Discussion Paper

Statistical Disclosure Control and special focus groups: a European perspective

Peter-Paul de Wolf and Eric Schulte Nordholt

February, 2023

Contents

1 Introduction 4
1.1 Background 4

2 Statistical Disclosure Control for special focus groups 5
2.1 Definitions 5
2.2 Concept of disclosure 5
2.3 Public interest 5
2.4 Utility aspects 6

3 Some European examples 7
3.1 Special focus groups 7
3.2 Small groups 9

4 Summary, conclusions and outlook 10

References 11

Summary

With the availability of larger and more diverse data sources, Statistical Institutes in Europe are inclined to publish statistics on smaller groups than they used to do. Moreover, high impact global events like the COVID-19 crisis and the situation in Ukraine may also ask for statistics on specific subgroups of the population. Publishing on small, targeted groups not only raises questions on statistical quality of the figures, it also raises issues concerning statistical disclosure risk. The main point we want to make is, that the *principle* of statistical disclosure control does *not* depend on the size of the groups the statistics are based on, but the *level* of disclosure risk *does* depend on the group size: the smaller a group, the higher the risk. Traditional ways to deal with statistical disclosure control and small group sizes include suppressing information and coarsening categories. These methods essentially increase the (mean) group sizes. More recent approaches include perturbative methods that have the intention to keep the group sizes small in order to preserve as much information as possible while reducing the disclosure risk sufficiently. In this paper we will mention some European examples of special focus group statistics and discuss the implications on statistical disclosure control. Additionally, we will discuss some issues that the use of perturbative methods brings along: its impact on disclosure risk and utility as well as the challenges in proper communication thereof.

Keywords

Disclosure, Risk, Focus groups, Perturbation.

The views expressed in this paper are those of the authors and do not necessarily reflect the policy of Statistics Netherlands.

The authors thank Naomi Schalken and Nynke Krol for their useful feedback when reviewing a preliminary version of this paper.

This paper was presented at the international Methodology Symposium **Data Disaggregation: Building a more representative data portrait of society**, organized by Statistics Canada, 2-4 November 2022.

1 Introduction

1.1 Background

In Europe National Statistical Institutes (NSIs) are connected within the European Statistical System (ESS). Within the ESS, Eurostat plays the role of a coordinating European institute that combines the national statistics to European statistics. To achieve that goal, Eurostat tries to harmonize the kind of data to be collected as well as the level of dissemination of statistics by the member states of the European Union. The latter mainly regarding the categories of variables used in tabular publications. Some of the statistics NSIs produce are mandatory in the sense that they are part of European regulations.

NSIs obviously will apply Statistical Disclosure Control (SDC) techniques to their national publications in order to prevent the release of information that can be related to identifiable individual units. As mentioned before, some of these data are sent to Eurostat to be combined into European statistics. These European statistics then need a second round of SDC techniques to prevent recalculation of data of one country by subtracting data of other countries from the total.

Traditionally, NSIs (and Eurostat) disseminate their statistics in a tabular format, i.e. as aggregated data. Nowadays these tabular data may sometimes also be presented using visualisations. In the latter case the visualisations are often based on tabular data to which SDC techniques have already been applied. A recent trend in dissemination is that more and more detail is being asked for. Researchers, policy makers and governmental institutes are all eager to receive detailed information. This leads to statistical information related to small groups of units. High impact global events, like the recent COVID-19 crisis and the situation in Ukraine may also ask for statistics on specific subgroups of the population.

Publishing on small, targeted groups of units not only raises questions on statistical quality aspects like accuracy, it also raises issues concerning statistical disclosure control. From the point of view of SDC, some ways of addressing the need for more detailed statistics were developed over the years. For example, for researchers it has become possible to directly access the microdata (in a safe way) in order to be able to apply their own models to the data. Examples are Scientific Use Files and Secure Use Files that were made available to authorized researchers. To prevent disclosure of individual information, Scientific Use Files did not always contain all survey variables, and some variables only in a less detailed format. Secure Use Files were usually richer in content compared with Scientific Use Files, but were accessible only in a research data centre (on-site) or via a secure remote access environment. Both Scientific Use Files and Secure Use Files are still in use by authorized researchers.

To facilitate more detailed information to others than authorized researchers (much) more detailed tabular data is provided, such as census information at the geographical level of 1 km × 1 km grid cells. Such grid cell data is for example mentioned in an implementing regulation on the European Census 2021 (see European Commission, 2018), which makes it mandatory for the member states to deliver statistics at the level of these grid cells to Eurostat.

2 Statistical Disclosure Control for special focus groups

2.1 Definitions

In this paper we use the term 'special focus groups' when we talk about either small groups of units (e.g. households on 1 km × 1 km grid cells in rural areas) or of specific groups of units (e.g. COVID-19 patients or Ukrainian refugees). Moreover, we use Statistical Disclosure Control (SDC) to denote the field of research that deals with assessing the risk of disclosing information on identifiable units on the basis of statistical output and with applying techniques to reduce that risk of disclosure. A synonym for Statistical Disclosure Control is Statistical Disclosure Limitation (SDL). For an overview of risk and utility measures and SDC methods, see e.g. Hundepool et al. (2012).

2.2 Concept of disclosure

The need for assessing the disclosure risk of statistical publications is partly driven by the fact that several laws and regulations demand this. In Europe e.g. the GDPR (see European Commission, 2016) concerns the protection of personal data, where personal data are defined as *any information relating to an identified or identifiable natural person ('data subject')*. Most member states of the EU have their own statistics acts that include the obligation for National Statistical Institutes (NSIs) to protect the privacy of their respondents. For example, articles 37 – 42 of the Dutch Statistics Act[1)] regulate the way Statistics Netherlands is allowed to process and publish statistical information. Article 37 for example states that data received by Statistics Netherlands *shall only be published in such a way that no recognisable data can be derived from them about an individual person, household, company or institution, unless, in the case of data relating to a company or institution, there are good reasons to assume that the company or institution concerned will not have any objections to the publication.*

The formulation of the way NSIs should deal with data provided to them and the way they are allowed to publish statistics based on these data, does not refer to the origin nor to the size of the group of units the statistic is about. The concept of statistical disclosure for special focus groups is thus not different from that for 'regular' statistics. Obviously, small groups may be more easily susceptible to disclosure: small groups and special focus groups will violate the restrictions of the GDPR and of national statistics acts more easily compared to regular statistics. But the *concept* is exactly the same.

2.3 Public interest

In the GDPR, article 6 (1)(e) states that data controllers may process personal data if *processing is necessary for the performance of a task carried out in the public interest*. Moreover, relating to special categories of personal data, article 9 (2)(g), states that processing is allowed if it is

[1)] An English translation can be found at https://www.cbs.nl/en-gb/about-us/organisation

necessary for reasons of substantial public interest, on the basis of Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject and article 9 (2)(i) states that it is allowed if it is *necessary for reasons of public interest in the area of public health*. This shows that under special circumstances, referring to public interest, processing of (special) personal data can be allowed under the GDPR.

The risk of disclosure is a combination of the probability that a disclosure occurs and the impact the disclosure may have on the related individual units. It is virtually impossible to publish useful information with zero risk of disclosure. Indeed, the fundamental law of Statistical Disclosure Control (as formalised in Dwork and Naor, 2010 as a rigorous impossibility of Dalenius's privacy goal), essentially states that any provision of *useful* statistical information *necessarily* entails a *nonzero* residual risk of disclosing information on some individuals. Therefore, NSIs should have a policy on how much residual risk is acceptable. The choices made for such policy may depend on the sensitivity of the variables in question, the importance of the results for the general public and many more arguments.

Even though this may be considered as an ethical issue, in case of special focus groups the importance to the public good may outweigh the impact on the individuals of those groups. Hence, it may be that the *concept* of Statistical Disclosure Control for (special) focus groups is not different from the regular situation, but the chosen *acceptable residual risk* may be. This may for example be reflected in the way parameters of risk measures are chosen.

2.4 Utility aspects

As a dual phenomenon to the just mentioned fundamental law of statistical disclosure control, application of any Statistical Disclosure Control (SDC) method reduces the amount of utility left in the published statistical information. Following the *concept* of disclosure, small groups often call for severe application of SDC methods. Consequently, utility is usually very much affected when publishing about small groups.

Application of such methods to small groups does not only affect the utility of statistics on those groups themselves, it may also affect the utility of other statistics. Especially when the statistics on small groups are asked for following unexpected events, i.e. as unplanned publications outside the regular publication scheme. These kind of topical phenomena are nonetheless often related to other, regular statistics. Applying SDC methods to protect the statistics on small groups may limit the possibilities of publishing the same statistics on larger, related groups in a safe way. Indeed, the publication of statistics on the larger groups should not only be safe on its own, but it should also not hamper the protection of the publication on the small groups. This interrelationship between the protection of statistics on small groups and other statistics not only entails statistics that are related topic wise, but also statistics that are related time wise. In all situations, publications should be protected consistent with or *conditionally on* the protection of related or previously published statistics.

Considering SDC as an art to publish as much as possible without disclosing individual information, NSIs will try to minimize information loss given a maximum allowed residual disclosure risk. As mentioned before, special focus groups often intrinsically ask for a rather

severe application of SDC methods, which may affect the utility of those statistics considerably. This can be implemented using 'traditional' SDC methods, like coarsening of categories and suppression of table cells, but just as well using more recent SDC methods like perturbation and noise addition. For an overview of 'traditional' methods, see Hundepool et al. (2012). Examples of perturbative methods are Controlled Tabular Adjustment (see e.g. Dandekar and Cox, 2002 or Castro, 2006), noise addition using the Cell Key Method (see e.g. Fraser and Wooton, 2006), Targeted Record Swapping (see e.g. Shlomo et al., 2010) and methods based on Differential Privacy (see e.g. Dwork and Naor, 2010).

An additional complexity comes from the observation that utility is not an absolute term: not all users have the same preferences. Furthermore, utility can often not be defined for a single publication: users often want to compare data over time and/or between different populations. Long term utility aspects could then limit short term utility aspects. This holds for example for population and housing census data that should be comparable over time and between countries (of the European Union).

3 Some European examples

3.1 Special focus groups

Within the European Union, recent examples of special focus group statistics entail statistics on Ukrainian refugees and statistics on the effects of the COVID-19 crisis. We will elaborate on these two examples in the following subsections.

3.1.1 COVID-19 crisis

The COVID-19 crisis has disrupted the life of many European Union (EU) citizens. In particular in the spring of 2020, as most EU governments imposed lockdowns on their citizens as a measure to stop the virus from spreading. During that time, working life took place online or got interrupted for a while. In order to counteract the effect of those lockdowns on their employees, most EU governments enacted also temporary relief measures. As the situation worsened again during autumn 2020, most governments decided to continue the relief measures. Examples of such governmental relief measures include:

- Short term work and assimilated schemes, where employees work a lower number of hours and are (partly) compensated for hours not worked;
- Temporary lay-offs, where employees keep an attachment to their jobs in the form of employment contracts, but are temporarily not working.

The funding of these measures could be directly between employee and government, or indirectly where the employer is compensated by the government and assigns the funding to its employees.

Statistics requested by the European Union on these measures were for example the number of jobs benefiting from these measures. Information was collected on the total number of jobs that requested to be supported, that authorized to be supported and that actually benefited from the

measures. The statistics were usually requested to be published per individual NACE[2)] sector. Additionally, statistics were asked about the number of local units using the benefits and the number of hours not worked. Where possible these statistics were normalized by the total number of jobs, total number of local units and total number of hours worked, in order to facilitate country comparisons. The data were collected on a monthly basis starting from January 2020, where the reference dates were the last day of each month.

Some of the complicating factors in producing these statistics on a European level include:

- The data were collected by different national public authorities. Data regarding the total number of jobs either came from the NSI or from the EU Labour Force Survey. Data regarding local units came either from national sources or from the EU Structure of Business Survey.
- There was no guarantee that the data were harmonized. There were important differences on the rules determining whether the funding would be direct or indirect.
- Participation varied by indicator and over time. The highest participation was on the total number of jobs supported by governmental measures, actually used or approved. At the beginning of the pandemic (April-June 2020) a total of 23 member states participated. The lowest participation was for the number of hours not worked, where only seven member states delivered data to Eurostat.

Not only the decreasing influence of the pandemic on the economy, but also the fact that access to the required data was sometimes problematic because NSIs were usually not the producers and owners of the data, resulted in ending this European data collection during 2021.

3.1.2 Ukrainian refugees

A Temporary Protection Directive was activated in March 2022 (see European Commission, 2022). This directive regulates the data collection on temporary protection statuses granted during the reference month to persons fleeing Ukraine as well as the stock of beneficiaries of temporary protection statuses at the end of the reference month. The data are to be transmitted to Eurostat within one month starting from the end of the reference month.

The data for the first quarter in 2022 was transmitted to Eurostat by the end of May 2022 and published at the beginning of June 2022. The annual data for 2022 will be transmitted to Eurostat by the end of March 2023 and published in the beginning of April 2023.

Although this legal framework has helped to start collecting European data on persons fleeing Ukraine, in practice it took quite a while before data of the Member States arrived in the Eurostat office. Moreover, the speed of implementing has led to issues regarding comparability of data between different countries. Nevertheless, in a relatively short time, relevant data could be published. These data are less detailed than regular detailed census outputs that are published less timely. This implies that different users are served using data with different levels of detail and with different timeliness. Finally, it is clear that the more detailed census data require more protection and that the way these more detailed data are protected depends on what has already been published earlier.

2) NACE = Nomenclature statistique des Activités économiques dans la Communauté Européenne

3.2 Small groups

A recent illustrative example of statistics on small groups are the statistics on small area groups in the European censuses. In the next subsection the newly introduced statistics at the level of 1 km × 1 km areas in the 2021 European census are described.

3.2.1 Small area groups in European censuses

The year 2021 was a European Census year. This implies that all member states of the European Union (EU) had to conduct a Population and Housing Census with a reference day in 2021 (Census Day). This is an important means to harmonise European census results. Moreover, all EU countries will at least publish a set of harmonised tables to make comparisons well possible. This set of linked high dimensional tables gives a precise description of the people living in the EU and their housing situation. This is called the set of European Census 2021 hypercubes.

Additionally, in the European Census 2021 for the first time also a set of 1 km × 1 km grid squares tables was mandatory. These tables are not detailed in content (for each of these tables only one characteristic is included), but detailed in structure (the number of grid squares is in all countries much larger than the number of municipalities). Moreover, grid squares and regional distributions are non-nested variables. This implies that countries have to check whether information about individuals can be disclosed by crossing these grid squares with municipalities (LAU2), the most detailed level of region in the European hypercubes. Other levels of region in the hypercubes – Country, NUTS1, NUTS2 and NUTS3[3)] – are combinations of LAUs. These regional levels are nested, i.e. they show a hierarchical structure. The grid squares however is a regional variable that is not nested in this hierarchical structure.

Classical non-perturbative methods like global recoding and cell suppression are for different reasons no solution to protect the European census tables. To make comparisons between countries possible the table formats are fixed and cannot be altered. Therefore, global recoding is not an option. Applying cell suppression to such a large set of high dimensional linked tables in an optimal way is practically impossible. Theoretically, it would be possible to apply cell suppression, but the many interrelationships between cells in rows, columns, layers, subtotals, etc. would lead to a lot of over-suppression. This would thus in turn cause a huge information loss which is unacceptable from a user's point of view. Another problem is the management of differencing risk between hypercube and grid level data, adding further to the complexity of cell suppression based protection concepts.

Based on experiences in many individual countries inside and outside the EU two methods were recommended: the pre tabular Targeted Record Swapping (TRS) and the post tabular Cell Key Method (CKM). Both methods essentially add noise to the table cells. Even though the methods are both recommended to be used to protect the census hypercubes in a harmonised way, different member states of the EU still have some freedom to decide on how to use these methods. Not only can they choose different parameter values, they can also decide to use only one of the methods or a combination of both methods. Indeed, given the different confidentiality rules applicable in different European countries as well as the difference in the sizes of these countries, it was advisable to recommend not just a single method. However, by limiting the number of recommended methods, it will be easier for Eurostat as well as other

3) NUTS = Nomenclature des Unités Territoriales Statistiques, LAU = Local Administrative Units

users, to compare protected census statistics between countries. Both methods do not lead to suppressed data, therefore the member states' data, if treated by these methods, can be combined into European-level data.

As mentioned before, the protection of the small area groups influences the way the other publications (the hypercubes) should be protected. Moreover, the publication of the grid cells and the general hypercubes take place at different points in time: the grid cells were supposed to be transmitted to Eurostat in December 2022, whereas the census hypercubes are expected to arrive at Eurostat in March 2024 at the latest.

Furthermore, there is a link between the census publications and the regular (national) population and demography statistics. The application of recent SDC methods like TRS and/or CKM to the census publications are not easily transferred to all related population and demography statistics. This again is an example of how SDC methods applied to focus groups and/or small groups of units will influence publications of other statistics.

4 Summary, conclusions and outlook

In this paper we discussed the case of statistics on (special) focus groups and/or small groups of units. We argued that the *concept* of Statistical Disclosure Control (SDC) is not different compared to the situation of 'regular' statistics. Only the *norm* that National Statistical Institutes (NSIs) may use in their policy on accepted residual disclosure risk may differ for statistics on (special) focus groups and 'regular' statistics. For example on the basis of the relevance and importance to the public interest of topical and timely information on specific groups of individuals. Some European examples were given of statistics on special focus groups (Ukrainian refugees) and on small groups (1 km × 1 km grid squares tables in the European Census 2021).

Using the (standard) concept of SDC on small groups of individual units shows that intrinsically such statistics often require severe application of SDC methods to make publications possible that have only limited disclosure risk. Recent developments go into the direction of using perturbative methods, like noise addition, as alternatives to the traditional SDC methods like coarsening of categories and suppression of table cells. Within the 2021 European Census projects, Targeted Record Swapping (TRS) and noise addition using the Cell Key Method (CKM) are proposed in an attempt to further harmonize the SDC techniques used by the member states of the European Union.

We argue that applying SDC methods to publications on special focus groups and small groups of individual units affects the protection of other, related publications. For example census hypercubes and grid cell tables can still have some common marginals, even though the respective regional variables are not nested: at least the national totals occur in both regional classifications. Additionally, the publication of the grid cell tables are to be published way ahead of time of the publication of the hypercubes. Moreover, there are other related statistics that are published by the NSIs themselves concerning population statistics and demography. In general, related publications should be protected consistently. When publications are not published at the same time, this means that the latest publications should be protected conditionally on the already published data. In case the related publications are published simultaneously, ideally the same SDC methods should be applied to protect the related publications.

In the near future, both demographic statistics and census statistics will fall under one European legal umbrella and thus need to be protected consistently. Lessons learned from the current experiences with the protection of special focus groups and small area groups in the census and the implications on other, related, statistics could and should be input to that more unified approach. Looking even further into the future, the amount of success of this unified approach might give rise to the attempt of defining unified approaches in other fields of European statistics as well.

Despite the way SDC is currently applied to special focus groups and small groups of individual units, there are still some open issues. Next to TRS and CKM as being suggested to be used in the European Census, the notion of differential privacy is used in the US Census. Essentially, that notion also adds noise to census tables. A comparison and possible pros and cons of each method should be investigated.

Adding noise to protect statistics for disclosure of individual information in general publications is not only relatively new to statistical offices, it is also new to users of the published data and to the providers of the individual data (respondents). This obliges NSIs to clearly and correctly inform the users and the respondents. It is not that easy to explain in what way noise addition really protects against disclosure of individual information nor in what way the perturbed statistics are still useful. Often users obviously consider 'official statistics' as the official figures and 'thus' as exact figures, even though 'official statistics' are estimates and thus have uncertainty by definition. On the protection side of the communication aspect it should be noted that perturbed figures may give rise to different opinions. One could be that a respondent concludes that the statistical office does not present his or her response correctly, especially in case a respondent claims to identify him or herself in the publication. It should then somehow be made clear that in general anyone who sees that information cannot be certain about the individual disclosure because of the (possibly) added noise: only the respondent himself can be sure about the correct information.

A second aspect concerning the use of perturbative methods and noise addition is the way researchers getting access to microdata for statistical or scientific research at the premises of an NSI or via a secure remote access connection should deal with the data. Should they also add noise to their results, consistently with the officially released publications? In case pre tabular methods are used, should the researcher use the perturbed microdata? In any case, their publications should be safe on their own as well as in relation to the already published statistics.

Despite the fact that perturbative methods like noise addition seem to become more popular when dealing with special focus groups and small groups on individual units, there is no room for complacency. The growing popularity of microdata research and the need for topical and timely statistics on specific groups in specific circumstances, still requires NSIs and Eurostat to continuously check and update their privacy measures. The aim will always be to improve the possibilities for scientific research and to serve the public interest by safe and useful official statistics.

References

Castro, J. (2006). Minimum-distance controlled perturbation methods for large-scale tabular data protection. *European Journal of Operational Research* 171, 39–52.

Dandekar, R. A. and Cox, L. H. (2002). *Synthetic Tabular Data: an alternative to complementary cell suppression for disclosure limitation of tabular data*. Technical report. Manuscript can be found at https://www.researchgate.net/publication/228907825_Synthetic_tabular_data_An_alternative_to_complementary_cell_suppression.

Dwork, C. and Naor, M. (2010). On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality* 2, 93–107.

European Commission (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union* L119, 1–88.

European Commission (2018). Commission Implementing Regulation (EU) 2018/1799 of 21 November 2018 on the establishment of a temporary direct statistical action for the dissemination of selected topics of the 2021 population and housing census geocoded to a 1 km2 grid. *Official Journal of the European Union* L296, 19–27.

European Commission (2022). Council Implementing Decision (EU) 2022/382 of 4 March 2022 establishing the existence of a mass influx of displaced persons from Ukraine within the meaning of Article 5 of Directive 2001/55/EC, and having the effect of introducing temporary protection. *Official Journal of the European Union* L71, 1.

Fraser, B. and Wooton, J. (2006). A proposed method for confidentialising tabular output to protect against differencing. In *Monographs of Official Statistics, Work session on Statistical Data Confidentiality*, pp. 299–302. Eurostat-Office for Official Publications of the European Communities, Luxembourg.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., and de Wolf, P. P. (2012). *Statistical Disclosure Control*. Wiley series in Survey Methodology. John Wiley & Sons, Ltd. ISBN: 978-1-119-97815-2.

Shlomo, N., Tudor, C., and Groom, P. (2010). Data Swapping for Protecting Census Tables. In J. Domingo-Ferrer and E. Magkos (Eds.), *International Conference on Privacy in Statistical Databases, PSD 2010, Corfu, Greece*, Lecture Notes in Computer Science (LNCS 6344), pp. 41–51. Springer, Berlin, Heidelberg.

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Grafimedia

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information