Discussion Paper

Privacy Preserving Techniques and Statistical Disclosure Control

Peter-Paul de Wolf and Bob van den Berg

November 25, 2021

Contents

Summary

This paper was discussed in the Advisory Board on Methodology in 2020. It discusses the privacy related fields of research Privacy Preserving Techniques (PPT) and Statistical Disclosure Control (SDC). PPT concerns techniques that can be used to share data among different parties in a privacy preserving way. This is sometimes referred to as Input Privacy. SDC on the other hand concerns techniques that can be used to limit the risk of disclosing information on individual entities from statistical publications. This is sometimes referred to as Output Privacy. These two fields of research thus obviously complement each other. In this paper we describe the relationship as we see it fit for use by a National Statistical Institute like Statistics Netherlands. Moreover, we provide two examples of recent research in SDC and PPT: 'SDC when publishing on thematic maps' and 'Privacy-Preserving Infrastructure for analyzing personal health data in a vertically partitioned scenario'. We end this paper with some issues that are still open for discussion.

Keywords

Privacy Preserving Techniques (PPT), Statistical Disclosure Control (SDC), thematic maps, health data.

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

1 Introduction

National Statistical Institutes (NSIs) have always had the obligation to protect the confidentiality of the information provided to them by respondents. Often this is regulated by the national Statistical Law. Other governmental institutes, health care facilities and in general data providers face similar challenges. Since the European General Data Protection Regulation (GDPR) came into force in May 2018, once again attention was drawn to the confidentiality of individual (personal) data. A recent discussion between Statistics Netherlands and the Dutch Data Protection Authority on the use of mobile phone data for analysis in view of the Covid 19 situation also shows that privacy and confidentiality are topical issues.

Other current developments concern Open Data initiatives, Big Data projects and the ever growing need for (detailed) 'facts' by local and national policy makers. These developments ask for new ways to cooperate: sharing data among different parties to reduce the response burden while taking advantage of the full potential of the data.

The challenge is to facilitate cooperation between Statistics Netherlands (CBS) and other research or policy making institutes, while taking the privacy and confidentiality restrictions into account. To that end, several initiatives have been started. In this report we will introduce some initiatives and we will contemplate about the effectiveness and usefulness of them for NSIs. We will focus on Privacy Preserving Techniques (PPT) and their connection with Statistical Disclosure Control (SDC).

2 General model

In general the statistical process can be split into an input phase, a throughput phase and an output phase. Similarly, cooperation between partners (i.e., CBS and external research or policy bodies) starts in the input phase and continues all the way to the output phase.

Article 25 of the GDPR discusses 'Data protection by Design' and 'Data protection by Default' in relation to processing personal data. One of the consequences is that we should make sure that the statistical process and the cooperation process by default takes into account data protection and is implemented when designing the process.

One way of cooperation is Multi Party Computation: multiple parties collaborate by using and combining each others data, leading to the desired output. Traditionally this is accomplished by making use of a Trusted Third Party (TTP). Such a TTP would collect all the data, combine or link the data, do the computations and share the results. That way, the TTP is the only party that has full access to all the data and is able to check the results for disclosure before sending them to the participating partners. Effectively, CBS often functioned as TTP in cooperating with external research institutes: we linked the data (if necessary), did the computations (or had external researchers perform the analyses via Remote Access) and we did the check of the output on disclosure.

The fields of Computer Science and Cryptography have been working on approaches to circumvent the need for a TTP. These approaches are often called Privacy Preserving Techniques (PPT). In essence these approaches make it possible to perform calculations without physically sharing readable versions of data among the partners. These approaches often involve considerably more calculations and/or communication rounds compared to the TTP approach. Recent increase in computer power has made those PPT more feasible solutions and thus drew more attention. Also SN has been investigating such approaches, e.g., making use of homomorphic encryption.

In PPT the main goal is to share data in such a way that the collaborating partners are not able to see the original data of the other partners. However, as we know from years of experience with Statistical Disclosure Control (SDC), statistical output in itself or in combination with other publicly available information may disclose (personal) information as well. Thus, making sure that the data are not readable during the computations is not sufficient to be able to claim 'Data protection by Design'. This shows that PPT and SDC are not competing fields but complementary to each other.

In Figure 2.1 we have drawn a high over view of a statistical process and the (possible) places where PPT and SDC come into play. Within PPT a possible distinction can be made between Privacy Preserving Data Sharing (PPDS) and Privacy Preserving Analysis (PPA). PPDS is targeted at safe ways of *sharing* data, whereas PPA aims at doing *analyses* in a safe way, preferably with shared data. Obviously these two are interwoven but could be considered separate techniques. For example, Secure Multiparty Computation (SMC) using Secret Sharing could be considered as a PPA technique since it typically involves calculations (analysis), whereas Data Virtualisation and Compartmentalisation could be considered as a PPDS technique.

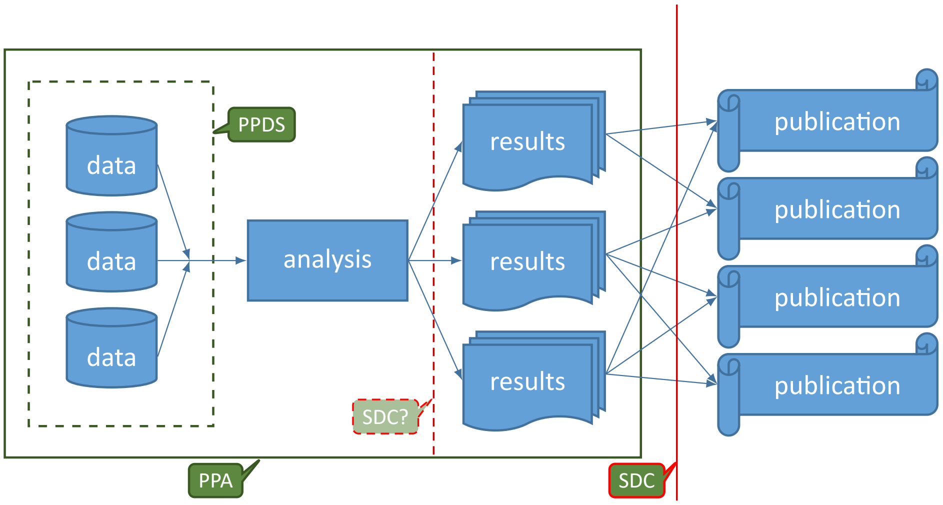

Figure 2.1 Where to apply PPT and SDC

In Figure 2.1 it is evident that SDC is still needed before publication of results, whether PPT is used or not. However, within the framework of PPA it may be needed that intermediate results are shared among the participating parties e.g., to decide on following steps in the research project. As a modern example where this is an issue one might think of Federated Learning.

However, these intermediate results may be disclosive (PPA does not prevent that by default). So it might be needed to include some SDC steps *inside* a PPA approach.

3 Examples

In this section we briefly mention examples of projects that have taken place at SN, either as pure research or as a Proof of Concept. We present one example in the field of SDC and one example in the field of PPT.

3.1 Statistical Disclosure Control when Publishing on Thematic Maps

SDC already has a rich history for application to ‘traditional’ NSI output, like tabular data and micro data files. More modern types of output like network information and cartographic information call for new approaches. This example shows recent developments in publishing safe cartographic information.

In the past, when regional data were plotted on a cartographic map it was based on tabular data that had passed tabular data protection (SDC applied to the tabular data). That way you end up with uniformly coloured (administrative) areas (choroplots) or with maps where circles whose size would reflect the size of a variable are plotted at certain locations. Another way to think about publishing cartographic information in a safe way, is to apply some SDC technique directly to the plot itself, without first constructing a table. That way it would be easier to plot the information independent of predefined administrative areas or locations.

Recently a master student from University of Twente devoted his thesis to SDC in relation to the use of kernel weighted averages. The working example he used was to plot the energy consumption by enterprises on a map (total consumption divided by number of enterprises per area):

$$m_h(\mathbf{r}) = \frac{\sum_{i=1}^{N} g_i k_h(\mathbf{r}-\mathbf{r}_i)}{\sum_{i=1}^{N} k_h(\mathbf{r}-\mathbf{r}_i)}$$

where g_i is the energy consumption of enterprise i at location $\mathbf{r}_i$ and $k_h(x) = k(x/h)$ a smoothing kernel.

He was able to show that, when an attacker is able to exactly get the value $m(\mathbf{r})$ at the locations of the enterprises, knows the used kernel k and bandwidth h, the attacker can exactly recalculate the originally observed energy consumptions. As a protective technique (SDC technique) addition of noise was considered. To be able to quantify the disclosure risk, the well known p%-rule for tabular data was extended to include the uncertainty of the added noise.

It turned out that the noise to be added was best taken to be inversely proportional to the distribution of the enterprises, i.e., to use

$$m_h(\mathbf{r}) = \frac{\sum_{i=1}^{N} g_i k_h(\mathbf{r}-\mathbf{r}_i) + \epsilon(\mathbf{r})}{\sum_{i=1}^{N} k_h(\mathbf{r}-\mathbf{r}_i)}$$

where $\epsilon(\mathbf{r})$ is generated as a Gaussian random field with mean 0 and covariance function $\sigma^2 k_h(\mathbf{r} - \mathbf{s})$.

Based on his working example, Figure 3.1 shows on the left a smoothed version of the energy consumption density and on the right a protected version. For more information on the used $(10\%, 0.1)$ sensitivity rule, see [1].

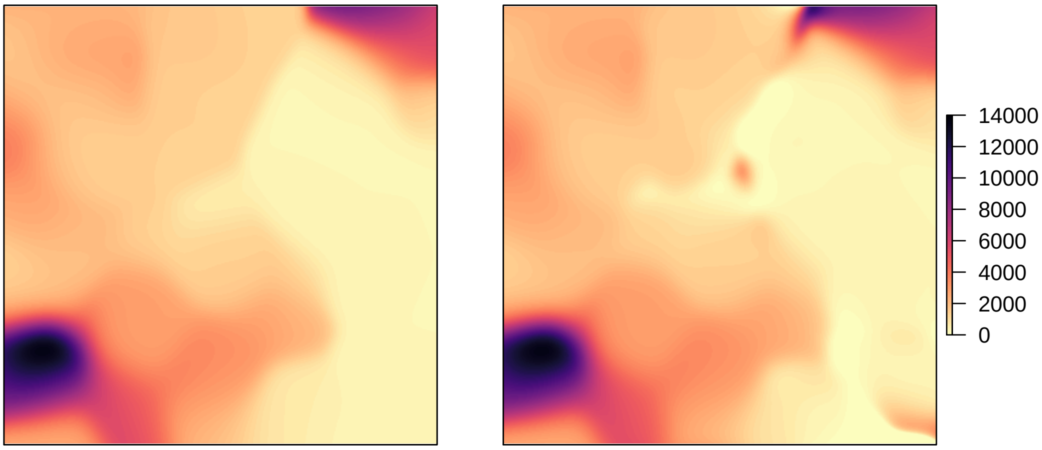

Figure 3.1 Unprotected (left panel) and protected (right panel) kernel weighted average of a part of our synthetic dataset, according to a $(10\%, 0.1)$ rule for a Gaussian kernel with bandwidth $h = 100$ m

3.2 Privacy-Preserving Infrastructure for Analyzing Personal Health Data in a Vertically Partitioned Scenario

Health 'Big Data' are extremely privacy sensitive. Using it responsibly is key to establish trust and unlock the potential of these data for the health challenges facing Dutch society now and in the future. One of the unique characteristics of Big Data in health is that they are extremely partitioned across different entities. Citizens, hospitals, insurers, municipalities, schools, etc. all have a partition of the data and nobody has the complete set. Sharing across these entities is not easy due to administrative, political, legal-ethical and technical challenges.

In this project, CBS partnered with Maastricht University to establish an infrastructure which supports secure and privacy-preserving analysis of personal health data from multiple providers with different governance policies. The objective is to use this infrastructure to explore the relation between Type 2 Diabetes Mellitus status and healthcare costs. We therefore analyze vertically partitioned data from the Maastricht Study, a prospective population-based cohort study, and data from CBS. This project seeks an optimal solution accounting for scientific, technical, and ethical/legal challenges. See [2] for more details on the results of this project.

4 Discussion

At Statistics Netherlands we are working on Statistical Disclosure Control issues related to new types of output as well as new techniques applied to traditional types of output. Moreover, we are experimenting with Privacy Preserving Techniques to safely share and analyze confidential data located at different institutes. We are aware of the fact that SDC and PPT are complementary: PPT cannot do without SDC and SDC cannot do without PPT when dealing with current ways of collaboration between different research parties.

However, a lot of issues are still for discussion.

CBS already provides good (and popular) remote access services for external researchers to work with data from CBS, in some cases combined with data from the researcher itself. PPT aims to develop a new service for situations where 1) the external researcher wants to combine the CBS data with its own data but either cannot or will not provide these data to CBS or 2) CBS wants to use sensitive data from (mostly) private companies for its own statistical products. In the latter case, CBS usually has a legal basis to request the data, but in some cases it might be preferable to use additional privacy preserving techniques to make this work.

PPT somehow tries to circumvent the idea of using a trusted third party (TTP). However, since confidentiality still needs to be checked when releasing (intermediate) results, you might still need a kind of TTP to do the actual output checking. Unless the SDC part could be integrated into the PPT-environment. This seems particularly challenging when intermediate results are to be shared among the different parties while performing the analysis. How feasible is it to include SDC into PPT-techniques? Are participating researchers willing to invest into new ways of programming their analyses (using ‘secure programming languages’)?

PPT has a sound foundation in computer science and cryptography. Computer science has also introduced the notion of Differential Privacy. In order to incorporate SDC into PPT, one could also think about differentially private PPT. How would concepts like ‘privacy budget’ interfere with a setup where PPT is used in cooperative research?

When working on the PPT example project, we have had long discussions on whether we should use encryption techniques that are ‘quantum proof’. Although it will take a considerable amount of time before quantum computers will be widely available, it might be necessary to account for situations where encrypted datasets could be stored and decrypted in that (near) future.

So far we have been looking at a few approaches in PPDS and PPA like Data Virtualisation and Compartmentalisation, SMC with homorphic encryption and SMC with secret sharing.

The SDC example we presented included a new sensitivity measure, the $(p\%, \alpha)$-rule as an extension of the traditional $p\%$-rule to include the uncertainty of the added random noise. Moreover, it introduced a way of plotting a heat-map-like representation of relative distributions. It assumed a particular attacker scenario, where the data snooper knows the exact locations of the points of interest as well as the used (fixed) bandwidth and kernel. Is this a realistic scenario? Is the $(p\%, \alpha)$-rule an adequate sensitivity measure?

References

[1] Hut, D., Goseling, J., van Lieshout, M.C., de Wolf, P.P. and de Jonge, E. (2020). *Statistical Disclosure Control when Publishing on Thematic Maps*. In: Privacy in Statistical Databases, J. Domingo-Ferrer and K. Muralidhar (Eds.), LNCS 12276, Springer, New York, pp. 195 – 205.

[2] Sun, C., Ippel, L., van Soest, J., Wouters, B., Malic, A., Adekunle, O., van den Berg, B., Mussmann, O., Koster, A., van der Kallen, C., van Oppen, C., Townend, D., Dekker, A. and Dumontier, M. (2019). *A Privacy-Preserving Infrastructure for Analyzing Personal Health Data in a Vertically Partitioned Scenario*. Studies in Health Technology and Informatics, 264, 373-377. https://doi.org/10.3233/SHTI190246.

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