UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Confidentiality

26-28 September 2023, Wiesbaden

Protecting High-Resolution Poverty Statistics against Disclosure using Differential Privacy

Raphaël de Fondeville (Federal Statistical Office, Switzerland), Michael Shoemate (Harvard University, USA), Wanrong Zhang (Harvard University, USA), Salil Vadhan (Harvard University, USA)

Abstract

The past few years have seen an explosion of the volume of geo-referenced data, a trend that can be observed in the world of official statistics: large scale imputation, generalizing survey results to the whole population, is made more and more common thanks to the efficiency and the flexibility of new machine learning algorithms. Official agencies are now capable of providing realistic estimates of population characteristics at lower than ever aggregation levels, but communicating survey results at always finer geographical scales strongly increases privacy risks. Thus, in order to maintain trust between populations and their administrations, official statistical offices must ensure highest levels of confidentiality.

In this context, Differential Privacy (DP) has been successfully applied to protect individual's privacy by addition of properly scaled random noise. We first discuss the specificities of DP applied to regionalized statistics and present a baseline framework minimizing the amount of noise necessary to successfully control disclosure risk when releasing spatial aggregates. The technical readiness of the framework is illustrated through a synthetic case study based on Swiss poverty statistics using the OpenDP Library. Finally, we discuss some limitations of the DP framework when controlling disclosure risk of geo-referenced data and present some ongoing themes of research.

1 Context: The SILC Survey and Poverty Indices

The SILC (Statistics on Income and Living Conditions) survey is a recurring annual survey conducted across more than 30 European countries [1]. Its implementation in Switzerland was initiated in 2007 [2]. The sampling methodology employed in this survey follows a proportional, stratified design, carefully organized into seven major geographical regions and currently reliant on the population register. A sample size of approximately 8,500 households, encompassing approximately 18,000 individuals, is typically included in the survey through a 4-year rotating panel design. Data collection is primarily conducted via telephone, although as of 2023, respondents have been provided with the alternative option of completing individual questionnaires online. To address issues related to non-response and loss to follow-up, as well as to ensure that the sample is appropriately calibrated to the reference population, weights are meticulously estimated and subsequently applied.

In conjunction with linked register and income data, the survey responses obtained from the SILC survey are employed to estimate multiple poverty indices, including Absolute Poverty, Relative Poverty, and Material and Social Deprivation. Further information pertaining to these indices can be found in the official report titled "Poverty Measurement in Switzerland" [3]. However, the primary focus of this work is exclusively directed towards the "Absolute Poverty Rate," which embodies a needs-based definition delineating a minimum subsistence level essential for ensuring both physical survival and a basic level of social participation. Individuals are deemed impoverished if they lack the financial means to procure goods and services necessary for a socially integrated life. The comparison between disposable household income and the costs associated with fundamental needs, such as food, clothing, personal care, transportation, entertainment, education, as well as housing and other indispensable expenditures like liability insurance, serves as the criterion for determining the Absolute Poverty Rate. This rate is estimated at the individual level, based on the poverty status of the household. Henceforth, in the subsequent sections of this paper, the terms poverty, poor, and poverty rate specifically pertain to this variable of absolute poverty.

Although the sample size of the SILC survey permits estimations at the level of grand regions (Nomenclature of territorial units for statistics 2), it is deemed insufficient for generating robust estimates at the cantonal level. Notably, Switzerland encompasses 26 cantons, exhibiting substantial disparities in terms of population size. However, the augmentation of the sample size to facilitate cantonal evaluations would entail considerable financial resources that are deemed impractical.

Building upon the pioneering work conducted by Statistics Austria, the Swiss Federal Statistical Office initiated an ambitious innovation project aimed at estimating poverty levels at fine-grained geographical units and furnishing reliable information across various regional levels, encompassing grids, census districts, municipalities, districts, and Nomenclature of territorial Units for Statistics (NUTS) levels. To realize this objective, a selection of supervised machine learning algorithms, namely Random Forest, Boosting, Support Vector Machines, and Neural Networks, were meticulously trained. These algorithms were employed to predict relative poverty levels by leveraging the SILC survey responses, which were appropriately linked with administrative and geographic data, while concurrently quantifying the associated uncertainty in a rigorous and systematic manner. Consequently, through the adoption of this approach, the process of imputation through algorithmic modeling was rendered feasible at the population level, thus enabling the provision of poverty statistics with robust uncertainty estimates at an unprecedented level of spatial resolution. However, it is crucial to acknowledge that such a commendable achievement inevitably amplifies the risk of potential disclosure, necessitating the comprehensive quantification and rigorous implementation of control measures to mitigate such risks effectively.

2 Disclosure Control using Differential Privacy

Differential privacy represents a concept of utmost significance within the realm of privacy preservation, offering a well-defined and mathematically rigorous framework for quantifying and constraining the extent of privacy loss. Primarily developed in the context of statistical disclosure control, differential privacy endeavors to furnish accurate statistical information pertaining to a group of respondents while simultaneously safeguarding the privacy of each individual within the group. Informally, it tackles the inherent paradox of acquiring valuable insights into the characteristics of a population without disclosing any specific details about an individual.

The fundamental principle underlying differential privacy is predicated upon the notion that if the influence of a singular substitution within the database remains sufficiently negligible, then the publicly released statistic cannot be exploited to glean substantial information about any individual within the dataset, thus ensuring privacy. Achieving control over the impact of an individual's contribution is accomplished through the introduction of noise during the release of computed statistics. The magnitude of the injected noise directly influences the resulting statistics: a larger amount of noise affords stronger guarantees of privacy at the expense of reduced reliability in the statistics obtained. This intricate interplay between privacy and utility is commonly referred to as the privacy-utility trade-off.

Within the framework of differential privacy, the magnitude of the noise applied is denoted by the budget parameter ε. A higher budget corresponds to more precise and accurate statistical outcomes, albeit at the expense of weaker privacy guarantees. Therefore, it is of utmost importance, particularly for national statistical offices, to optimize the utilization of the budget and provide the strongest privacy guarantees while maintaining the desired level of utility.

The application of differential privacy to geo-referenced data has witnessed significant advancements, with the United States Census Bureau emerging as a trailblazer in this domain. Notably, the "OnTheMap" initiative aimed to disseminate commuting patterns of the U.S. population while preserving the privacy of sensitive data [4]. Subsequently, during the 2020 Census, the Census Bureau leveraged the inherent data structure offered by geo-referenced data to optimize the allocation of the differential privacy budget [5]. In our current project, we endeavor to replicate a comparable strategy to disclose the Swiss poverty statistics at an unprecedented level of geographical resolution: Drawing inspiration from the approach utilized in [6] for the disclosure of a synthetic dataset pertaining to U.S. broadband coverage, we put forth a novel strategy aimed at optimizing budget allocation. This strategy entails dividing the national territory into progressively smaller and mutually exclusive regions, thereby affording a natural means of partitioning the dataset. As a result of this partitioning, the budget is judiciously allocated among these distinct regions, as opposed to assigning it to each individual statistic separately. However, as highlighted by [7], the practical implementation of differential privacy poses considerable challenges, necessitating the use of mature, robust, and secure software solutions.

3 Implementation Challenges

OpenDP [8] is an ongoing collaborative initiative hosted at Harvard University, which aims to develop a reliable suite of open-source tools for facilitating privacy-preserving analysis of sensitive personal data. The primary focus lies in the creation of a comprehensive library of algorithms designed to generate differentially private statistical releases. The intended applications of OpenDP encompass governmental, industrial, and academic sectors, wherein secure and confident sharing of sensitive data is paramount to support scientifically oriented research and exploration in the public interest.

The OpenDP project offers an open-source library featuring efficient implementations of differential privacy algorithms utilizing the Rust programming language. Contributions to the library undergo rigorous scrutiny and undergo an extensive vetting process to ensure their correctness and reliability. Currently, the library provides ready-to-use mechanisms for injecting noise, with bindings available in Python and partially in R. However, when dealing with geo-referenced data, complex manipulation and partitioning of datasets are necessary to optimize the utilization of the differential privacy budget. Regrettably, such operations could not yet be performed within the confines of the OpenDP framework.

To address this limitation, we integrated the capabilities of the Polars library [9] into the OpenDP framework. Developed in Rust, Polars offers exceptional speed and efficiency for data-frame representation and manipulation, and it is increasingly regarded as the successor to Pandas, a widely used library among data scientists analyzing data with Python. This integration enables end-to-end data pipelines within the OpenDP framework, encompassing crucial tasks such as partitioning, thereby facilitating secure and private data release while maximizing the utility of the differential privacy budget. In the near future, thanks to Polars integration, it will be feasible to conduct all necessary data manipulation and transformation directly within OpenDP, starting from the processing of raw data and culminating in the generation of differentially private releases of computed statistics. This contribution represents a significant stride towards the realization of fully secure and production-ready data analysis pipelines, offering a solution to some of the challenges outlined in [7].

4 Limitations of DP for geo-referenced Data

Differential privacy, as a privacy-enhancing concept, offers the assurance that an attacker's ability to identify individual contributions is inherently limited, regardless of the information at their disposal. However, when dealing with geo-referenced data, it is not uncommon to encounter potentially sizable groups that share the same geographical location and possess identical sensitive attributes. For instance, within tall buildings, numerous individuals may fall below the poverty line, causing the observed poverty rate at that specific location to deviate significantly from the national estimate and approach a value close to 1. Likewise, in the context of mobility data derived from live mobile phone networks, the identification of regions or buildings with an unusually low number of active users, juxtaposed with the presence of multiple households indicating primary residency, can divulge the absence of residents.

In both examples, the privacy risk is directly linked to the geographical localization or absence thereof, of a group of individuals, whose size cannot be predetermined. Traditional differential privacy mechanisms could offer some level of disclosure control by increasing the magnitude of noise proportionally to the maximum group size that needs protection. However, this solution remains partial in its effectiveness and incurs a notable loss in usability.

To preclude any potential discrimination or harm stemming from geographical information, it is imperative to implement confidentiality protection measures that not only curtail the identification of individual contributions but also impede an attacker from pinpointing the exact geographical coordinates associated with potentially sensitive attributes. Currently, our research endeavors revolve around the development of such mechanisms, with the overarching goal of generating accurate and finely detailed maps of geo-referenced data. These maps will serve as invaluable resources for meticulous policy piloting, while upholding stringent confidentiality safeguards.

Acknowledgement

Michael Shoemate, Wanrgon Zhang and Salil Vadhan were supported in part by Cooperative Agreement CB20ADR0160001 with the U.S. Census Bureau and a grant from the Sloan Foundation. Salil Vadhan was also supported in part by a Simons Investigator Award. The views expressed in this paper are those of the authors and not those of the U.S. Census Bureau, the Swiss Federal Statistical Office or any other sponsor.

[1] EU statistics on income and living conditions - Microdata - Eurostat (europa.eu)

[2] Swiss Federal Statistical Office - Poverty and deprivation

[3] Poverty Measurement in Switzerland

[4] Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., & Vilhuber, L. (2008). Privacy: Theory meets Practice on the Map. *2008 IEEE 24th International Conference on Data Engineering, Cancun, Mexico*, 277–286. https://doi.org/10.1109/ICDE.2008.4497436

[5] Haney, S., Sexton, W., Machanavajjhala, A., Hay, M., & Miklau, G. (2021). Differentially Private Algorithms for 2020 Census Detailed DHC Race & Ethnicity. *ArXiv:2107.10659*, 1–18. http://arxiv.org/abs/2107.10659

[6] Pereira, M., Kim, A., Allen, J., White, K., Ferres, J. L., & Dodhia, R. (2021). U.S. Broadband Coverage Data Set: A Differentially Private Data Release. *ArXiv:2103.14035v2*, 1–7. http://arxiv.org/abs/2103.14035

[7] Garfinkel, S. L., Abowd, J. M., & Powazek, S. (2018). Issues encountered deploying differential privacy. *Proceedings of the ACM Conference on Computer and Communications Security*, 133–137. https://doi.org/10.1145/3267323.3268949

[8] The OpenDP team. (2020). *The OpenDP White Paper*. https://projects.iq.harvard.edu/files/opendp/files/opendp_white_paper_11may2020.pdf

[9] Polars: Lightning-fast DataFrame library for Rust and Python