will multiply its detail level, resulting in rapidly increasing number of cells, which *ceteris paribus* will increase the number of risky cells. Highly disaggregated dimensions have the same effect.

Producing these tables is discouraging for the compiler since a taxing process of producing the tables is followed by an equally taxing process of 'destroying' their utility in the confidentiality process. Generally, when designing tables built on micro-linked data, our analysis shows that there are clear limits to the level of granularity of the tables. We realize that the appropriate level of granularity is likely to be much lower than the user (and the advanced user in particular) would prefer, but confidentiality cannot be compromised. We also agree that the use of micro data is essential in increasing our knowledge of a given issue. We, however, believe that a more appropriate avenue for pursuing this ambition is to further cross-border access to official micro data, e.g. by giving trusted international organizations access to micro data as is already the norm domestically in many countries. Designing highly disaggregated, multidimensional tables is not a cost-effective use of resources and it might very well be meet with disappointment by the users due to extensive confidentiality issues.