CSIS Discussion Paper No. 140

Method for Noise Addition for Individual Record Preserving Privacy and Statistical Characteristics: Case Study of Real Estate Transaction Data

Yuzo Maruyama Ryoko Tone and Yasushi Asami

The University of Tokyo
e-mail:
maruyama@csis.u-tokyo.ac.jp; ryo-t@ua.t.u-tokyo.ac.jp; asami@csis.u-tokyo.ac.jp

Abstract: We propose a new method to perturb a major variable by adding noise, so that the result of regression analysis is not affected by the noise addition. The extent of the perturbation can be controlled by a single parameter, which will ease the perturbation operation in the actual application. Through a numerical experiment, we recommend an appropriate value of the parameter to achieve both the sufficient perturbation to mask the original values and the sufficient coherence between the perturbed data and the original data.

1. Introduction

More and more detailed information becomes circulated thanks to the recent advancement of digitalization. Personal information protection also becomes important. Personal information, which can identify the person, cannot be publicized nor utilized without the person's consent. In case of information regarding real estates, the correct position can be identified by combining several sources of information, which in turn may be used to identify the person, such as owners or residents. Spatial information such as information on real estates can be said to be information such that special attention is necessary to protect personal information.

In dealing with privacy sensitive information, the following two aspects are important. First, if the information is leaked, then the organization who is responsible for the information may have risk of claim for compensation due to privacy protection failure. Second, due to this fact, publicized data tend to become very rough or vague to avoid possible troubles, which often hinders the usefulness of analyses in real estate to understand the market.

To deal with this situation, one promising way will be to add noise to the acquired data, so that the personal information is protected. One typical example of sensitive information is the transaction data. The transaction data may include transacted price, characteristics of the real estate, characteristics of transacting persons, and information of the condition of the transaction. Publicized data tend to omit the information on characteristics of transacting persons, and hence such contents are not assumed to be included in the database. In this case, one of the most sensitive data will be the transacted price. If we add noise on the price data, then the private information will be protected. However, if the noise is added tactlessly, the results of the analysis of the data will be seriously distorted. Therefore, a way to add noise, achieving that the analyses of the data are not distorted and that the privacy is protected, is very important. This paper is devoted to propose such a method as well as its application assuming that the main concern of the analyses is the hedonic analysis, i.e., regression analysis with the transacted price being the response variable.

Takemura (2003) reviewed statistical issues in publicizing individual data. To protect personal information, he lists up several methods, such as (1) direct hiding by making the information secrete, (2) global categorization by classifying the values into several coarse classes, (3) disturbance by replacing with different values (such as swapping by exchanging individual values, post randomization method (PRAM), addition of noise). Direct hiding and global categorization are not appropriate for releasing the data for detailed analyses, for the resolution of the information becomes too coarse sometimes. Disturbance method is superior in this aspect, but usually it brings some errors into analyses, and such effects have to be carefully examined.

One famous method to protect the personal information is the statistical disclosure limitation (SDL) method. SDL method is a general term for methods to protect identification of personal data by adding perturbation, modification or summarization (Shlomo (2010)). The main concern is to reduce the risk of identification as well as to keep the usability of the data.

Typically, to reduce the risk of identification, three kinds of methods are often utilized, namely, (1) method to make coarse categorization, (2) method to generate new data such that statistical characteristics are similar to the original data, and (3) method to add noise to the original data (Karr et al. (2006); Oganian and Karr (2011)).

Lots of research are done regarding the method to make coarse categorization. In particular, population uniqueness is extensively studied, which is the feature that a combination of attributes becomes unique in the parent population. For example, Manrique-Vallier and Reiter (2012) estimates the risk of suffering from the population uniqueness for discrete data.

Regarding the method to generate new data, swapping method is well known, in which categorical data are probabilistically exchanged. This is called PRAM (post-randomization method) by perturbing the exchanging categorical data (Gouweleeuw et al. (1998); Willenborg and Waal (2001)). In this method, transition probability matrix is first determined, and the categorical data are exchanged based on the matrix, so that the proportion of each category remains the same.

A variety of methods to add noise are proposed, such that the qualitative feature is carefully maintained. For example, Oganian and Karr (2011) focuses on features such as the positivity of values and magnitude relation between two values. They propose a method to add noise such that the positivity of values, mean values and variance-covariance matrix remain the same. One remarkable idea is to use multiplicative noise addition to avoid getting the negative values. Moreover, they show the stability of results after the regression analyses. A similar method to maintain characteristics of attributes is proposed by Abowd and Woodcock (2001). Another method to add noise avoiding the risk of identification is to introduce random noise, which is distributed following a peculiar symmetric distribution with a hole in the central part. With this method, the perturbed value is never close to the original value, and therefore the risk of identification is drastically reduced. In the actual application of this method, the noise distribution is not publicized, which hinders the analyses using the distribution (Reiter (2012)).

In general, addition of noise may influence the quality of subsequent analyses. Fuller (1993) points out that the addition of noise has similar influence as introducing measurement errors to explanatory variables. To minimize the influence in particular analyses, several methods are devised. For example, some methods maintain the original mean values and variance-covariance matrix (Ting, Fienberg and Trottini (2008); Shlomo and De Waal (2008)). In our paper, by focusing on the regression analysis, a method is proposed so that the results are robust by adding noise.

The paper is organized as follows. In Section 2, a method is proposed to add noise to response variable, and show some important statistics do not change with the addition of noise. In Section 3, numerical experiments are conducted to examine how the results of multivariate analyses, apart from the assumed regression analysis, may change. Finally, Section 4 concludes with the summary and possible extension of method.

2. Theoretical results

We assume that the $n \times (p+1)$ design matrix $\boldsymbol{X}$ is given by $(\mathbf{1}_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, where $\mathbf{1}_n$ is n-dimensional vector of ones, and the n-dimensional response vector is $\boldsymbol{y}$. We also assume that n is sufficiently larger than p and that the rank of $\boldsymbol{X}$ is $p+1$. Then the OLS estimator is

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)' = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{y}.$$

A decomposition of $\boldsymbol{y}$ based on the OLS estimator $\hat{\boldsymbol{\beta}}$ is $\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e}$ where

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{y}$$

is the predictive vector and

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}')\boldsymbol{y}$$

is the residual vector. Then the coefficient of determination defined by

$$R^2 = 1 - \frac{\|\boldsymbol{e}\|^2}{\|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^2} \tag{2.1}$$

where $\bar{y}$ is the sample mean of $\boldsymbol{y}$, measures the goodness of the fit under the use of the OLS estimator $\hat{\boldsymbol{\beta}}$. The coefficient of determination, R^2, is hence regarded as a key quantity of regression analysis. The t-value of regression coefficient β_j for $j = 0, 1, \ldots, p$, is another key quantity, which is defined by

$$t_j = \frac{\sqrt{n-p-1}}{d_j} \frac{\hat{\beta}_j}{\|\boldsymbol{e}\|} \tag{2.2}$$

where d_j is the $(j+1)$-th diagonal component of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Under Gaussian linear regression setting, t_j has Student's t-distribution with $n-p-1$ degrees of freedom under the null hypothesis $\beta_j = 0$.

In this paper, we are interested in adding perturbation to the original response vector together with tractable tuning of R^2 and t-values. We start with any n-dimensional random vector

$$\boldsymbol{v} = (v_1, \ldots, v_n)'.$$

Since n is sufficiently larger than p, $\boldsymbol{v}$ cannot be expressed by a linear combination of $\boldsymbol{e}, \boldsymbol{1}_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ with probability 1. In other words,

$$\boldsymbol{u} = \left(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' - \boldsymbol{e}\boldsymbol{e}'/\|e\|^2\right)\boldsymbol{v}, \tag{2.3}$$

cannot be the zero vector. The noise vector we consider in this paper is a linear combination of $\boldsymbol{e}$ and $\boldsymbol{u}$, given by

$$\boldsymbol{\epsilon} = \frac{a\|\boldsymbol{e}\|}{1+b}\left\{\frac{\boldsymbol{e}}{\|\boldsymbol{e}\|} + \sqrt{b}\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}\right\}, \tag{2.4}$$

where $a \neq 0$ and $b \geq 0$.

When $\boldsymbol{y} + \boldsymbol{\epsilon}$ is used instead of the original response vector $\boldsymbol{y}$, we have a following result.

Theorem 2.1. *1. The sample mean of $\boldsymbol{y} + \boldsymbol{\epsilon}$ is $\bar{y}$ for any a and b.*

2. The OLS estimator for the response vector $\boldsymbol{y} + \boldsymbol{\epsilon}$ keeps the same for any a and b, that is,

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{y} + \boldsymbol{\epsilon}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

3. The t-values for the response vector $\boldsymbol{y} + \boldsymbol{\epsilon}$ are given by

$$\tilde{t}_j = \left\{\frac{1+b}{1+b+a(a+2)}\right\}^{1/2} t_j,$$

for $j = 0, \ldots, p$.

4. The coefficient of determination for the response vector $\boldsymbol{y} + \boldsymbol{\epsilon}$ is

$$\tilde{R}^2 = \left\{\frac{1+b}{1+b+a(a+2)(1-R^2)}\right\} R^2.$$

5. *The correlation coefficient of* $\boldsymbol{y}$ *and* $\boldsymbol{y}+\boldsymbol{\epsilon}$ *is*

$$r_{y,y+\epsilon} = \frac{1+b+a(1-R^2)}{(1+b)^{1/2}\{1+b+a(a+2)(1-R^2)\}^{1/2}}.$$

Proof. By Part 3 of Lemma 2.1, we have $\boldsymbol{X}'\boldsymbol{\epsilon} = \boldsymbol{0}$, the first component of which is $\boldsymbol{1}_n'\boldsymbol{\epsilon} = 0$. Hence Part 1 follows.

Since $\boldsymbol{X}'\boldsymbol{\epsilon} = \boldsymbol{0}$, we have

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{y}+\boldsymbol{\epsilon}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \quad (2.5)$$

which completes the proof of Part 2.

Note that the t-values are defined by (2.2). By (2.5), any component of the OLS estimator keeps the same. Further $\sqrt{n-p-1}/d_j$ does not depend on the response vector. Hence Part 3 follows from Part 6 of Lemma 2.1.

Note the coefficient of determination is defined by (2.1). Since the sample mean of $\boldsymbol{y}+\boldsymbol{\epsilon}$ is also $\bar{y}$ as in Part 1 of this theorem, the coefficient of determination for the response vector $\boldsymbol{y}+\boldsymbol{\epsilon}$ is

$$1 - \frac{\|(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')(\boldsymbol{y}+\boldsymbol{\epsilon})\|^2}{\|\boldsymbol{y}+\boldsymbol{\epsilon}-\bar{y}\boldsymbol{1}_n\|^2},$$

which is rewritten as

$$1 - \frac{\{1+a(a+2)/(1+b)\}\,\|\boldsymbol{e}\|^2}{\|\boldsymbol{y}-\bar{y}\boldsymbol{1}_n\|^2 + \{a(a+2)/(1+b)\}\,\|\boldsymbol{e}\|^2}$$

by Parts 5 and 6 of Lemma 2.1. By the definition of R^2, we have

$$1 - R^2 = \|\boldsymbol{e}\|^2/\|\boldsymbol{y}-\bar{y}\boldsymbol{1}_n\|^2, \quad (2.6)$$

which completes the proof of Part 4.

The correlation coefficient of $\boldsymbol{y}$ and $\boldsymbol{y}+\boldsymbol{\epsilon}$ is

$$\frac{(\boldsymbol{y}-\bar{y}\boldsymbol{1}_n)'(\boldsymbol{y}+\boldsymbol{\epsilon}-\bar{y}\boldsymbol{1}_n)}{\|\boldsymbol{y}-\bar{y}\boldsymbol{1}_n\|\|\boldsymbol{y}+\boldsymbol{\epsilon}-\bar{y}\boldsymbol{1}_n\|}.$$

By Parts 3 and 4 of Lemma 2.1 as well as (2.6), we have

$$\begin{aligned}(\boldsymbol{y}-\bar{y}\boldsymbol{1}_n)'(\boldsymbol{y}+\boldsymbol{\epsilon}-\bar{y}\boldsymbol{1}_n) &= \|\boldsymbol{y}-\bar{y}\boldsymbol{1}_n\|^2 + (\boldsymbol{y}-\bar{y}\boldsymbol{1}_n)'\boldsymbol{\epsilon} \\ &= \|\boldsymbol{y}-\bar{y}\boldsymbol{1}_n\|^2 + (\hat{\boldsymbol{y}}+\boldsymbol{e}-\bar{y}\boldsymbol{1}_n)'\boldsymbol{\epsilon} \\ &= \|\boldsymbol{y}-\bar{y}\boldsymbol{1}_n\|^2 + (\boldsymbol{X}\hat{\boldsymbol{\beta}}-\bar{y}\boldsymbol{1}_n)'\boldsymbol{\epsilon} + \boldsymbol{e}'\boldsymbol{\epsilon} \\ &= \|\boldsymbol{y}-\bar{y}\boldsymbol{1}_n\|^2 + \boldsymbol{e}'\boldsymbol{\epsilon} \\ &= \|\boldsymbol{y}-\bar{y}\boldsymbol{1}_n\|^2\left[1+\{a/(1+b)\}(1-R^2)\right].\end{aligned}$$

Further, by Part 5 of Lemma 2.1, we have

$$\|\boldsymbol{y}-\bar{y}\boldsymbol{1}_n+\boldsymbol{\epsilon}\|^2 = \|\boldsymbol{y}-\bar{y}\boldsymbol{1}_n\|^2\left[1+\{a(a+2)/(1+b)\}(1-R^2)\right],$$

which completes the proof of Part 5. □

The lemma below summarizes fundamental properties related to $\boldsymbol{e}$ and $\boldsymbol{\epsilon}$, which are needed in the proof of Theorem 2.1.

Lemma 2.1. *1.* $\boldsymbol{e}$ *is orthogonal to* $\mathbf{1}_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ *or equivalently* $\boldsymbol{X}'\boldsymbol{e} = \mathbf{0}$.

2. $\boldsymbol{u}$ *is orthogonal to* $\boldsymbol{e}, \mathbf{1}_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ *or equivalently* $\boldsymbol{X}'\boldsymbol{u} = \mathbf{0}$ *and* $\boldsymbol{e}'\boldsymbol{u} = 0$.

3. $\boldsymbol{X}'\boldsymbol{\epsilon} = \mathbf{0}$.

4. $\boldsymbol{e}'\boldsymbol{\epsilon} = a\|\boldsymbol{e}\|^2/(1+b)$ *and* $\|\boldsymbol{\epsilon}\|^2 = a^2\|\boldsymbol{e}\|^2/(1+b)$.

5. The sum of squared deviation of $\boldsymbol{y} + \boldsymbol{\epsilon}$ *is*

$$\|\boldsymbol{y} + \boldsymbol{\epsilon} - \bar{y}\mathbf{1}_n\|^2 = \|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^2 + \{a(a+2)/(1+b)\}\,\|\boldsymbol{e}\|^2.$$

6. The residual sum of squares for $\boldsymbol{y} + \boldsymbol{\epsilon}$ *is*

$$\|(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')(\boldsymbol{y} + \boldsymbol{\epsilon})\|^2 = \{1 + a(a+2)/(1+b)\}\,\|\boldsymbol{e}\|^2.$$

Proof. Since $\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{X}'$, we have

$$\begin{aligned}\left(\mathbf{1}_n'\boldsymbol{e}\ \boldsymbol{x}_1'\boldsymbol{e}\ \cdots\ \boldsymbol{x}_p'\boldsymbol{e}\right)' &= \boldsymbol{X}'\boldsymbol{e} \\ &= \boldsymbol{X}'\left(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)\boldsymbol{y} \\ &= \left(\boldsymbol{X}' - \boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)\boldsymbol{y} \\ &= \mathbf{0},\end{aligned}$$

which completes the proof of Part 1. In the same way, Part 2 can be proved.

Recall $\boldsymbol{\epsilon}$ is given by a linear combination of $\boldsymbol{e}$ and $\boldsymbol{u}$,

$$\boldsymbol{\epsilon} = \frac{a\|\boldsymbol{e}\|}{1+b}\left\{\frac{\boldsymbol{e}}{\|\boldsymbol{e}\|} + \sqrt{b}\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}\right\}. \tag{2.7}$$

Then Part 3 follows from Parts 1 and 2. Part 4 follows from the orthogonality of $\boldsymbol{e}$ and $\boldsymbol{u}$ together with (2.7).

Since the sample mean of $\boldsymbol{y} + \boldsymbol{\epsilon}$ is $\bar{y}$ by Part 1 of Theorem 2.1, the sum of squared deviation of $\boldsymbol{y} + \boldsymbol{\epsilon}$, $\|\boldsymbol{y} + \boldsymbol{\epsilon} - \bar{y}\mathbf{1}_n\|^2$, is expanded as

$$\|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^2 + 2(\boldsymbol{y} - \bar{y}\mathbf{1}_n)'\boldsymbol{\epsilon} + \|\boldsymbol{\epsilon}\|^2.$$

By Part 3, we have

$$(\boldsymbol{y} - \bar{y}\mathbf{1}_n)'\boldsymbol{\epsilon} = (\hat{\boldsymbol{y}} + \boldsymbol{e} - \bar{y}\mathbf{1}_n)'\boldsymbol{\epsilon} = (\boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{e} - \bar{y}\mathbf{1}_n)'\boldsymbol{\epsilon} = \boldsymbol{e}'\boldsymbol{\epsilon} = a\|\boldsymbol{e}\|^2/(1+b).$$

Then Part 5 follows from Part 4.

Since $\boldsymbol{X}'\boldsymbol{\epsilon} = \mathbf{0}$ by Part 3, we have

$$(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')(\boldsymbol{y} + \boldsymbol{\epsilon}) = \boldsymbol{e} + \boldsymbol{\epsilon}.$$

From Part 4, the residual sum of squares is

$$\|\boldsymbol{e} + \boldsymbol{\epsilon}\|^2 = \|\boldsymbol{e}\|^2 + 2\boldsymbol{e}'\boldsymbol{\epsilon} + \|\boldsymbol{\epsilon}\|^2 = \|\boldsymbol{e}\|^2 + 2a\|\boldsymbol{e}\|^2/(1+b) + a^2\|\boldsymbol{e}\|^2/(1+b),$$

which completes the proof of Part 6. □

By Theorem 2.1, we see that $a = -2$ is a special case as follows.

Theorem 2.2. *Assume $a = -2$. Then, we have the followings.*

1. *For any $b > 0$, the coefficient of determination for $\boldsymbol{y} + \boldsymbol{\epsilon}$ is equal to R^2, the coefficient of determination for the original $\boldsymbol{y}$.*
2. *For any $b > 0$, the t-value of β_j $(j = 0, 1, \ldots, p)$ for the response vector $\boldsymbol{y} + \boldsymbol{\epsilon}$ is equal to t_j.*
3. *The correlation coefficient of $\boldsymbol{y}$ and $\boldsymbol{y} + \boldsymbol{\epsilon}$ is*

$$r_{y,y+\epsilon} = 1 - \frac{2(1 - R^2)}{1 + b}. \tag{2.8}$$

Recall that $\boldsymbol{\epsilon}$ is a function of $\boldsymbol{v}$, any random n-dimensional vector, through the relationship, (2.3) and (2.4), that is,

$$\boldsymbol{u} = \left(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' - \frac{\boldsymbol{e}\boldsymbol{e}'}{\|\boldsymbol{e}\|^2}\right)\boldsymbol{v}, \ \boldsymbol{\epsilon} = \frac{a\|\boldsymbol{e}\|}{1 + b}\left\{\frac{\boldsymbol{e}}{\|\boldsymbol{e}\|} + \sqrt{b}\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}\right\}.$$

In Parts 1 and 2 of Theorem 2.2, the choice $a = -2$ guarantees that the coefficient of determination and t-value remain the same regardless of random $\boldsymbol{v}$.

By Part 3 of Theorem 2.2, $r_{y,y+\epsilon}$ is increasing in b for fixed R^2. The coefficient of correlation between the original response $\boldsymbol{y}$ and the perturbed response $\boldsymbol{y} + \boldsymbol{\epsilon}$ with $a = -2$, some $b \geq 0$ and R^2, is illustrated in Table 1. In the actual application, it is desirable to have relatively high correlation, because the users of the data may assume that the perturbed response are close to the original response. But if the correlation is very high, then the perturbed response is very close to the original response, and the aim to conceal the actual response cannot be achieved. Thus we need to determine the value of b, so that the perturbed response is not too close, which we will discuss through the analysis of real data in the next section.

Remark 2.1. When $a = -2$ and $b = 0$, we have $\boldsymbol{\epsilon} = -2\boldsymbol{e}$ as the noise or equivalently

$$\boldsymbol{y} - 2\boldsymbol{e} = \hat{\boldsymbol{y}} - \boldsymbol{e} \tag{2.9}$$

as the perturbed response. In this particular case, it is clear that the coefficient of determination and t-value remains the same since y_i and $y_i - 2e_i$ for $i = 1, \ldots, n$ are symmetric with respect to the point $\hat{y}_i = y_i - e_i$. Needless to say, the noise $\boldsymbol{\epsilon} = -2\boldsymbol{e}$ does not depend on $\boldsymbol{v}$ and hence there is no randomness on the noise. Theorem 2.2 ensures that with randomness through $\boldsymbol{v}$ as in (2.4), we can construct the noise $\boldsymbol{\epsilon}$ so that the coefficient of determination and t-value remains the same.

Remark 2.2. As in Theorem 2.2, we found that the choice $a = -2$ with random $\boldsymbol{v}$ surprisingly keeps R^2 and t values. Here are some remarks for the other choices. For $a \in (-\infty, -2) \cup (0, \infty)$, both R^2 and the absolute value of t values are

Table 1

Correlation coefficient of $\boldsymbol{y}$ and $\boldsymbol{y}+\boldsymbol{\epsilon}$ with $a=-2$

$R^2\backslash b$	0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0
0.4	-0.2	0.04	0.2	0.31	0.4	0.47	0.52	0.56	0.6
0.6	0.2	0.36	0.47	0.54	0.6	0.64	0.68	0.71	0.73
0.8	0.6	0.68	0.73	0.77	0.8	0.82	0.84	0.85	0.87

reduced. For example, $b>0$ and $a=-1\pm\sqrt{b+2}\in(-\infty,-2)\cup(0,\infty)$ give

$$\tilde{t}_j=\frac{1}{\sqrt{2}}t_j,\tilde{R}^2=\frac{R^2}{2-R^2}<R^2. \qquad (2.10)$$

Note that R^2 and t values can be completely controlled. The data provider safely provide the data with the relation between $\{t_j,R^2\}$ and $\{\tilde{t}_j,\tilde{R}^2\}$ described by (2.10) and practitioners can restore the original R^2 and t-value by themselves. An efficient way to open data with reduced accuracy will be reported elsewhere.

3. Numerical experiment

In the previous section, a method is proposed to add noise on the response variable. This method can be applied when real estate database is released into the public domain, by adding noise to the transacted price, which is supposed to be a sensitive information in Japan. As Theorem 2.2 in the previous section ensure, the result of regression analysis using the perturbed data will not change. However, in the actual application, a variety of analyses will be devised and the theorems do not apply for cases with unexpected applications. Thus we need to confirm if the proposed method remains appropriate even in unexpected applications.

The precision of the results may be degraded, if analytical operations are applied, which are not assumed in the theory. In such cases, acceptable levels should be decided upon which the error due to the perturbation is permitted. In the following, to know the relationship between the perturbation and the precision level, a numerical experiment is conducted.

3.1. The data used in the experiment

The data source used for the numerical experiment is provided by At Home Co., Ltd.. This data contains real estate advertisement information in 2008. By supplementing some spatial variables, the database for the experiment is created. The database contains 1320 cases for newly built detached houses in Setagaya Ward in Tokyo Prefecture[1]. The variables include price of the property (yen),

[1] We selected data which have information on designated floor area ratio and designated building coverage ratio. These elements are thought to be important in real estate analysis in Japan. In the original database, when a property owner changes its price in the advertisement, then a new record is added. In this situation, we selected only the newest record.

time to (nearest) railway station (minute), dummy variable to use bus, area of the site (square meter), floor area (square meter), dummy variable to signify leased land, designated building coverage ratio, designated floor area ratio, time to Shinjuku by railway (from the nearest station) (minute), time to Shibuya by railway (minute), time to Yokohama by railway (minute), time to Tokyo by railway (minute), width of the nearest road (meter), dummy variable to signify the nearest road in south. Note that Shinjuku, Shibuya, Yokohama and Tokyo are four major railway stations in the study region. Among these variables, time to railway stations, width of nearest road (meter) and dummy variable to signify the nearest road in south are spatial variables, as are described in the next subsection.

3.2. Creation of spatial variables

As variables to signify spatial relationship, time to major railway stations from the nearest station (minute), width of road which is the nearest from the representative point of the property (meter), and dummy variable signifying if the nearest road is located to the south of the property are added to the original database. The width of road which is the nearest from the representative point of the property is regarded as the width of adjacent road. This is because the precise digital data for lots are not available. Accordingly, the dummy variable signifying if the nearest road is located to the south of the property is regarded as the dummy variable to signify adjacency to road in south.

The time to major railway stations from the nearest station is calculated with the search system for guiding transferring railways provided by NAVITIME Japan, Co. Ltd.. This system automatically calculates the time required to go to the major railway stations, i.e., Shinjuku station, Shibuya station, Yokohama station and Tokyo station, from the nearest railway station from the property. In searching the time required, the departing time is set to be at $12:00$ (noon) on August 2nd in 2010.

The width of road which is the nearest from the representative point of the property is calculated as follows. Mapple 10000 digital data produced by Shobunsha, Publications, Inc. have digital road data classified by the road width in the categories, such as 4 to 5 meters, 5 to 6 meters, etc.. The median of each class is assigned as the width of the road. For example, the width is 4.5 meters for the class from 4 to 5 meters. With ArcGIS 10, a GIS software, the nearest road is assigned for each property, and the width of the road calculated as the above is set to be the width of road which is the nearest from the representative point of the property.

In real estate market in Japan, if the residential lot is adjacent to road in south, then the lot tends to be evaluated highly. This is because getting much sunlight is preferred in Tokyo. For example, The Real Estate Transaction Modernization Center (1986) deals with the properties adjacent to road in south as a favorable merit in appraising properties. With this in mind, the dummy variable signifying if the nearest road is located to the south of the property is also added to the database.

TABLE 2
Summary statistics of variables

	min	max	mean	s.d.
price of the property (yen)	34800000	330000000	72431491	25539447
time to (nearest) railway station (minute)	0	25	10.60	4.83
d.v.[a] to use bus	0	1	0.07	0.26
area of the site (square meter)	29.53	211.49	88.56	25.48
floor area (square meter)	47.07	228.48	98.94	20.06
d.v.[a] to signify leased land	0	1	0.03	0.17
designated building coverage ratio	40	80	54.18	7.70
designated floor area ratio	80	300	141.43	47.10
time to Shinjuku by railway (minute)	5	32	18.72	5.29
time to Shibuya by railway (minute)	3	29	14.86	6.01
time to Yokohama by railway (minute)	17	64	44.30	10.99
time to Tokyo by railway (minute)	23	48	34.09	4.90
width of adjacent road (meter)	4.5	35	5.80	2.25
d.v.[a] to signify adjacency to road in south	0	1	0.28	0.45

[a] d.v. stands for "dummy variable".

This dummy variable is constructed as follows. By using ArcGIS 10, the direction to the nearest road is calculated, such that 0 degree is to the east, that the value increases up to 180 degrees counter-clockwise and that the value decreases up to -180 degrees clockwise. If the degree is in the range from -135 to -45, then it is judged to the south, and the dummy variable is set to 1, and otherwise it is set to 0.

The summary statistics of the variables are shown in Table 2.

3.3. Numerical experiment with perturbed property price

The perturbed property price, which is generated by adding noise to response variable through the method described in the previous section, is numerically tested in this subsection. The explanatory variables are the rest 13 variables in Table 2.

3.3.1. Some statistics of the perturbed property price

Although Part 1 of Theorem 2.1 guarantees that the mean of perturbed response variable is exactly equal to the mean of the original response, the equality or similarity of other statistics including the minimum value, the maximum value, the first and third quantile, and so on, cannot be theoretically controlled. In this section, we generate four sets of quasi response variable with different $\boldsymbol{v}$, $a = -2$ and $b = 1$, and will see the degree of perturbation of some statistics among five sets including the original response (original, quasi1, quasi2, quasi3, quasi4).

First of all, Figure 1 provides the boxplot of five sets. We see that when the original and quasi response variables are compared, the median are very similar

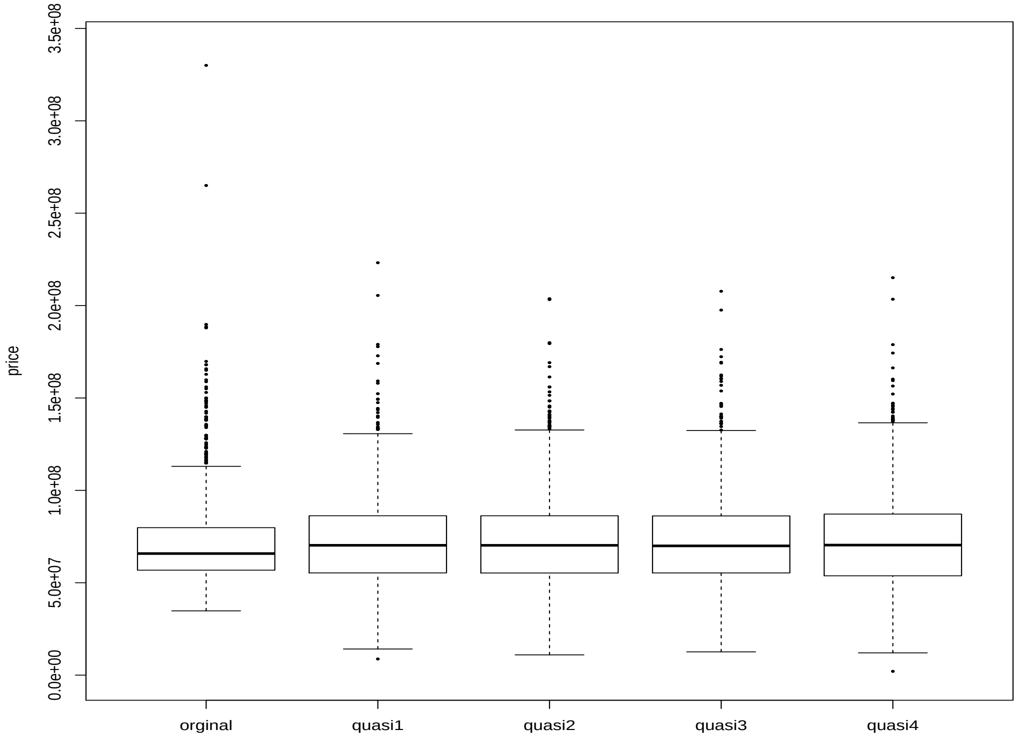

Fig 1. *Boxplot of 5 sets of response variable*

each other, but the quantiles, minimum and maximum are quite different. We also see that among four sets of quasi variable, all statistics are similar.

Figure 2 gives scatterplot of the original and 4 sets of quasi variables. Although four plots look very similar each other, it goes without saying that different $\boldsymbol{v}$ implies different quasi variable, as we explained in Section 2. Table 3 gives the correlation matrix of five sets of response variables. By Part 3 of Theorem 2.2, the correlation coefficient of the original and quasi variable is theoretically given by $1 - 2(1 - R^2)/(1 + b)$, which equals to R^2 for $b = 1$. Among quasi variables, the correlation in all cases is around 0.78.

Remark 3.1. In this particular data set, the response variable is property price, which should be positive. Hence the positive perturbed price is strongly desirable. As we claimed in Remark 2.1, for sufficiently small b, we have

$$\boldsymbol{y} + \boldsymbol{\epsilon} \approx \hat{\boldsymbol{y}} - \boldsymbol{e}.$$

Suppose there exist individuals i with relatively expensive price y_i while rela-

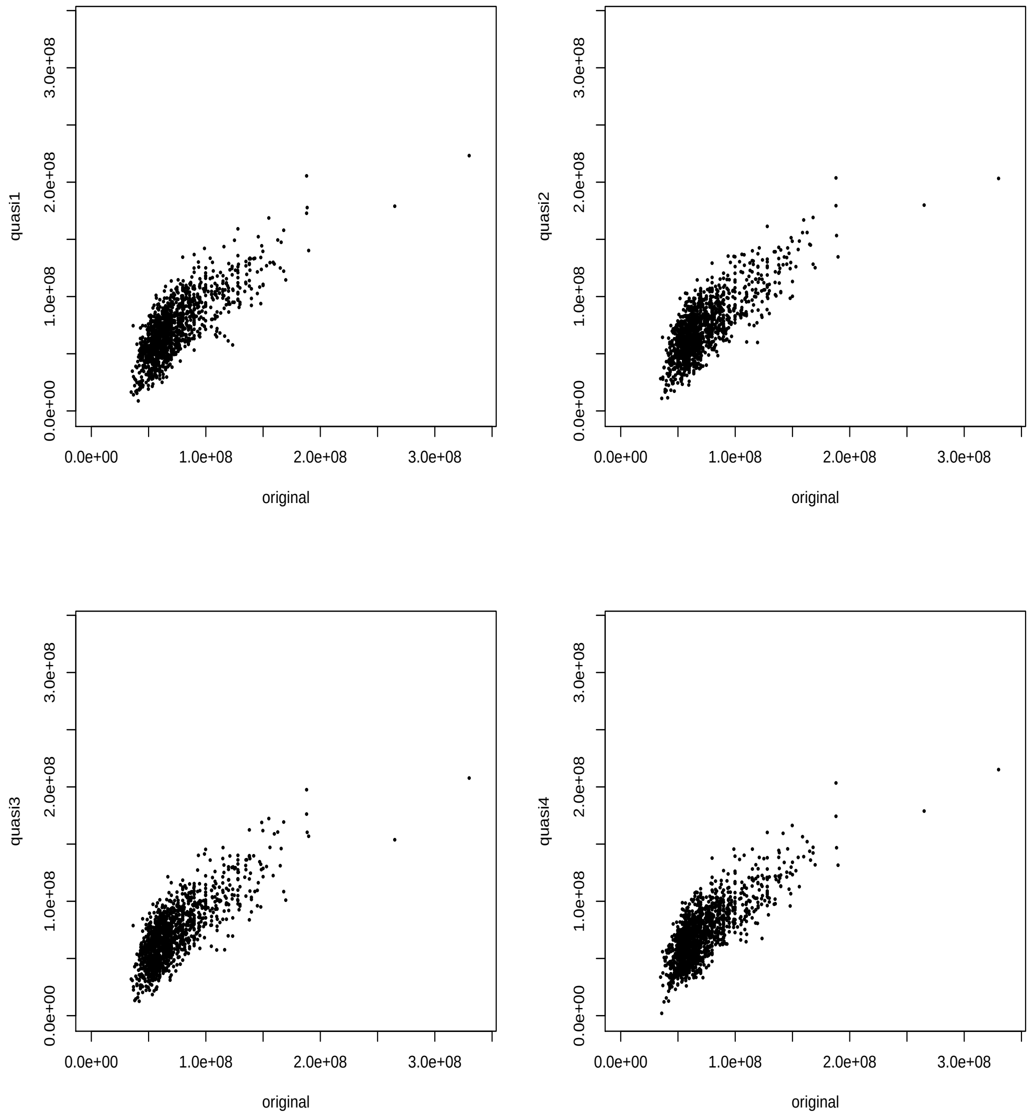

FIG 2. *Scatterplots of 5 sets of response variable*

TABLE 3
Correlation among original response and 4 sets of quasi response

	orig	quasi1	quasi2	quasi3	quasi4
orig	1	0.7748	0.7748	0.7748	0.7748
quasi1	0.7748	1	0.7693	0.7749	0.7814
quasi2	0.7748	0.7693	1	0.7870	0.7812
quasi3	0.7748	0.7749	0.7870	1	0.7733
quasi4	0.7748	0.7814	0.7812	0.7733	1

tively lower price $\hat{y}_i$ is expected. Then e_i becomes larger and as a result

$$y_i + \epsilon_i \approx \hat{y}_i - e_i < 0 \tag{3.1}$$

can happen. In our data set, this rarely happens for b smaller than 1.2, while it never happen for $b = 1.3$ or more. As far as we know, it is not theoretically controllable through the choice of b and $\boldsymbol{v}$. When (3.1) happens, generation of $\boldsymbol{\epsilon}$ with different $\boldsymbol{v}$ until $\min\{y_i + \epsilon_i\} > 0$ is achieved, is simply recommended.

3.3.2. Regression analysis with only a part of the database

The theory assumes that all the data will be used for the analysis. If only a part of the perturbed data are used, then the theorems do not apply exactly. In the actual analyses for real estate data, this is the case that part of the (perturbed) database is used for the analysis. In such a case, we need to know how the results may differ from the genuine results and make some guidance for choosing the appropriate value for b.

To do so, critical value for b is obtained so that the difference is not significant statistically, between the regression model for the perturbed property price as the response variable and that for the original property price.

3.3.3. Chow test

From 1320 cases (total database), 20% (i.e., 264 cases) are selected randomly, and for 13 variation of b values (i.e., $b = 0.5$, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 2.0, 2.5), the perturbed prices are generated. Chow test is applied to test if the regression model with the original price and the regression model with the perturbed price are regarded as the same model. For each value of b, 1000 independently chosen samples are made and analyzed.

As a result, for each value of b, 1000 values of F value from Chow test are derived. Putting them in the order of magnitude, 5%, 10%, 50%, 90% and 95% points (i.e., the 50^{th}, 100^{th}, 500^{th}, 900^{th} and 950^{th} value) of F value are derived. Figure 3 demonstrate that the larger the value of b, the smaller the variation of F value.

From the aim to choose the appropriate value of b to generate properly perturbed property price, the minimum value of b, such that the null hypothesis

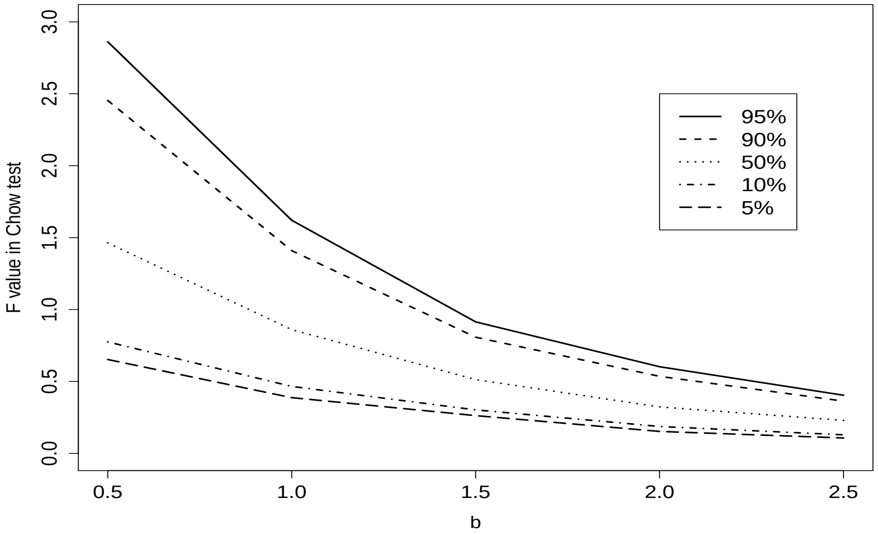

FIG 3. *The relation between F value and b*

of Chow test, H_0:"There is no statistically significant difference between two models", is not rejected, can be considered as the critical value for b. Note that the F value in F distribution with degrees of freedom 14 and 500 for significance level 0.05 is $F = 1.71$. Hence if F is smaller than 1.71, then the null hypothesis cannot be rejected, and therefore two models can be regarded statistically the same.

For each value of b, the percentage of F values among 1000 trials, which satisfies the acceptance condition that F value is less than 1.71, is calculated. In our numerical experiment, it is 65.0% for $b = 0.5$, 97.0% for $b = 1.0$, and 100% for $b \geq 1.4$, as are seen in Table 4.

3.3.4. Recommended standard for b value

In the numerical experiment described above, for the case that 20% is selected randomly, Chow tests testing the identity of two models, i.e., the regression model with the actual property price as the response variable and the regression model with the perturbed property price as the response variable, show that F value satisfies the acceptance condition with 97.0% probability when $b = 1.0$, and 100% for $b \geq 1.4$. By assuming that about 5% is the permissible level for rejection (i.e., that two models cannot be regarded as the same), $b = 1.0$ will be judged appropriate, satisfying that the perturbed price is perturbed enough and nonetheless that the regression model with the perturbed price can be regarded identical to the regression model with the original price.

Table 4

Percentage of sample that accepted the null hypothesis that the perturbed sample can be regarded statistically identical to the original sample for sample selection percentage, q, and b value

$b\backslash q$	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
0.5	0.591	0.576	0.650	0.728	0.827	0.918	0.969	0.992	0.996	1.000
0.6	0.686	0.670	0.744	0.808	0.903	0.947	0.988	0.997	0.999	1.000
0.7	0.729	0.764	0.808	0.873	0.943	0.976	0.994	1.000	0.999	1.000
0.8	0.815	0.845	0.858	0.928	0.966	0.988	0.996	1.000	1.000	1.000
0.9	0.867	0.867	0.931	0.966	0.989	1.000	0.998	1.000	1.000	1.000
1.0	0.930	0.925	0.970	0.987	0.995	0.998	1.000	1.000	1.000	1.000
1.1	0.960	0.968	0.983	0.998	0.997	1.000	1.000	1.000	1.000	1.000
1.2	0.978	0.987	0.994	0.999	0.999	1.000	1.000	1.000	1.000	1.000
1.3	0.994	0.994	0.995	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1.4	0.996	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1.5	0.999	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

For other cases of percentage in selecting the sample, appropriate value for b changes. To see this, let q be the percentage in selecting the sample. For example, in the above numerical experiment, $q = 0.2$ (20%). By assuming that 5% is the possible level for rejecting, the critical value of b, b_*, such that for b less than b_*, the probability for the rejecting becomes larger than 5%, is calculated by changing q. Table 4 summarizes the result. For all cases of q, that are experimented here, $b_* = 1.0$ looks a reasonable choice by balancing the similarity and the perturbation to the original price.

4. Conclusion

This paper proposed a new method to perturb a major variable by adding noise, so that the result of regression analysis is not affected by the noise addition. The extent of the perturbation can be controlled by a single parameter, b, which will ease the perturbation operation in the actual application. Moreover, $b = 1.0$ can be regarded as an appropriate value to achieve both the sufficient perturbation to mask the original values and the sufficient coherence between the perturbed data and the original data.

The proposed method only masks one major variable. But in the actual application, we may encounter many situations in which only one variable is critical to put the whole dataset in the public domain. Our method will be useful in these situations. There will be other possible uses of the perturbed data, and appropriateness of b value should be examined by testing more variety of data use cases. Admittedly, the application of the proposed method is limited, for other variables are assumed to be maintained as the original values. Thus another method to perturb the explanatory variables is necessary to broaden the application of this method. Such an extension will be made in the subsequent paper.

Acknowledgements

This work was partly supported by Research Grant from the Secom Science and Technology Foundation. This work of the first and third author was partly supported by Grant-in-Aid for Scientific Research from the JSPS (Japan Society for the Promotion of Sciences) (No. 25330035) and (No. 26590036), respectively.

References

Abowd, J. M. and Woodcock, S. D. (2001). Disclosure limitation in longitudinal linked data. *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies* 215–277.

Fuller, W. A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics* **9** 383–383.

Gouweleeuw, J. M., Kooiman, P., Willenborg, L. C. R. J. and de Wolf, P. P. (1998). Post randomisation for statistical disclosure control: Theory and implementation. *Journal of Official Statistics* **14** 463–478.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P. and Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician* **60** 224–232.

Manrique-Vallier, D. and Reiter, J. P. (2012). Estimating identification disclosure risk using mixed membership models. *Journal of the American Statistical Association* **107** 1385–1394.

Oganian, A. and Karr, A. F. (2011). Masking methods that preserve positivity constraints in microdata. *Journal of Statistical Planning and Inference* **141** 31–41.

Reiter, J. P. (2012). Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences. *Public opinion quarterly* **76** 163–181.

Shlomo, N. (2010). Releasing microdata: Disclosure risk estimation, data masking and assessing utility. *Journal of Privacy and Confidentiality* **2** 73–91.

Shlomo, N. and De Waal, T. (2008). Protection of Micro-data Subject to Edit Constraints Against Statistical Disclosure. *Journal of Official Statistics* **24** 229–253.

Takemura, A. (2003). Current trends in theoretical research of statistical disclosure control problem. *Proceedings of the Institute of Statistical Mathematics* **51** 252.

The Real Estate Transaction Modernization Center (1986). How to evaluate the price for residential area: Manual for appraising land price Technical Report, The Real Estate Transaction Modernization Center, Tokyo. Real estate properties distribution series, No.8.

Ting, D., Fienberg, S. E. and Trottini, M. (2008). Random orthogonal matrix masking methodology for microdata release. *International Journal of Information and Computer Security* **2** 86–105.

Willenborg, L. and Waal, T. (2001). Elements of statistical disclosure control.