The Usability of Perturbed Data Created based on Differential Privacy for Japanese Population Census – a Comparative Study

Shinsuke Ito[1,4], Masayuki Terada[2], Shunsuke Kato[3]
[1] Chuo University, Tokyo, JAPAN
[2] NTT DOCOMO, INC./Kyoto Tachibana University, Tokyo, JAPAN
[3] National Statistics Center/Ministry of Internal Affairs and Communications, Statistics Research and Training Institute, Tokyo, JAPAN
[4] Corresponding author: Shinsuke Ito, e-mail: ssitoh@tamacc.chuo-u.ac.jp

Abstract

Recent international trends in privacy-protecting techniques for official statistics include the use of perturbative methods. The U.S. Census Bureau is implementing perturbative methods based on the methodology of differential privacy, which was originally developed in the field of computer science in order to prevent "database reconstruction attacks", where attackers attempt to identify personal information by combining multiple published statistical tables. The U.S. Census Bureau has created and published statistical tables that use differential privacy for the 2020 United States Census.

Exploring the applicability of differential privacy techniques to Japanese official statistics is worthwhile both from the viewpoint of expanding the future scope of creation and publication of official statistical tables, and shaping the future direction of secondary use of official statistics. Several empirical studies have examined the potential of differential privacy as an anonymization method for detailed geographical data from the Japanese Population Census.

This paper investigates the possibility of adapting differential privacy to cross-tabulated data created using individual data from the 2020 Japanese Population Census, and conducts a comparative analysis of data usability at different geographical levels for perturbed statistical tables created based on differential privacy.

Keywords: differential privacy, U.S. Census Bureau, Japanese Population Census, individual data, data usability, security

1. Introduction

Many national statistical agencies have actively adopted perturbative methods as privacy protection techniques when publishing official statistics. The U. S. Census Bureau has considered differential privacy as a countermeasure to database reconstruction attacks (Abowd (2018)). In database reconstruction attacks, attackers expose personal information by merging (seemingly privacy-protected) datasets generated from a database, setting up constraint satisfaction problems. Differential privacy provides the means to reduce the risk that the original database can be reconstructed.

Prior to applying differential privacy to 2020 census data in preparation for publishing statistical tables, the U.S. Census Bureau used 2010 census data to investigate the practicality of differential privacy. Specifically, based on the top-down algorithm it adopted, the Bureau set a privacy loss budget ε to be consumed by publishing statistical tables, and examined how to appropriately allocate the parameter

ε at the geographical level (Garfinkel et al. (2019))[1].

In examining the applicability of differential privacy for Japanese official statistics, Ito et al. (2023) used individual data from the 2015 Japanese population census to quantitatively evaluate data utility after various differential privacy methods are applied to produce statistical tables at different geographical levels. Data utility was compared using mean absolute error (MAE) as the evaluation metric. The results showed that when differential privacy is applied to population census data, generating noise at the highest geographical level and then allocating it to the cells of statistical tables in a top-down manner with proper adjustments produces more accurate figures than a bottom-up manner which adds noise to the cells of aggregate data tables for the lowest geographical level and then aggregates the resulting figures to produce tables for a higher geographical level.

However, Ito et al. (2023) only focused on the granularity of geographical levels as the basis of comparison, and further empirical research for investigating data utility also of higher-dimensional, differentially private statistical tables is needed.

This study uses individual-level data from the 2020 Japanese Population Census to suggest a method for quantitatively evaluating the utility of differentially private aggregate data tables. It also evaluates the effect of adding variables and performing apportionment based on the distributional characteristics of variables on the utility of differentially private aggregate data tables.

2. Assessing Utility for Differential Private Census Data

There are several methods to evaluate utility for data created using disclosure limitation methods. Examples include MAE and RMSE, which are calculated as indicators to assess the extent of differences in distribution characteristics between aggregated data created from noise-added data and original individual data. When the methodology of differential privacy is applied, these indicators can be defined as the difference between the values based on original data and those with added noise generated by differential privacy. At the same time, utility metrics can also be used to evaluate values at different levels of granularity of geographical classifications, including the correlation between variables based on individual data and metrics calculated from aggregated data.

When differential privacy methods are applied, data is created at specific geographical levels. Therefore, in order to ensure the most similar data characteristics for the aggregated tables created from the original individual data, the appropriate geographical level should be selected among the various geographical classifications. This suggests that attention should also be paid to the difference between the distribution characteristics of the aggregated data and those of the original individual data from the standpoint of both noise addition based on differential privacy and granularity of geographical classifications.

For each of the geographical categories of different granularity, the correlation between variables based on individual data can be compared with that based on aggregated data. Robinson (1950) conceptualized the relationship between the distributional characteristics of aggregated data and those of the original individual data as the “ecological fallacy”. The ecological fallacy arises from inferring the relationship

---

[1] The top-down algorithm implemented by the Bureau produces statistical tables through the following process which includes the application of differential privacy. First, national-level aggregation is performed, noise is injected based on a mathematically optimized privacy loss budget (ε), and differentially private statistical tables are created. Next, state-level, noise-added differentially private statistical tables are created with consideration given to both the strength of privacy and data utility. Similarly, differentially private statistical tables are created for hierarchical geographical categories, namely at the county level, tract level, and block level, in that order (Ito and Terada (2020)).

between individual-level socioeconomic attributes based on an ecological correlation between area-level characteristics (Robinson 1950).

In evaluating the utility of differentially private aggregate data tables, taking into account geographical granularity, two types of errors must be considered: errors attributable to differential privacy which are errors between aggregate data with noise added for differential privacy and aggregate data created based on the original data and errors causing the ecological fallacy, i.e., differences between the distributional characteristics of the original data and the distributional characteristics of aggregate data for different geographical categories.

Researchers obtain data with the most useful distributional characteristics by finding the combination of variables and geographical granularity (used in the aggregate data tables) which minimizes the sum of errors attributable to differential privacy and errors causing ecological fallacy.

As the geographical granularity becomes finer, the distribution characteristics of the aggregate data table tend to become more similar to those of individual-level data, and ecological-fallacy errors are therefore smaller. However, because the frequencies of the cells in the aggregate data table are lower, if noise is added to each cell for differential privacy, the impact of the noise addition on the frequencies is relatively large.

It is necessary to formalize the sum of errors attributable to differential privacy and ecological-fallacy errors. Furthermore, a utility indicator is needed to quantitatively evaluate the sum of these two types of errors.

3. Proof-of-Concept Experiment on the Application of Differential Privacy to 2020 Census Data

This experiment examines the relationship between geographical granularity and the variables used in cross-tabulated tables. In the experiment, individual data from the 2020 Japanese Population Census is used. Various cross-tabulated tables for three variables including gender, age, and type of residence as well as tables aggregated by small regions with different levels of geographical granularity are created.

The experiment also investigates how noise added based on the methodology of differential privacy affects the utility of the cross-tabulated tables. To achieve this, various differential privacy methods are applied to these tables. The experiment also examines the effects of adding new aggregation items or performing proration based on the distribution characteristics of survey items on the aggregated tables. For this, differentially private full cross-tabulated tables and prorated cross-tabulated tables containing gender (two categories), age (18 categories), and residential type (three categories) are created.

We use the following methods for implementing differential privacy: (a) Laplace mechanism (with negative value rounding), (b) bottom-up composition method, and (c) top-down composition method. Additionally, we use eight values for the privacy loss budget (ε): 0.1, 0.2, 0.7, 1, 1.1, 5, 10, and 20. To ensure that the geographical divisions are structured hierarchically, we set the following hierarchy: (A) prefectures, (B) municipalities, (C) towns/villages, and (D) basic units. For each geographical division, we calculate MAE and RMSE as indicators of the utility of the statistical values to which differential privacy is applied.

We apply the top-down method as follows. First, noise is added, and optimization is applied to the population of each prefecture, which is the top-level geographical classification, with the total population of Japan as the total constraint. Next, noise is added, and optimization is applied to the population of each municipality, using the refined population of each prefecture obtained in the previous step as the total constraint. The same process is repeated for the population of each town/village and basic unit.

The procedure of experimental application of differential privacy to the 2020 Census data is as follows:

(1) A cross-tabulated table using all variables for each geographical level (hereinafter referred to as the "full cross-tabulated table") is created.
(2) All possible cross-tabulated tables using the same combination of variables as full cross-tabulated table are prepared and prorated based on the method of applying distribution from higher-level geographical classifications (hereinafter referred to as the “prorated cross-tabulated table").
(3) For each of (1) and (2), differential privacy methods are applied to cross-tabulated tables created for all types of geographical areas while varying the value of ε.
(4) Cells in full cross-tabulated tables and those in the prorated cross-tabulated tables from the standpoint of effectiveness of differential privacy and ecological fallacy are compared.

4. Experimental Results

Tables 1 and Tables 2 show the comparison between MAE calculated for full cross-tabulated tables created using three-variable (gender, age, type of residence) and MAE calculated for prorated cross-tabulated tables for the prefecture, municipality, town/village, and basic units. In each table, “Laplace,” “BottomUp,” and “TopDown” refers to the Laplace mechanism (plus negative-value truncation), and bottom-up composition method, and top-down composition method. The values in bold indicate the differential privacy method with the smallest MAE under the given conditions. Also, a cell highlighted in light blue in a full cross-tabulated tables indicates that the MAE in the cell is smaller than the corresponding MAE in the relevant prorated cross-tabulated table.

Creating cross-tabulated tables using all variables tends to increase noise relatively, but among these, the MAE for the top-down approach was found to be generally smaller than the MAE for other methods. It was confirmed that as the geographical classification becomes larger in the order of basic unit district, town/village, municipality, prefecture, and nationwide, the relative noise assigned tends to increase. In the top-down approach, creating cross-tables using all variables generally demonstrated higher utility, whereas in the bottom-up approach and Laplace mechanism (with negative values rounded up), cross-tabulated tables created by allocating from higher-level categories generally showed better MAE results.

Focusing on the MAE at the basic unit level, it is interesting to note that, with some exceptions, when $\varepsilon \leq 1.1$, the results obtained using the prorated method generally have a relatively smaller MAE than those obtained using any of the other methods. Conversely, when $\varepsilon \geq 5.0$, regardless of differential privacy methods, the cross-tabulated tables created using all the target variables show a relatively smaller MAE and higher utility compared to the prorated cross-tabulated tables.

5. Conclusions

This paper explores the applicability of differential privacy methodologies to Japanese Census data and offers a preliminary analysis aimed at further investigating methods for evaluating their utility. Using aggregated tables created from individual-level data from the 2020 Japanese Population Census, we quantitatively assessed the utility of various differential privacy implementation methods.

When prorating was performed based on the distribution characteristics of survey items in higher-level geographical categories, it was empirically confirmed that the MAE of the aggregated tables created using certain differential privacy methods can be smaller than for full cross-tabulated tables created using all the target variables when $\varepsilon \leq 1.1$. On the other hand, the full cross-tabulated tables created using all variables show a relatively smaller MAE than the prorated cross-tabulated tables for any of differential privacy method used in this study when $\varepsilon \geq 5.0$.

For data to which differential privacy methods were applied, the results can

potentially be affected by not only the noise introduced by the application of differential privacy but also the discrepancy between the distribution characteristics of the aggregated data subject to noise and the original individual data. Further consideration of evaluation methods from the perspective of ecological fallacy is required.

References

Abowd, J. M. (2018). Staring-down the database reconstruction theorem, Joint Statistical Meetings, Vancouver, BC, Canada.

Garfinkel, S. Abowd, J. M., and Martindale, C. (2019) “Understanding Database Reconstruction Attack in Public Data”, *Communications of the ACM*, Vol. 62 No. 3, ACM, pp. 46-53.

Ito, S. and Terada, M. (2020) “An Evaluation of Anonymization Methods for Creating Detailed Geographical Data” (in Japanese), *Journal of the Japan Statistical Society*, Vol. 50, No. 1, pp.139-166.

Ito, S., Terada, M., Kato, S. (2023) “The Potential of Differential Privacy Applied to Detailed Statistical Tables Created Using Microdata from the Japanese Population Census”, Paper presented at UNECE Expert Meeting on Statistical Data Confidentiality 2023, pp.1-10.

Robinson, W. S. (1950) “Ecological Correlations and the Behavior of Individuals", *American Sociological Review*, vol.15, pp.351-357.

Table 1 Experimental Results: MAE for Full Cross-tabulated Tables by Gender, Age, and Type of Residence

ε	Method	Nationwide	Prefecture	Municipality	Town/Village	Basic Unit District
0.1	(a)Laplace	9368343.77	199326.46	4938.51	91.30	5.16
	(b)BottomUp	11415.90	2706.27	113.71	**4.90**	**0.66**
	(c)TopDown	**52.50**	**36.58**	**19.80**	5.79	0.77
0.2	(a)Laplace	4580701.41	97461.73	2414.72	44.77	2.63
	(b)BottomUp	6553.29	1865.22	79.71	**3.41**	**0.51**
	(c)TopDown	**24.88**	**18.30**	**11.55**	3.80	0.72
0.7	(a)Laplace	1239727.50	26377.18	653.57	12.27	0.79
	(b)BottomUp	2043.03	657.09	29.25	**1.38**	**0.23**
	(c)TopDown	**7.42**	**5.51**	**4.09**	1.56	0.54
1	(a)Laplace	855047.25	18192.49	450.80	8.50	0.56
	(b)BottomUp	1448.49	479.82	21.44	**1.05**	**0.18**
	(c)TopDown	**4.07**	**3.79**	**2.99**	1.19	0.47
1.1	(a)Laplace	774578.91	16480.40	408.38	7.71	0.51
	(b)BottomUp	1260.85	437.03	19.67	**0.97**	**0.16**
	(c)TopDown	**3.93**	**3.41**	**2.74**	1.10	0.45
5	(a)Laplace	165950.78	3530.87	87.51	1.67	0.11
	(b)BottomUp	305.43	109.69	4.99	**0.27**	**0.04**
	(c)TopDown	**1.05**	**0.79**	**0.67**	0.33	0.17
10	(a)Laplace	82949.81	1764.89	43.74	0.84	0.06
	(b)BottomUp	150.59	55.30	2.51	**0.13**	**0.02**
	(c)TopDown	**0.48**	**0.39**	**0.34**	0.18	0.10
20	(a)Laplace	41467.23	882.28	21.87	0.42	0.03
	(b)BottomUp	68.13	27.51	1.25	**0.07**	**0.01**
	(c)TopDown	**0.23**	**0.20**	**0.17**	0.09	0.05

Table 2 Experimental Results: MAE for Prorated Cross-tabulated Tables by Gender, Age, and Type of Residence

ε	Method	Nationwide	Prefecture	Municipality	Town/Village	Basic Unit District
0.1	(a)Laplace	910658.64	19375.72	484.49	11.14	0.78
	(b)BottomUp	5915.24	1127.86	51.94	**3.50**	**0.36**
	(c)TopDown	**64.12**	**59.14**	**28.07**	7.39	0.58
0.2	(a)Laplace	437610.58	9311.27	234.33	5.75	0.51
	(b)BottomUp	2942.42	583.23	28.65	**2.10**	**0.30**
	(c)TopDown	**34.87**	**29.77**	**16.70**	5.10	0.46
0.7	(a)Laplace	115895.64	2466.44	63.18	1.73	0.32
	(b)BottomUp	999.59	208.31	10.71	**0.81**	**0.26**
	(c)TopDown	**10.76**	**8.53**	**6.09**	2.20	0.31
1	(a)Laplace	79628.75	1694.99	43.66	1.23	0.30
	(b)BottomUp	558.41	157.58	8.17	**0.61**	**0.26**
	(c)TopDown	**6.93**	**6.07**	**4.48**	1.69	0.29
1.1	(a)Laplace	71933.97	1531.35	39.51	1.12	0.29
	(b)BottomUp	635.96	145.43	7.59	**0.57**	**0.26**
	(c)TopDown	**6.23**	**5.43**	**4.17**	1.58	0.29
5	(a)Laplace	14897.50	317.46	8.32	0.25	0.26
	(b)BottomUp	133.98	38.98	2.09	**0.15**	**0.25**
	(c)TopDown	**1.50**	**1.25**	**1.05**	0.48	0.26
10	(a)Laplace	7415.91	158.04	4.15	0.13	0.25
	(b)BottomUp	59.51	19.66	1.07	**0.08**	**0.25**
	(c)TopDown	**0.78**	**0.62**	**0.54**	0.27	0.25
20	(a)Laplace	3705.32	78.97	2.07	0.06	0.25
	(b)BottomUp	28.72	9.93	0.54	**0.04**	**0.25**
	(c)TopDown	**0.35**	**0.32**	**0.28**	0.15	0.25