The Potential of Anonymization Methods for Creating Detailed Geographical Data in Japan

Shinsuke Ito, Chuo University, Japan

Masayuki Terada, NTT DOCOMO, Inc

1. Introduction: Secondary Use of Official Statistics in Japan
2. Differential Privacy Initiatives for Official Statistics
3 Experiment on the Application of Differential Privacy to Small-Area Statistics
4. Conclusion and Outlook

1. Introduction: Secondary Use of Official Statistics in Japan

・In Japan, provision of confidential data, creation and provision of anonymized microdata, and tailor-made tabulation services is carried out under the Japanese Statistics Act.

・The use of official statistical data, administrative data and private “big data” to promote evidence-based policymaking (EBPM) is receiving increased attention in Japan.

・A revised Statistics Act was enacted on 1 June 2018 and came into force on 1 May 2019. One feature of the revised act is an expanded scope for provision of questionnaire information.

1. Introduction: Secondary Use of Official Statistics in Japan

・The revisions state that the “significant public benefit” in Articles 33 (Provision of Questionnaire Information) and 36 (Provision of Anonymized Data) should be of a level to which “the preparation of statistics, etc., does not compromise citizen confidence regarding official statistical surveys” and is “objectively reasonable and appropriate”.

・As a result, it is necessary to place usage purpose-based limitations on questionnaire information (original microdata) and anonymized microdata at the operational level.

2. Differential Privacy Initiatives for Official Statistics

・The Statistics Bureau provides 6 types of Anonymized microdata from Japanese official statistics including the Population Census.

Current Situation for Population Census Data:

- 2000 and 2005 census data is currently available.
- Geographic information is limited (prefecture level and geographical areas with 500,000 people and above)

Current Anonymization Methods for Population Census Data:

- Sampling based on household units (at a sampling rate of 1%)
- Non-perturbative methods incl. deletion of direct identifiers, recoding, top and bottom coding
- Deletion of unique records
- Data swapping

2. Differential Privacy Initiatives for Official Statistics

・In order to promote a broader use of anonymized Japanese official microdata, several empirical studies on the effectiveness of disclosure limitation methods for official microdata have been conducted by the National Statistics Center (Ito and Murata (2011), Ito and Hoshino (2012, 2013, 2014)) and by the Statistics Bureau of Japan and the National Statistics Center (Ito et al. (2015, 2016, 2017, 2018)).

・Empirical research on using top coding and recoding with the aim of creating Japanese anonymized microdata that contain more detailed geographical information was conducted by Ito et al. (2015, 2016).

・Empirical research on the potential of perturbative methods such as data swapping and PRAM (=Post RAndomization Methods) to reduce disclosure risk for official microdata was conducted by Ito et al. (2017, 2018).

2. Differential Privacy Initiatives for Official Statistics

・In various countries, differential privacy methods have been proposed as a method for controlling noise according to the standards of secrecy. In the area of computer science, differential privacy has been developed based on the concept of formal privacy.

・ Differential privacy (Dwork (2006)) is an indistinguishably-based definition of privacy where the output distribution of a random query to database D1 is almost identical to that to any neighbouring database, and thus can limit the disclosure of private information from an individual record contained in the database.

2. Differential Privacy Initiatives for Official Statistics

・Abowd (2018) described the possibility of database reconstruction attacks on statistical tables. A reconstruction attack exposes personal information contained in confidential data that is the source of a query by combining a small number of random queries, even without having to look carefully at the queries themselves (Dinir and Nissim, 2003).

・Abowd (2018) showed that when many geographic categories are used to create detailed statistical tables, combining these tables can increase the risk of identifying individuals even if the tables do not contain identifying information.

・The U.S. Census Bureau is performing verifications using 2010 census data, and investigating the potential of differential privacy as a way of maintaining data accuracy.

・To examine the potential of differential privacy, the U.S. Census Bureau is applying visualization of optimal values for privacy loss and data accuracy based on the concept of the “production probability frontier” from economics.

2. Differential Privacy Initiatives for Official Statistics

・Differential privacy has so far been rarely discussed in the context of Japanese official statistics, whereas in other countries, discussions of differential privacy in the field of official statistics have advanced.

・The process undertaken by the U.S. Census Bureau, which applies differential privacy to high-dimensional cross tables at sub-regional levels and tries to expand this into higher-level regional level aggregation result tables for publishing, may be an approach worth considering when it comes to exploring the potential adoption of differential privacy in Japan.

3. Experiment on the Application of Differential Privacy to Small-Area Statistics

・In order to explore the potential of differential privacy for official statistics in Japan, we applied differential privacy to meshed population data from the Japanese Population Census.

・The simplest method to apply differential privacy to the meshed population data would be to add random values following a double exponential distribution (a Laplace distribution) to all population values in the mesh (known as the “Laplace mechanism”) (Dwork et al. (2006)).

・The value of the privacy loss budget ε for differential privacy security is determined according to the scale of the Laplace noise to be added. Specifically, letting Lap(λ) be the Laplace noise at scale λ for a given mesh population value pi, the population values with added Laplace noise $pi' = pi + Lap(\lambda)$ satisfy differential privacy $\varepsilon = 1/\lambda$ also written (1/λ) differential privacy.

3. Experiment on the Application of Differential Privacy to Small-Area Statistics

Three issues in particular have emerged in past research on the subject.

(1) Violation of nonnegativity constraints
(2) Inflation of data amounts
(3) Degraded partial-sum accuracy

・Deviation from nonnegativity constraints could be simply addressed by replacing all cells containing negative values with 0. This would cause a positive bias to the overall dataset, further exacerbating the problem of partial-sum accuracy.

・A method based on nonnegative wavelet transformation with top-down refinement has been proposed as an attempt to solve the problems described above (Terada et al. (2015)). This method applies a Haar wavelet transform to the original population data, and Laplace noise is added to the resulting wavelet coefficients.

3. Experiment on the Application of Differential Privacy to Small-Area Statistics

・In order to handle two-dimensional data such as meshed population data, the method introduces a Morton order map (Morton (1966)) (a type of locality-preserving map) for transforming the data into one-dimensional data before applying Haar wavelet transform, rather than use multi-dimensional wavelets as in the case of Privelet.

・The characteristics of wavelet transformation control the accuracy of partial sums without violating nonnegativity constraints. Further, the process for correcting wavelet coefficients to avoid violating nonnegativity constraints restores data sparseness.

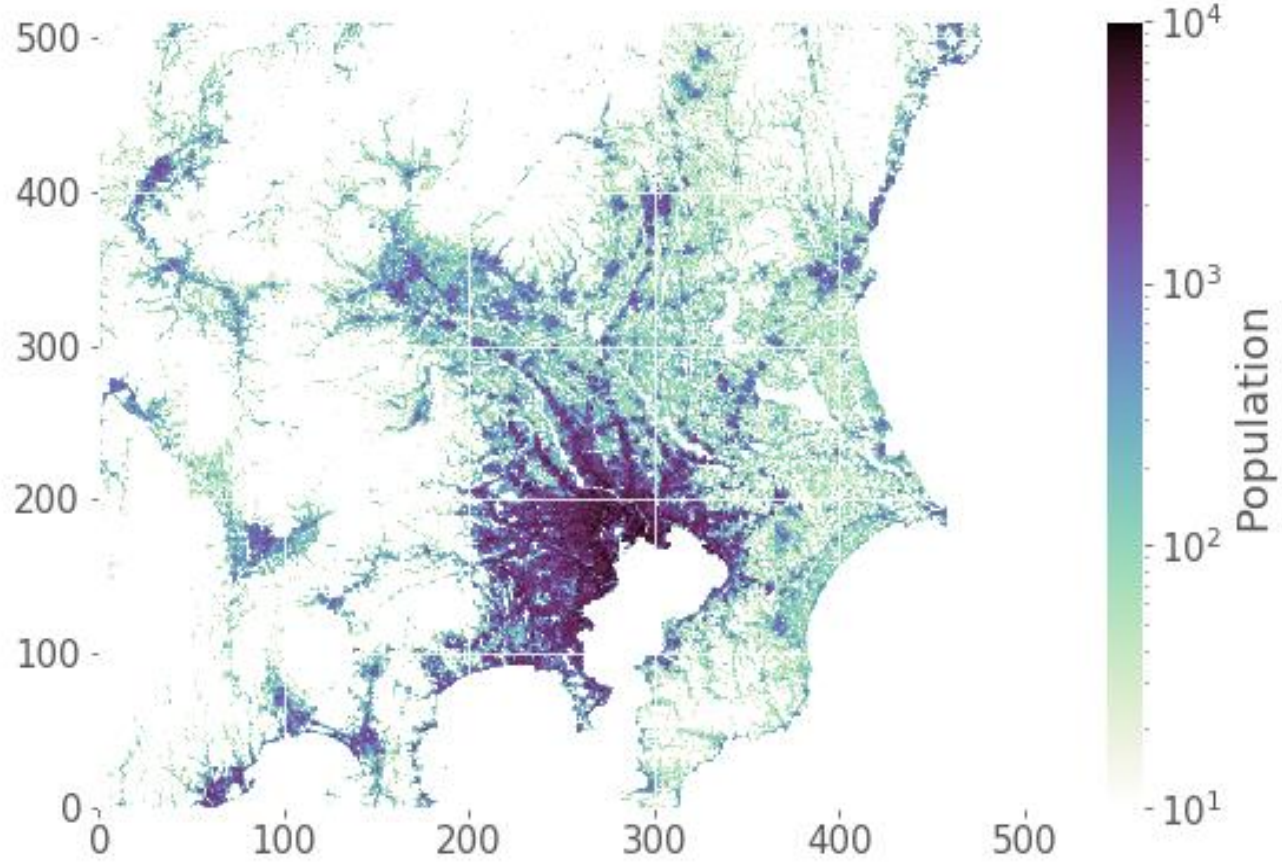
3. Experiment on the Application of Differential Privacy to Small-Area Statistics

・In this experiment several methods of differential privacy were applied to data from the 2010 Population Census for a 1/2 standard regional mesh (n=512×512=262,144 meshes, 500m per side) over a 256-km2 area in the Tokyo region.

・Results from applying the Laplace mechanism and the Laplace mechanism with nonnegativity correction were compared with results from applying the mechanism proposed by Terada et al.(2015).

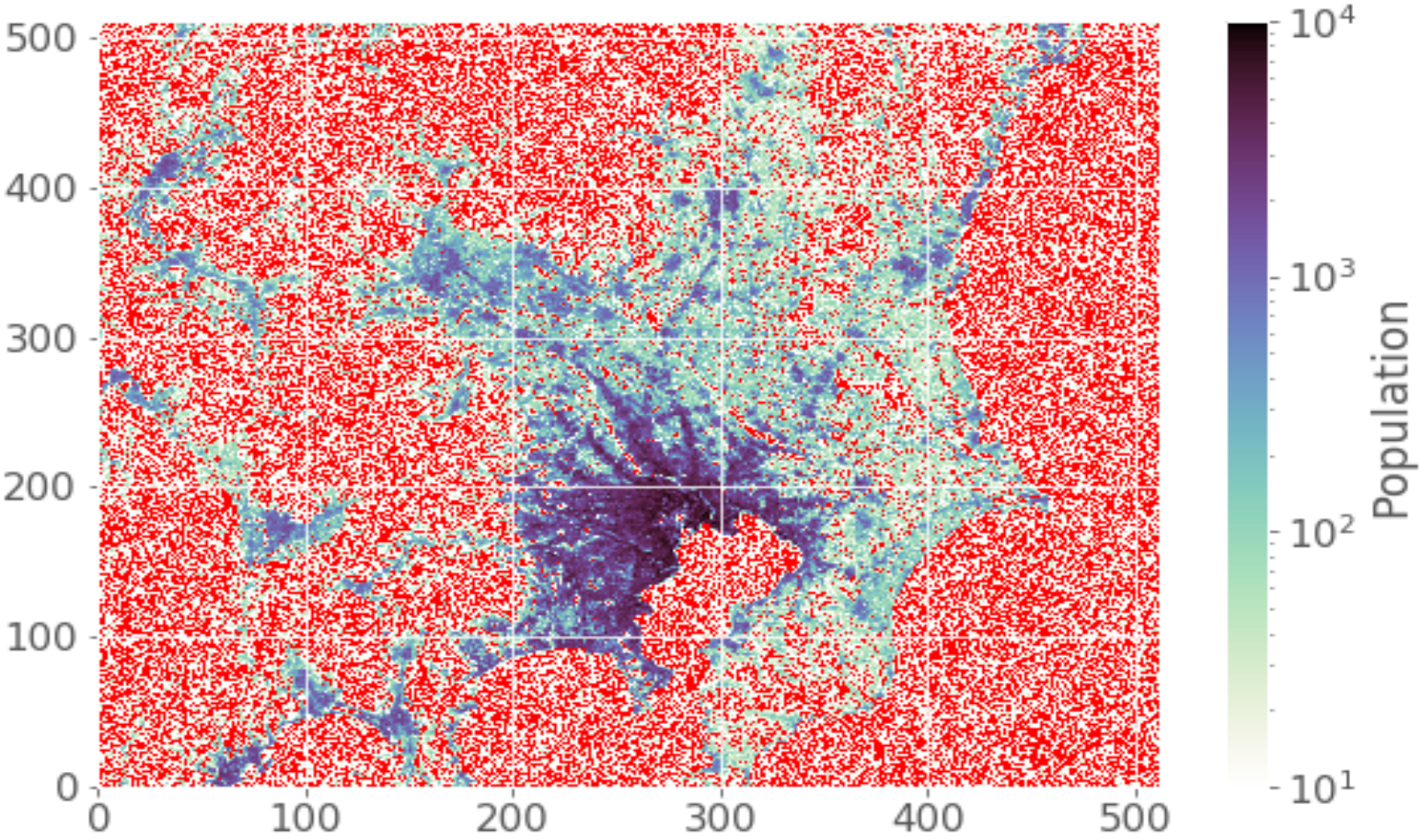
Regarding partial sums error in the output for each method, we changed the privacy loss budget ε to values in {0.1, 0.2, ln 2, ln 3, 3, 5} and compared the results.

Fig. 1 Results based on the original data



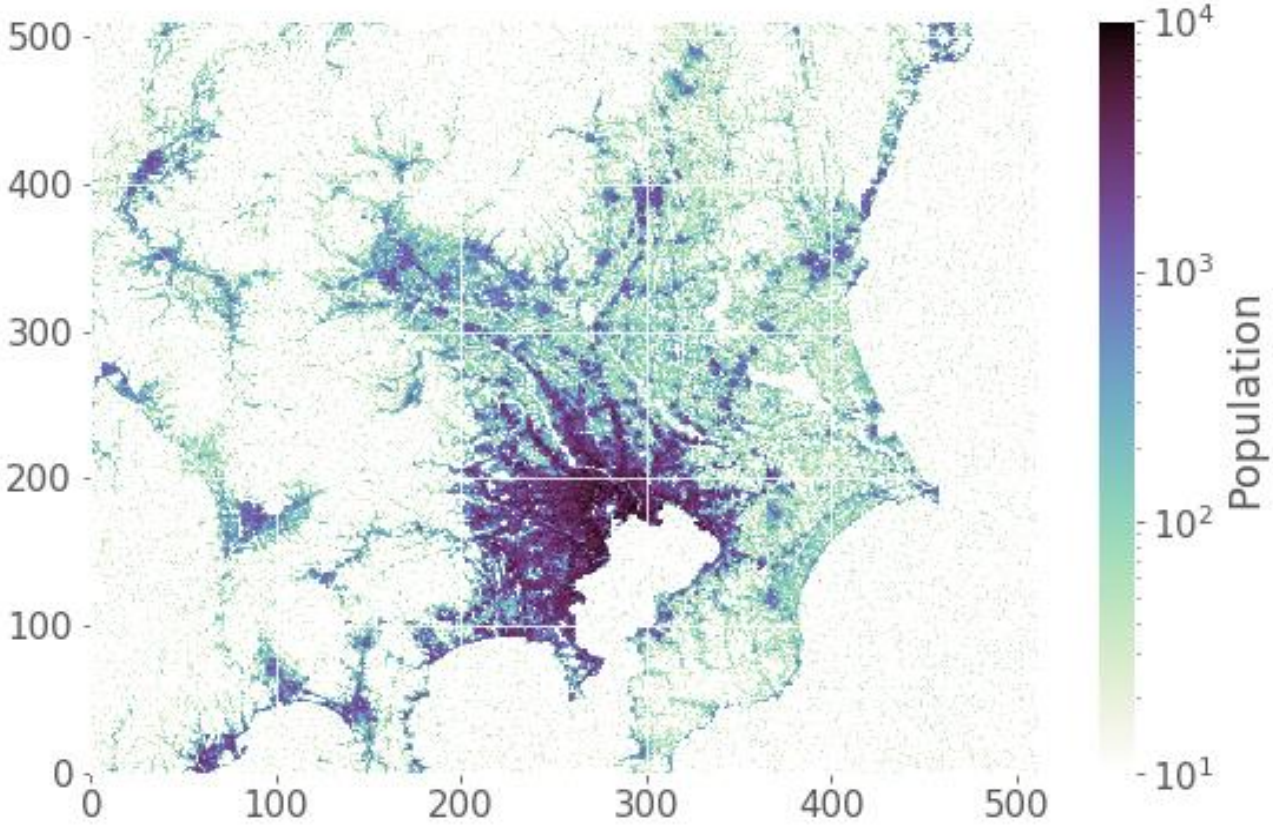
This is a remarkably high density for a Census dataset as the region includes areas with the highest population densities in Japan, while it also includes uninhabitable areas.

Fig. 2 Results based on the Laplace mechanism, ε=0.1



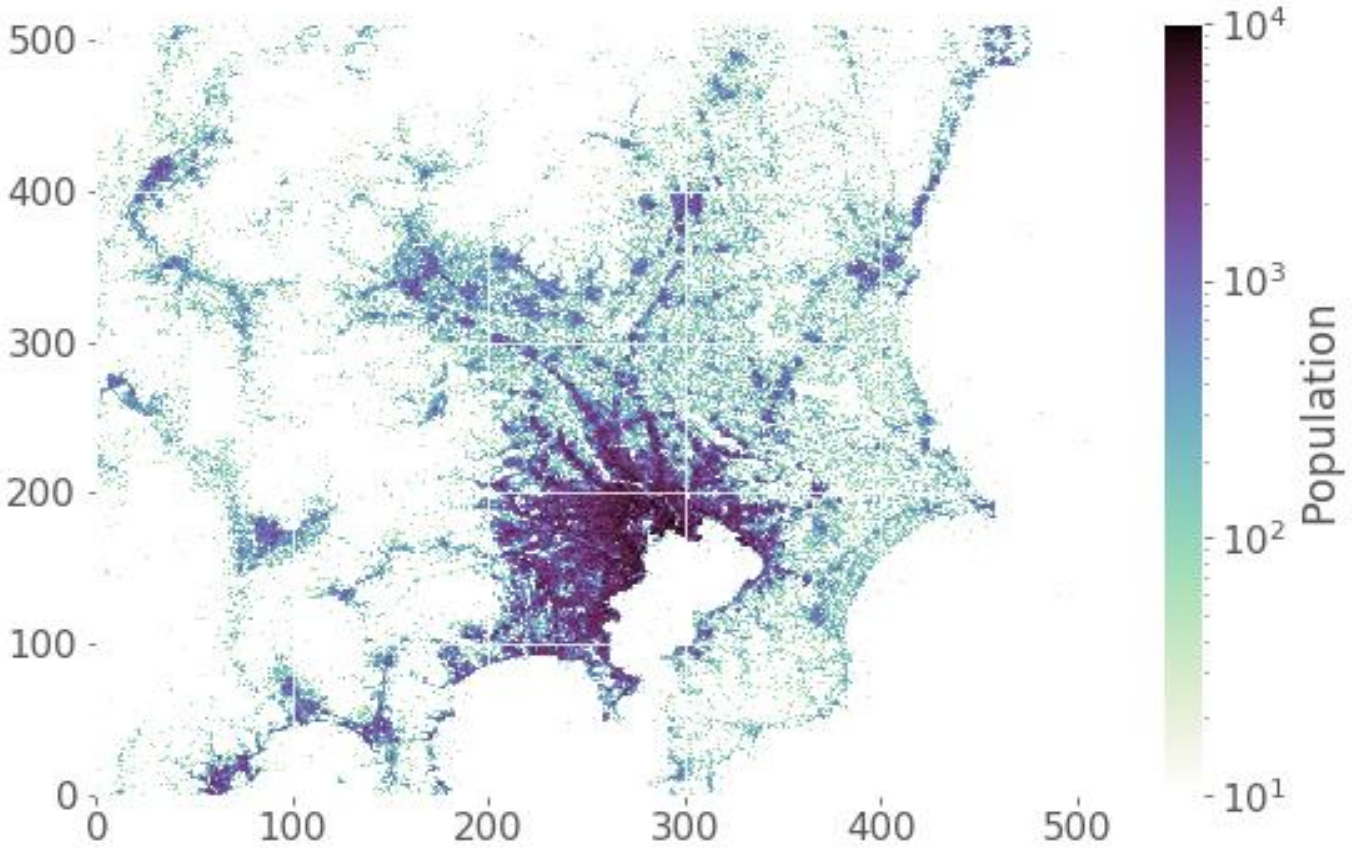
This reflects the nature of the Laplace mechanism, in which the sparsity of output data will be almost 100% lost, regardless of the sparsity of the original data.

Fig. 3 Results based on the Laplace mechanism with nonnegativity correction, ε=0.1.



All negative values in the output (values for 88,399 cells) are replaced with zeros, density decreased to approximately 64%. However, this still represents an approximate doubling of the density compared to the original data.

Fig. 4 Results of applying the method in Terada et al. (2015), with ε=0.1



There are no red regions indicating violation of nonnegativity constraints, so this issue is resolved. Further, sparsity is approximately 28%, so sparsity of the original data is largely maintained.

Table 1 Output error under the Laplace mechanism

Area size (km^2)	1^2	2^2	4^2	8^2	16^2	32^2	64^2	128^2	256^2
$\varepsilon = 0.1$	21.9	44.8	90.2	180.4	362.1	738.9	1,483.3	3,068.9	6,547.7
$\varepsilon = 0.2$	10.9	22.4	45.0	90.4	181.3	363.7	723.9	1,389.5	2,807.1
$\varepsilon = \ln 2$	3.2	6.5	13.0	26.0	51.7	104.1	208.8	415.3	865.7
$\varepsilon = \ln 3$	2.0	4.1	8.2	16.4	32.9	65.4	131.0	282.0	625.0
$\varepsilon = 3$	0.7	1.5	3.0	6.0	12.0	23.8	48.8	102.7	209.0
$\varepsilon = 5$	0.4	0.9	1.8	3.6	7.2	14.4	28.9	59.4	131.6

The error is small for partial sums over small regions, but greatly increases for larger regions. Furthermore, results retain the above-described issues of violated constraints and increased data quantity.

Table 2 Output error under the Laplace mechanism with nonnegativity correction

Area size (km^2)	1^2	2^2	4^2	8^2	16^2	32^2	64^2	128^2	256^2
$\varepsilon = 0.1$	20.1	64.5	229.4	876.9	3,464.1	13,824.3	55,291.9	221,167.5	884,669.9
$\varepsilon = 0.2$	10.1	32.0	112.7	428.2	1,686.6	6,728.5	26,910.5	107,641.9	430,567.6
$\varepsilon = \ln 2$	2.9	9.2	32.1	121.1	475.3	1,895.0	7,578.1	30,312.5	121,249.8
$\varepsilon = \ln 3$	1.9	5.8	20.2	76.1	298.8	1,190.4	4,760.4	19,041.7	76,166.6
$\varepsilon = 3$	0.7	2.1	7.4	27.8	109.1	434.6	1,738.0	6,952.0	27,807.8
$\varepsilon = 5$	0.4	1.3	4.4	16.7	65.5	260.9	1,043.2	4,172.7	16,690.8

The Laplace mechanism with nonnegativity correction does not violate constraints, and the increase in data is somewhat mitigated, but as the area of partial sums becomes larger, the error becomes even larger than under the normal Laplace mechanism.

Table 3 Output error under the method of Terada et al. (2015).

Area size (km^2)	1^2	2^2	4^2	8^2	16^2	32^2	64^2	128^2	256^2
$\varepsilon = 0.1$	44.5	60.1	75.0	88.3	100.1	104.3	106.3	119.7	147.8
$\varepsilon = 0.2$	24.6	31.9	38.8	44.8	49.4	53.8	60.2	64.3	109.4
$\varepsilon = \ln 2$	7.9	9.8	11.6	13.1	14.5	15.7	16.6	17.2	23.4
$\varepsilon = \ln 3$	5.1	6.3	7.4	8.3	9.1	9.9	10.5	11.3	16.6
$\varepsilon = 3$	1.9	2.4	2.7	3.1	3.4	3.6	3.9	4.4	6.3
$\varepsilon = 5$	1.2	1.4	1.6	1.9	2.0	2.2	2.4	2.5	3.7

Under the method of Terada et al. (2015), the error increase accompanying expansion of the partial sum region is much better controlled.

This experiment demonstrates that the method proposed in Terada et al. (2015) avoids constraint violations and increases in data, and furthermore addresses the issue of degraded partial-sum accuracy.

4. Conclusion and Outlook

(1) This paper examines the current state of secondary use of official statistics and the applicability of differential privacy for data anonymization in Japan.

(2) Our experiment using Japanese Census data demonstrates that applying noise to statistical tables based on the concept of differential privacy results in changes of cell values according to the value of ε.

(3) Results show that identical values of ε can result in different degrees of noise addition depending on the underlying mechanism used for applying differential privacy.

(4) While the adoption of differential privacy for official statistics in Japan could be viewed as challenging, it is worthwhile to consider the concept of differential privacy in order to further advance developments in official statistics in Japan.