# The Potential of Differential Privacy Applied to Detailed Statistical Tables Created Using Microdata from the Japanese Population Census

*Shinsuke Ito, Chuo University, Japan*

*Masayuki Terada, NTT DOCOMO, Inc*

*Shunsuke Kato, Statistics Bureau of Japan*

# 1. Introduction

・Recent international trends in privacy-protecting techniques applied to official statistics include the **active use of perturbative methods**: The U.S. Census Bureau has investigated the applicability of perturbative methods based on the methodology of differential privacy.

・Several empirical studies on the effectiveness of perturbative methods for Japanese official microdata have been conducted in Japan (ex. Ito et al. (2018)).

Other studies have investigated the possibility of adapting differential privacy for detailed geographical data from the Japanese Population Census (Ito and Terada (2019) and Ito et al. (2020)).

# 1. Introduction

・Investigating the applicability of differential privacy techniques to Japanese official statistics is worthwhile from the standpoint of discussing the future creation and publication of official statistical tables and the future direction of secondary use of official statistics.

・This paper explores the applicability of differential privacy to data from the **Japanese Population Census** not only by empirically demonstrating the characteristics of data obtained under various perturbative methods, but also by examining the utility of these methods for Japanese official statistical data.

# 2. Application of Differential Privacy to Census Statistics

・Differential privacy is a **privacy protection framework aimed at achieving comprehensive (ad omnia) data security against arbitrary attacks including unknown attacks**.

・Differential privacy does not refer to a certain privacy protection method, but rather is a **framework for defining the level of data security** provided.

・ The **Laplace mechanism** is a typical mechanism for achieving differential privacy. Some traditional statistical disclosure control methods such as **PRAM (Post RAndomization Methods)** are also known to provide data security based on differential privacy.

# 2. Application of Differential Privacy to Census Statistics

・A random number from the Laplace distribution centered on zero (**Laplace noise**) can be added independently to each cell (even if its value is 0) in the contingency table.

・Not all mechanisms are suitable for statistical tables from the Japanese Population Census. Even if some mechanisms are suitable, using the wrong mechanism would significantly reduce the utility of the output statistics.

・Simply applying the Laplace mechanism to a large-scale contingency table, such as those from the Population Census, causes **the issues listed below** and reduces the utility of statistics (Terada et al. (2015), Ito and Terada (2020)).

**(1) Deviation from the nonnegative constraint**
**(2) Loss of sparseness**
**(3) Loss of accuracy in partial sums**

# 2. Application of Differential Privacy to Census Statistics

・A method based on wavelet transform has been shown to be useful for mesh population statistics. The **Privelet method** (Xiao et al. (2011)) introduces the **Haar wavelet transform** in the process of noise injection so that the noises of neighboring cells offset one another, and thereby increases the accuracy of the partial sum for a continuous domain.

・Terada et al. (2015) propose a method based on the **Morton order mapping and the Wavelet transform** with nonnegative refinement (hereinafter, “nonnegative wavelet method”). In this method, noise is injected over the wavelet space as in the Privelet method. By applying an inverse wavelet transform while correcting coefficient values to prevent the output from deviating from the nonnegative constraint, the nonnegative wavelet method produces population data that satisfies the nonnegative constraint and guarantees differential privacy.

・An empirical experiment (Ito and Terada (2020)), in which the nonnegative wavelet method is applied to mesh statistics from the 2010 Population Census, shows that the **nonnegative wavelet method solves the three problems** discussed above.

# 2. Application of Differential Privacy to Census Statistics

・In order to solve the above problems such as deviation from the **nonnegative constraint, loss of sparseness and loss of accuracy in partial sums**, applying the **constrained optimization method** which searches for nearest neighborhood vectors, satisfying with a total-number constraint and nonnegative constraint, to population statistics other than mesh statistics is useful for achieving differential privacy.

・There are two types of approach, **bottom-up data construction approach** and **top-down construction approach** as the methods for the application of constrained optimization.

# 2. Application of Differential Privacy to Census Statistics

(1) **Bottom-up data construction approach**: the method is applied to only the population of the smallest geographical unit (the basic unit district, in the case of the Japanese Population Census); and the resulting district-level population data are summed to obtain the population at the municipal or prefectural level in a bottom-up manner.

(2) **Top-down data construction approach**: the methods is applied to the prefecture-level population data with the total national population setting the total-number constraint in order to obtain privacy-protected prefecture-level population data and the method is applied recursively in order of municipality-level, town/village-level, and basic unit district-level, based on the same approach used for US 2020 Census data.

# 3. Applying Differential Privacy to the 2015 Japanese Population Census Data

## 3.1 Data Used in the Experiment

・The experiment is based on three small area statistics (Aggregation Tables 1 to Aggregation Tables 3) with different aggregation categories that were created from individual data from the 2015 Population Census.

・The experiment is based on three aggregate data tables for

**(1) basic unit district-level total population (Aggregation Table 1)**

**(2) basic unit district-level population by gender (Aggregation Table 2)**

**(3) basic unit district-level population by gender and 5-year age group (Aggregation Table 3).**

・ For each aggregate data table, the focus is on population size; the level of aggregation is at **the basic unit district (minimum geographic district) level**; and the three aggregation categories considered are **“all”, “males and females”, and “males and females** in **5-year age groups**”.

## 3. Applying Differential Privacy to the 2015 Japanese Population Census Data

### 3.2 Experimental Methods

・This empirical study aims to gain knowledge by implementing various differential privacy methods for the Aggregation Tables 1 to Aggregation Table 3, and evaluates the utility of the data.

・**(1) PRAM**, **(2) Laplace mechanism**, **(3) top-down data construction method**, and **(4) bottom-up data construction method** were used as methods for achieving differential privacy.

・Output of the Laplace mechanism can include negative population values. **These were zeroed out as a post-processing adjustment**. The constraint optimization method used with the top-down data construction approach or the bottom-up data construction approach was the one proposed by Terada et al. (2017).

## 3. Applying Differential Privacy to the 2015 Japanese Population Census Data

・Four methods were applied with each of the following eight values for the privacy loss budget ($\varepsilon$) set for the experiment: **0.1, 0.2, 0.7, 1.0, 1.1, 5, 10, and 20**(The values 0.7 and 1.1 were chosen as approximations of $log_e 2$ and $log_e 3$, respectively, which are frequently used as values of the privacy loss budget).

・Though PRAM and the top-down data construction approach can be configured to allocate different privacy loss budgets to different geographical levels or attribute categories, in this experiment, each value of the privacy loss budget was evenly allocated.

・Utility of the data was evaluated for population data at the basic unit district level, and for population data at higher geographical levels (partial sums). Specifically, **the errors in the prefecture-level, municipality-level, town/village-level, and basic unit district-level population data were quantitatively compared**.

・**The mean absolute error (MAE)** was used as an error index in this study.

# Table 1: Evaluation Results for Aggregation Table:Basic Unit District-level Total Population

| ε | Method | Prefecture | Municiparity | Town/Village | Basic Unit District |
|---|---|---|---|---|---|
| 0.1 | (a)PRAM | 0.00 | 14408.44 | 520.10 | 48.65 |
| | (b)Laplace | 98607.22 | 2490.96 | 83.50 | 17.60 |
| | (c)BottomUp | 0.00 | 855.53 | 72.06 | 17.38 |
| | (d)TopDown | 0.00 | 79.09 | 73.35 | 49.83 |
| 0.2 | (a)PRAM | 0.00 | 14399.53 | 520.26 | 48.64 |
| | (b)Laplace | 30844.00 | 817.25 | 39.15 | 9.23 |
| | (c)BottomUp | 0.00 | 367.38 | 36.86 | 9.21 |
| | (d)TopDown | 0.00 | 41.12 | 37.78 | 30.42 |
| 0.7 | (a)PRAM | 0.00 | 14401.57 | 520.09 | 48.65 |
| | (b)Laplace | 5433.01 | 157.62 | 11.06 | 2.75 |
| | (c)BottomUp | 0.00 | 97.37 | 10.81 | 2.75 |
| | (d)TopDown | 0.00 | 11.62 | 11.26 | 10.42 |
| 1 | (a)PRAM | 0.00 | 14406.80 | 520.17 | 48.65 |
| | (b)Laplace | 3483.00 | 104.64 | 7.65 | 1.92 |
| | (c)BottomUp | 0.00 | 65.78 | 7.52 | 1.91 |
| | (d)TopDown | 0.00 | 7.84 | 7.83 | 7.38 |
| 5 | (a)PRAM | 0.00 | 14351.58 | 518.16 | 48.47 |
| | (b)Laplace | 609.34 | 19.08 | 1.54 | 0.39 |
| | (c)BottomUp | 0.00 | 13.10 | 1.51 | 0.38 |
| | (d)TopDown | 0.00 | 1.60 | 1.57 | 1.52 |
| 10 | (a)PRAM | 0.00 | 10215.78 | 359.65 | 34.59 |
| | (b)Laplace | 314.63 | 9.65 | 0.77 | 0.19 |
| | (c)BottomUp | 0.00 | 6.43 | 0.76 | 0.19 |
| | (d)TopDown | 0.00 | 0.84 | 0.79 | 0.76 |
| 20 | (a)PRAM | 0.00 | 3.53 | 0.23 | 0.02 |
| | (b)Laplace | 152.12 | 4.80 | 0.38 | 0.10 |
| | (c)BottomUp | 0.00 | 3.15 | 0.38 | 0.10 |
| | (d)TopDown | 0.00 | 0.42 | 0.39 | 0.38 |

The errors for the prefecture-level population data are zero for the three methods other than the Laplace mechanism. PRAM, the bottom-up data construction method, and the top-down data construction method have the characteristic that the total number of records in the input data is preserved in the output data.

# Table 2: Evaluation Results for Aggregation Table 2: Basic Unit District-level Total Population by Gender

| ε | Method | Prefecture | Municiparity | Town/Village | Basic Unit District |
|---|---|---|---|---|---|
| 0.1 | (a)PRAM | 35301.70 | 7277.54 | 261.95 | 24.80 |
| | (b)Laplace | 158482.08 | 3940.62 | 96.57 | 16.10 |
| | (c)BottomUp | 4800.07 | 977.40 | 68.02 | 15.55 |
| | (d)TopDown | 60.08 | 79.80 | 69.48 | 36.40 |
| 0.2 | (a)PRAM | 34432.49 | 7273.24 | 261.97 | 24.81 |
| | (b)Laplace | 50430.20 | 1271.48 | 41.92 | 8.77 |
| | (c)BottomUp | 2081.00 | 443.80 | 36.11 | 8.68 |
| | (d)TopDown | 24.09 | 39.48 | 36.56 | 24.83 |
| 0.7 | (a)PRAM | 29972.11 | 7260.04 | 261.72 | 24.79 |
| | (b)Laplace | 7016.17 | 193.09 | 11.17 | 2.72 |
| | (c)BottomUp | 432.39 | 100.55 | 10.83 | 2.71 |
| | (d)TopDown | 9.27 | 11.83 | 11.16 | 9.67 |
| 1 | (a)PRAM | 27463.17 | 7255.24 | 261.68 | 24.80 |
| | (b)Laplace | 4239.02 | 120.71 | 7.69 | 1.90 |
| | (c)BottomUp | 310.38 | 68.29 | 7.50 | 1.89 |
| | (d)TopDown | 7.14 | 8.10 | 7.77 | 7.00 |
| 5 | (a)PRAM | 5492.60 | 7215.53 | 261.27 | 24.78 |
| | (b)Laplace | 653.28 | 19.92 | 1.54 | 0.38 |
| | (c)BottomUp | 59.69 | 12.58 | 1.51 | 0.38 |
| | (d)TopDown | 1.19 | 1.58 | 1.57 | 1.51 |
| 10 | (a)PRAM | 530.02 | 7197.73 | 260.39 | 24.70 |
| | (b)Laplace | 315.57 | 9.95 | 0.77 | 0.19 |
| | (c)BottomUp | 26.27 | 6.56 | 0.76 | 0.19 |
| | (d)TopDown | 0.54 | 0.82 | 0.78 | 0.76 |
| 20 | (a)PRAM | 7.51 | 5115.43 | 180.82 | 17.74 |
| | (b)Laplace | 159.90 | 4.92 | 0.39 | 0.10 |
| | (c)BottomUp | 14.20 | 3.23 | 0.38 | 0.10 |
| | (d)TopDown | 0.25 | 0.40 | 0.39 | 0.38 |

# Table 3: Evaluation Results for Aggregation Table 3: Basic Unit District-level Total Population by Gender and 5-year Age Group

| ε | Method | Prefecture | Municiparity | Town/Village | Basic Unit District |
|---|---|---|---|---|---|
| 0.1 | (a)PRAM | 15899.43 | 587.27 | 18.50 | 2.05 |
| | (b)Laplace | 376314.96 | 9323.57 | 173.38 | 10.77 |
| | (c)BottomUp | 14008.77 | 544.73 | 25.62 | 3.23 |
| | (d)TopDown | 81.64 | 76.50 | 30.75 | 3.44 |
| 0.2 | (a)PRAM | 15874.93 | 586.70 | 18.49 | 2.05 |
| | (b)Laplace | 175973.88 | 4359.93 | 81.69 | 5.69 |
| | (c)BottomUp | 11884.21 | 447.52 | 19.48 | 2.80 |
| | (d)TopDown | 41.25 | 39.09 | 20.72 | 3.28 |
| 0.7 | (a)PRAM | 15703.82 | 584.13 | 18.46 | 2.05 |
| | (b)Laplace | 40261.21 | 997.68 | 19.59 | 1.90 |
| | (c)BottomUp | 5618.71 | 203.17 | 8.85 | 1.55 |
| | (d)TopDown | 11.51 | 11.42 | 8.38 | 2.66 |
| 1 | (a)PRAM | 15573.72 | 582.26 | 18.44 | 2.05 |
| | (b)Laplace | 25349.47 | 628.32 | 12.71 | 1.38 |
| | (c)BottomUp | 4077.82 | 146.72 | 6.56 | 1.20 |
| | (d)TopDown | 7.94 | 7.93 | 6.20 | 2.37 |

The errors caused by PRAM to the basic unit district-level data vary very little across different values of the privacy loss budget. The MAE calculated for PRAM at $\varepsilon$=0.1, 0.2 are smaller than for the other three methods.

## 3. Applying Differential Privacy to the 2015 Japanese Population Census Data

### 3.3 Experimental Methods

・For tables 1 to 3, **(a) PRAM, (b) Laplace, (c) BottomUp, and (d) TopDown** refer to PRAM, **the Laplace mechanism (plus zeroing out of negative values)**, the bottom-up data construction method, and the top-down data construction method, respectively.

・The result that errors at the prefectural level in Table 1 are zero is **attributable to the conditions of this experiment**, and the same result would not be obtained if the highest geographical level were the national level (instead, errors for total population at the national level would become zero).

・In the case of PRAM, **the basic unit district-level aggregate data table used for Table** 3 is quite **sparse.** Therefore, at first glance its accuracy does not appear undesirable at the basic unit district level. However, this is a **false accuracy and not statistically meaningful**. In fact, the errors caused by PRAM at the municipality level and the town/village level shown in Table 3 reveal that **the accumulation of errors greatly degrades the accuracy of the partial sums** and there is significant **degradation** of the characteristics of the original aggregate data table.

# 4. Discussion

- For a given privacy loss budget, differential privacy guarantees the same level of privacy protection regardless of the differential privacy method used, but **the utility of the resulting data depends on the method and the use of the data**.

- If the **errors at the basic unit district level** are taken as indices of the utility of the relevant data, then the output data from the bottom-up data construction method and the Laplace mechanism seem superior.

- **For partial sums at the municipality level and town/village level, the errors tend to be larger for both the bottom-up method and the Laplace mechanism**, and the tendency is particularly noticeable with the Laplace mechanism.

- The top-down data construction method is inferior to the bottom-up data construction method in terms of the errors at the basic unit district level. **However, for the top-down method, the errors are not accumulated at higher geographical levels**.

# 4. Discussion

・Satisfying **the nonnegative constraint is a problem for a simple Laplace mechanism.** Even if an attempt is made to satisfy the constraint by zeroing out negative values as in this experiment, **it is still difficult to obtain a practically usable aggregate data table because of the large overestimation bias affecting partial sums**.

・**PRAM clearly fails to achieve both a reasonable level of privacy protection and data utility**. Under a given set of conditions, in most cases **PRAM is significantly inferior in terms of privacy protection efficiency**. For small values of the privacy loss budget, the results of applying PRAM at the basic unit district level seem to be superior to the results of other methods. **However, this is attributable to false accuracy**.

・Regarding the bottom-up data construction method and the top-down data construction method, the errors at the basic unit district level show that the **bottom-up method** provides higher data utility, but **its errors for partial sums increase as the range of cells used for the partial sums expands, which indicates decreasing data utility**.

# 4. Discussion

・**For the top-down method, the errors for partial sums remain small**. Therefore, when partial sums are calculated for higher geographical levels, **the degree of data utility is maintained**.

・Judging from the errors for the output data at the basic unit district level, **data utility is relatively high for the bottom-up method, but deteriorates for partial sums since the errors associated with them tend to increase significantly**.

**For the top-down method,** errors at the basic unit district level **are larger than for the bottom-up method**, and if the level of privacy protection is properly set, **the utility of different output data considered in this study is maintained**, even when the effect on partial sums is taken into account.

## 5. Conclusion and Outlook

(1) This paper evaluates the utility of statistical tables for different geographical levels which were created using individual data from the Population Census and by **applying various differential privacy methods**.

(2) The results of this study show that in applying differential privacy to Japanese Population Census data, **the top-down data construction method yields a higher level of data utility than the other methods**.

(3) This study also suggests that given a hierarchical geographical structure, **reasonable results from the standpoint of data utility can be obtained by top-down**, consistent allocation of the noise generated based on differential privacy to the cells of a statistical table.

(4) Our future research agenda also includes further investigation into the effectiveness of differential privacy based on aggregate data tables created with **various Population Census variables**.