DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 30, No. 4, 2014, pp. 839–857, http://dx.doi.org/10.2478/JOS-2014-0050

Data Smearing: An Approach to Disclosure Limitation for Tabular Data

Daniell Toth[1]

Statistical agencies often collect sensitive data for release to the public at aggregated levels in the form of tables. To protect confidential data, some cells are suppressed in the publicly released data. One problem with this method is that many cells of interest must be suppressed in order to protect a much smaller number of sensitive cells. Another problem is that the covariates used to aggregate and level of aggregation must be fixed before the data is released. Both of these restrictions can severely limit the utility of the data. We propose a new disclosure limitation method that replaces the full set of microdata with synthetic data for use in producing released data in tabular form. This synthetic data set is obtained by replacing each unit's values with a weighted average of sampled values from the surrounding area. The synthetic data is produced in a way to give asymptotically unbiased estimates for aggregate cells as the number of units in the cell increases. The method is applied to the U.S. Bureau of Labor Statistics Quarterly Census of Employment and Wages data, which is released to the public quarterly in tabular form and aggregated across varying scales of time, area, and economic sector.

Key words: Cell suppression; contingency tables; synthetic data; confidentiality; multiple imputation; nearest neighbor.

1. Introduction

Statistical agencies often collect data under a confidentiality agreement and are bound to protect the identity and/or the provided information of individual respondents. To accomplish this, a disclosure limitation method (DLM) is chosen to protect the sensitive data while allowing the provided data set to retain as much of the utility of the original data as possible. Because quantifying the level of protection and the utility a given DLM provides is difficult (Lambert 1993), comparing DLMs (and thus choosing a method) is not straightforward. Indeed, the level of protection offered by a DLM usually depends on characteristics of the data being published and is usually only quantified with certain restrictions on how the data can be accessed (see, for example Wasserman and Zhou 2010). Measures of the utility, on the other hand, often depend on the intended purpose of the data.

[1] Bureau of Labor Statistics, Office of Survey Methods Research, Suite 1950, Washington, DC 20212, U.S.A. Email: toth.daniell@bls.gov

Acknowledgments: The author would like to thank the editors and referees for their careful review of this article. Their helpful comments and suggestions have materially improved this article. I also thank Michail Sverchkov for many discussions and insights while developing the method, Michael Buso for his expert help with the QCEW data, Polly Phipps and John Eltinge for their helpful comments on the article and especially Randall Powers for his hard work during the evaluation of the method on the QCEW data set.

Sometimes, the sensitive information is collected with the intent to provide data only at certain aggregated levels in a way that still protects the sensitive data. For instance, income may be collected at the household level, while only the mean wages by geographic location such as state or county are reported; or an individual's opinion on a given topic is collected, but only percentages by gender and age category are reported.

When disseminating the data through published tables, cell suppression (CS) is one DLM that is often used by statistical agencies to protect the data of individual respondents. This method requires that cell entries deemed risky be withheld (usually because they represent only a few units or have estimates dominated by one or two large units). Protecting the privacy of responders using CS comes at the cost of withholding values of aggregated cells for which the data was intended to provide information. Often, this results in statistics for the gross aggregates being published, while more refined aggregates are suppressed. Depending on the sample size and level of refinement desired, this usually leads to tables with many holes, reducing the utility of the published data. In addition to the holes in the table resulting from the cell suppressions, CS requires an even larger number of secondary cell suppressions when the data is published as hierarchical contingency tables with more than one dimension.

Take, for example, complex data releases such as the Quarterly Census of Earnings and Wages (QCEW) published by the Bureau of Labor Statistics (BLS). The QCEW aims to provide time-series data with multiscale aggregations (by area and industry classifications). In order to protect against disclosure risks that arise from additive relationships within a table, additional (secondary) cell suppressions are required. Cox (1995) provides background on secondary cell suppressions and a solution to the problem of selecting these cells. Though these secondary suppressions are necessary, they further reduce the utility of the provided data. Over sixty percent of the possible QCEW table cells are suppressed.

Additionally, assuming that all of the risks of disclosure are accounted for through primary and secondary cell suppressions is problematic. For example, the BLS consistently applies both primary and secondary cell suppressions, yet additional risks still arise from the additive relationships in the table along with serial correlation. Holan et al. (2010) showed that it was possible to impute many of the suppressed values within one percent accuracy. Their approach takes advantage of the additive relationships of the QCEW tables (multiscale aggregations) and the serial correlation of the longitudinal data.

Another limitation of the CS method is that the cells defined by the published contingency tables must be fixed by the statistical agency in advance. This limits the potential utility of the data by preventing the release of different cells when other variables are available for further conditioning. For example, an agency may release tables of wages aggregated across only industry and occupation while area data is also available.

For these reasons, the BLS has considered replacing CS with another method for protecting the data (Yang et al. 2012). Any chosen DLM would have to protect the sensitive values (employment count and wages) while allowing for the publishing of total estimates for cells defined by industry, area, and ownership. The published estimates for the main cells with high-level of aggregation should be close to the true collected totals, while sufficiently protecting cells representing few establishments. Ideally, the new method would not require any cell suppressions and would allow estimates for user-defined cells. In addition, the employment and wage trends, which are very important for users of QCEW data, would be preserved by the estimates obtained from the new method.

One way to accomplish this might be to use a synthetic data approach on the microdata where the synthetic data is generated from random draws from a specified distribution. Using synthetic data to deal with disclosure limitation was proposed by Rubin (1993). Using a synthetic approach, the agencies can provide (or allow users to produce) any requested slice of the data, allowing them to produce any contingency table, without fear of disclosing confidential information.

Fully synthetic data approaches usually focus on trying to produce a data set with a distribution that matches the distribution of the observed microdata as closely as possible in order to allow valid inference while protecting sensitive information. A model is estimated using the sampled data and then values for the entire population (including sample values) are produced using draws from the estimated model distribution. The data obtained for either the entire population or for a sample from these random draws is released to the public (Reiter 2002; Reiter 2004; Reiter and Raghunathan 2007; and Graham et al. 2009).

Since the identity of units contained in the sample is generally unknown, the synthetic values could be as close to the true values as possible without risk of disclosure. However, the QCEW is a census of establishments, the location and identity of most establishments is already public knowledge, so the chosen DLM will have to protect the data without the benefit of anonymity. Because all population values are known, a model could be obtained that produces synthetic values very close to the true values, providing good utility, but not much protection. In addition, these synthetic approaches have the potential to impose associations between the data that do not exist, while reducing or eliminating legitimate associations (Graham et al. 2009). Eliminating this possibility or even the perception of this possibility is a major concern for the production of official statistics.

A related approach is to instead publish the true values with values masked by adding a random noise factor (Fuller 1993; Evans et al. 1998). A complication to applying this approach to establishment data is that the distribution of establishment wages and employment are extremely skewed, making it impossible to use the same noise factor for all establishments. Yang et al. (2012) determine that it is not possible to directly apply a standard noise model of Evans et al. (1998) to the QCEW data because of the inherent skewness of establishment data. In an attempt to modify the method, they propose three different noise factors (multiplicative as well as additive). Unlike the original method of additive noise, this new procedure results in biased marginal totals. To correct for this, a raking procedure is used to guarantee unbiased marginal totals. The cumulative effect of these adjustments on each value becomes unclear and could potentially result in removing the noise from some sets of establishments.

We propose a simple, more specialized DLM (data smearing), which is guaranteed to protect an individual's sensitive data by replacing it with an average value of surrounding units. This allows users to obtain aggregated estimates for any cell, which under a set of given conditions is shown to be asymptotically consistent. To accomplish this, the proposed method relies on a sampling scheme and a weighted estimator to divide the data for a sampled unit among its nearest neighbors. Essentially the method acts to "smear" the data of each unit around an “area” defined by a unit’s characteristics so that each individual unit’s data is replaced by data that represents an area’s average.

Advantages of this method include those of the synthetic approaches since all microdata values will be replaced, without the risk of disclosure or inducing nonexistent relationships among variables. However, the data released under this method no longer represents the microdata, but instead an average of the data of surrounding units, so the data no longer has the distribution of the original data but can be used only for providing statistics that are functions of totals.

The remainder of the article is organized as follows. Section 2 presents the proposed method and contains a discussion of some properties of the method. Section 3 contains results from an application of the method to QCEW microdata, while Section 4 includes a discussion of the results and mentions future areas of research.

2. The Proposed Disclosure Limitation Method

Suppose a data set consists of elements $\mathbf{u}_i = (\mathbf{Y}_i, \mathbf{X}_i)$ and is indexed by the set $U = \{i = 1. . .N\}$, where $\mathbf{Y}_i$ are the protected variables and $\mathbf{X}_i$ is a vector of unprotected auxiliary data. We assume these vectors include the data used to form cells of a table for release to the public. For example, the sensitive variables in the QCEW are the total employment and total wages paid, while the auxiliary information includes the establishment’s industry and geographic location.

In this article, we assume that $\mathbf{Y}_i$ contains the sensitive information of the i-th unit. Since QCEW represents a census of establishments, the establishment’s identity and inclusion in the sample must be considered known. Therefore a disclosure limitation procedure for the QCEW must account for this. This is in contrast to the situation of protecting sample data, where a disclosure limitation procedure can often exploit the fact that the identity of units that have been included in the sample is unknown. Therefore, sensitive sample data can be released for an individual unit as long as there are enough units in the population with similar characteristics to mask their identity or if characteristics are changed slightly.

2.1. Description of the Method

The first step in the procedure, is to define a metric $\|\cdot\|$ on the data elements which will determine the distance between each unit $d(\mathbf{u}_i, \mathbf{u}_j) = \|\mathbf{u}_i - \mathbf{u}_j\|$. We use this distance function to find the k-nearest neighbors for each element in the population. In case of ties, we include a small real-valued noise variable to be used in the distance function. These neighbors define the units the method will use to select a sample in order to produce an average value. A neighborhood is found for each unit in the population. Neighborhoods are defined so that the units contained within them are likely to be included in the aggregate cells to be produced from the estimates. For example, the metric may include geographic location and industry when applying the procedure to business surveys.

Dummy variables are used to handle categorical variables like industry and political borders by assigning “penalties” to units not in the same category. For instance, one could add a ν-mile “penalty” to the geographic distance between units that are not in the same state. That is, if $state_i$ is the state in which unit i is located and $geo(\mathbf{u}_i, \mathbf{u}_j)$ is the geographic distance between units i and j in miles, then the distance between units i and j defined by the metric is

$$d(\mathbf{u}_i, \mathbf{u}_j) = geo(\mathbf{u}_i, \mathbf{u}_j) + \nu \mathbb{1}_{\{state_i \neq state_j\}}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function and $\nu \in [0, \infty)$. A value of $\nu < \infty$ would allow the "smearing" over categories while $\nu = \infty$ would require that neighbors be in the same category.

The next step of the procedure is to find the k-nearest neighbors for each unit. That is, for each $i \in U$, define r_i as the smallest real number such that the set

$$\{j \neq i \in U \mid \|\mathbf{u}_j - \mathbf{u}_i\| \leq r_i\}$$

has k elements. Note that r_i exists for every $i \in U$, as long as $k \leq N$, and we assume that in most practical situations $k \ll N$. We define $K(i)$ as the k-nearest neighborhood of unit i,

$$K(i) = \{j \neq i \in U \mid \|\mathbf{u}_j - \mathbf{u}_i\| \leq r_i\}.$$

To make sure that the data for each element is spread out among enough other units, we extend the k-nearest neighborhood $K(i)$ to be the k-network, $\overline{K(i)}$, defined by also including every unit j for which unit i is a k-nearest neighbor. Formally,

$$\overline{K(i)} = K(i) \cup \{j \mid i \in K(j)\}.$$

Figure 1 illustrates why extending the network could be necessary for some establishments. In this example, units j, k, and l are in $K(i)$, but i is not in $K(j)$, $K(k)$, or $K(l)$. In fact, there are no units in the population shown that contain unit i in their k-nearest neighborhood. The completed network ensures that the information of unit i gets represented in the other synthetic units produced by the method.

For each $i \in U$, draw a random sample of size $n \leq k$ from unit $i's$ network, $\overline{K(i)}$. For example, for the applications of the method in this article, we used a simple random sample without replacement (SRSWOR). Let $\delta_j(i) = 1$ if unit j is selected in the sample from $\overline{K(i)}$ which will be used to protect element i, and 0 otherwise. To produce a fully synthetic data set, we replace $\mathbf{Y}_i$ for each unit $i \in U$ with the weighted average

$$\tilde{\mathbf{Y}}_i = w_i \mathbf{Y}_i + \sum_{j \in \overline{K(i)}} w_j \delta_j(i) \mathbf{Y}_j, \qquad (1)$$

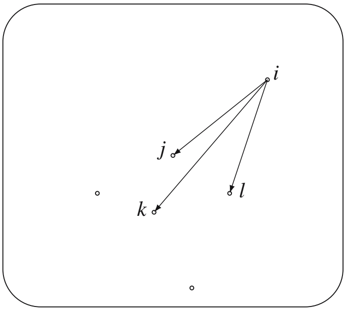

Fig. 1. Illustration shows K(i), the k-nearest neighbor of unit i in the population, where k = 3.

for a given fixed set of weights $\{w_i \mid i \in U\}$. The properties of the method depend on the choice of weights. We will present a choice of weights which are shown to produce asymptotically consistent estimates for cells satisfying certain conditions.

Note that the sampling is done to give an extra level of protection by not allowing users to guess the members included in the unit's network; however if the network is defined large enough, this would not be necessary. Instead, defining the synthetic value as a weighted average of every unit in the network would remove this uncertainty from the synthetic value. Another way to remove some of this uncertainty is to produce a number ($m > 1$) of synthetic values for each unit i ($\tilde{\mathbf{Y}}_{i1}, \tilde{\mathbf{Y}}_{i2}, \ldots, \tilde{\mathbf{Y}}_{im}$) independently and use the average of these as the synthetic value

$$\tilde{\mathbf{Y}}_i = \frac{(\tilde{\mathbf{Y}}_{i1} + \tilde{\mathbf{Y}}_{i2} + \ldots + \tilde{\mathbf{Y}}_{im})}{m}. \tag{2}$$

Alternatively, the multiple sets of imputed values can be released directly to allow the user to estimate the variance of any estimates produced using the synthetic value. By releasing multiply-imputed synthetic values for each establishment, the agency would be giving a data user some indication of the reliability of each estimate being produced.

In addition, a referee pointed out that the variance of the synthetic values could be controlled by defining $\tilde{\mathbf{Y}}_i = \alpha Y_i + (1 - \alpha)\bar{Y}_i$. This would give the data providers another option to provide more accurate estimates, and all the consistency properties described below hold for synthetic values defined this way or by Equations (1) or (2). However, α would have to be chosen carefully to balance the added utility with a loss of protection.

2.2. *Properties of the Method*

The required aggregated data for the released tables are produced using these new synthetic values. The differences in the values and the properties of the synthetic data that is produced depend on the distance function defined by the agency and the moment structure of $\mathbf{Y}_i$ in the k-nearest networks. Therefore, the protection that is afforded the individual units depends on the distance function and the set of networks it produces. The protection also depends on the other parameters of the method, including the value of k, the size of the sample n selected from the k-networks, and the set of weights.

Now we define the notation used to investigate some of the theoretical properties of the synthetic data produced by the proposed method. First, define any subset of units $C \subseteq U$ to be a closed area if it is equal to

$$\bar{C} = \underset{i \in C}{\cup} \overline{K(i)}. \tag{3}$$

The circle in Figure 2 displays a hypothetical user-defined area that is an example of a closed area. Note that given any subset C, there exists a closed area that contains C. We will use $\bar{C}$ to denote the smallest of these.

Let $|\overline{K(i)}|$ denote the total number of units in $\overline{K(i)}$. The following property of the method states that if we define the weights in Equation (1) correctly, then the cells in the table will be unbiased for large enough levels of aggregation. That is, the expected value of the aggregated value of a cell produced from the synthetic values will be equal to the aggregated value of the cell using the original microdata, if the cell defines a closed network using the given distance function. The original microdata are considered fixed values and the expectation is with respect to the random samples from the k-networks.

The number of times a given unit's value is used to produce different synthetic values depends on the size of the unit's network and the probability of selection used in the sampling process. With this in mind, we show in the next result, that if we select a weight for each unit that is the inverse of the expected number of times the unit's value will appear in other synthetic values, then we will get unbiased cell totals for those cells defined by closed networks.

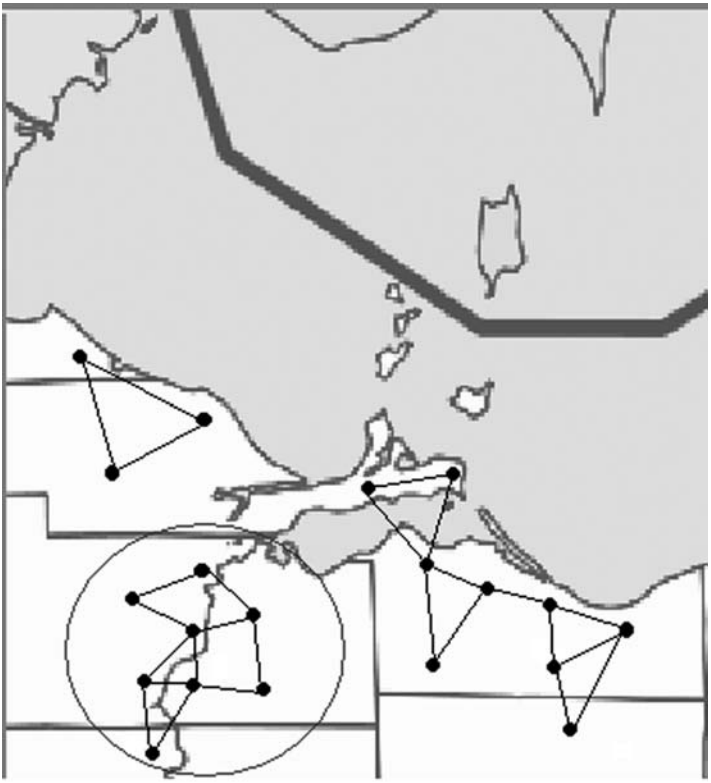

Fig. 2. Illustration with hypothetical establishments in a given location and the k-nearest network that would result from these establishments if k = 2. The circle representing a selected area is an example of a closed area. Every establishment in the closed k-network is included in the selected area.

Lemma 2.1 *If a cell C is a closed area and the weights, w_i used in Equation (1) are defined as*

$$w_i = \left(1 + n \sum_{j \in \overline{K(i)}} \frac{1}{|\overline{K(j)}|}\right)^{-1}, \tag{4}$$

then the synthetic data produced from the method satisfies

$$E\left[\sum_{i\in C}\tilde{\mathbf{Y}}_i\right]=\sum_{i\in C}\boldsymbol{Y}_i.$$

Proof

From Equation (1)

$$\sum_{i\in C}\tilde{\mathbf{Y}}_i=\sum_{i\in C}w_i\mathbf{Y}_i+\sum_{i\in C}\sum_{j\in\overline{K(i)}}\delta_j(i)w_j\mathbf{Y}_j. \tag{5}$$

Note that for any i and j if $j\in\overline{K(i)}$, then $i\in\overline{K(j)}$, by the definition of a nearest network. Also, since C is a closed area, if $i\in C$, then for all $j\in\overline{K(i)}, j\in C$. Therefore, we can re-write the sum in Equation (5) as

$$\sum_{i\in C}\tilde{\mathbf{Y}}_i=\sum_{i\in C}w_i\mathbf{Y}_i+\sum_{i\in C}\sum_{j\in\overline{K(i)}}\delta_i(j)w_i\mathbf{Y}_i=\sum_{i\in C}\left(1+\sum_{j\in\overline{K(i)}}\delta_i(j)\right)w_i\mathbf{Y}_i.$$

Since each sample from the k-networks is drawn using a SRSWOR,

$$E[\delta_i(j)]=n|\overline{K(j)}|^{-1},$$

so the expectation of Equation (5) is

$$\sum_{i\in C}\left(1+\sum_{j\in\overline{K(i)}}E[\delta_j(i)]\right)w_i\mathbf{Y}_i=\sum_{i\in C}\left(1+n\sum_{j\in\overline{K(i)}}\frac{1}{|\overline{K(j)}|}\right)w_i\mathbf{Y}_i.$$

The proof follows by substituting Equation (4) for w_i.

□

Note, that if the neighborhood of unit i does not contain any units that have an extended neighborhood, then $\forall j\in\overline{K(i)},\ |\overline{K(j)}|=k$. This means that the weight for unit i given by Equation (4) is simply $1/(n+1)$.

The result of Lemma 2.1 applies only to table cells that are closed areas. In general we can expect that many cells of interest will not necessary be closed areas. The next result states that we can still expect to obtain reasonable estimates for any area as long as most of the data of the area being estimated are contained in a closed area.

Define the boundary of area C as the set $\partial C=\overline{C}-C$. This is the set of elements that contribute information to the estimate of area C, but are not located in the area. Property 2.1 states that the ratio of the area total estimated using the synthetic data over the total estimated using the real data asymptotically goes to one as long as the data in the interior of the area of interest increases sufficiently fast compared to the data in the boundary.

Property 2.1 Assume $|Y_i-E[\tilde{Y}_i]|<M<\infty$ *for all* i. *If* $|\partial(C)|=o(\sum_{i\in C}\boldsymbol{Y}_i)$ *and that the weights are defined by Equation (4), then the synthetic data produced from the* *method satisfies*

$$\lim_{|C|\to\infty}\left(\sum_{i\in C}\boldsymbol{Y}_i\right)^{-1}E\left[\sum_{i\in C}\tilde{\boldsymbol{Y}}_i\right]=1.$$

Proof By the definition of the boundary and Lemma 2.1

$$\begin{aligned}E\left[\sum_{i\in C}\tilde{\mathbf{Y}}_i\right]&=E\left[\sum_{i\in\overline{C}}\tilde{\mathbf{Y}}_i\right]-E\left[\sum_{i\in\partial C}\tilde{\mathbf{Y}}_i\right]\\&=\sum_{i\in\overline{C}}\mathbf{Y}_i-E\left[\sum_{i\in\partial C}\tilde{\mathbf{Y}}_i\right]=\sum_{i\in C}\mathbf{Y}_i+\sum_{i\in\partial C}\mathbf{Y}_i-E\left[\sum_{i\in\partial C}\tilde{\mathbf{Y}}_i\right]\\&=\sum_{i\in C}\mathbf{Y}_i+\sum_{i\in\partial C}\left(\mathbf{Y}_i-E[\tilde{\mathbf{Y}}_i]\right).\end{aligned}$$

Since

$$\sum_{i\in\partial C}\left(\mathbf{Y}_i-E[\tilde{\mathbf{Y}}_i]\right)\leq M|\partial C|$$

we can divide by $\sum_{i\in C}\mathbf{Y}_i$ to get the result.

□

Figures 3 and 4 give examples of two different user-defined areas. Figure 3 is an illustration of an area that is likely to satisfy the condition $|\partial(C)|=o\left(\sum_{i\in C}\mathbf{Y}_i\right)$. On the other hand, the area shown in Figure 4 has a boundary that would likely grow faster than the contained area as the area expands. The difference is that the first area is a sphere in the coordinates used to define the metric whereas the second area is a very elongated shape with respect to those coordinates.

3. Application to QCEW Data

The Bureau of Labor Statistics (BLS) Quarterly Census of Employment and Wages (QCEW) program aims to publish a near census of wage and employment data for every industry at the national, state, county and metropolitan statistical area (MSA) levels. Industry is defined by the establishment's assigned six-digit code from the North American Industrial Classification Systems (NAICS). The codes are organized hierarchically, where higher digit codes aggregate to fewer digit codes. For instance, the three-digit industry codes 423 (merchant wholesales, durable goods), 424 (merchant wholesalers, nondurable goods), and 425 (electronic wholesale markets) aggregate to the two-digit industry code 42 (wholesale trade).

The QCEW collects the number of employees on the payroll of an establishment each month and the total payroll of an establishment every quarter. Every quarter, QCEW publishes employment and wage data in tabular form aggregated across varying cells defined by these location and industry categories. Less aggregated-level data can only be published if disclosure restrictions are met. Currently, over 60% of the possible cells are

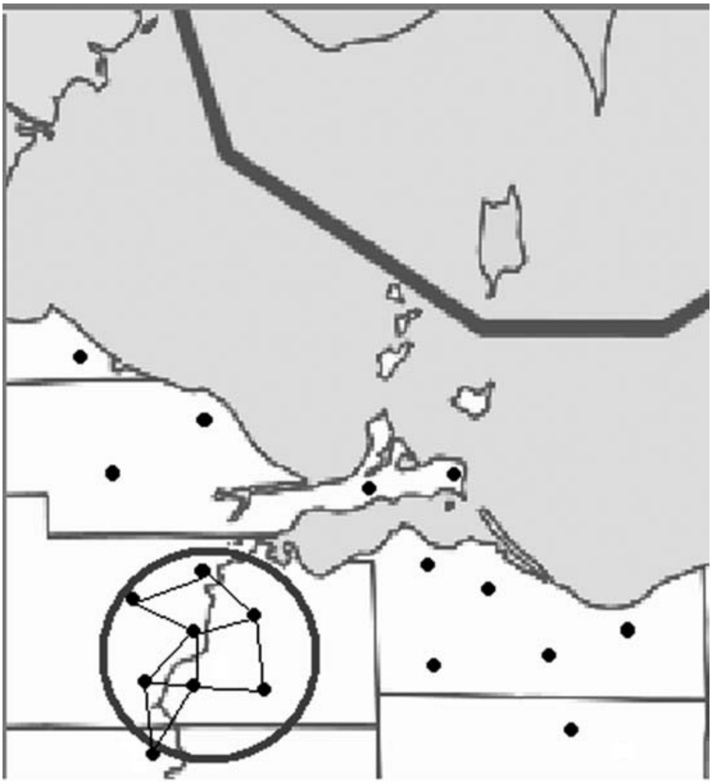

Fig. 3. The circle is an example of a selected area that is not closed but likely to satisfy the conditions of Property 2.1. Though there is one establishment in the closed k-network that has been excluded from the selected area, the number of establishments from the closed k-network that are included in the selected area are likely to dominate the total estimate for the selected area.

suppressed as a result of the current use of CS as the DLM for QCEW. In addition, requested aggregate estimates for areas not published cannot easily be accommodated under CS without risk of disclosure. Using the proposed data smearing DLM, all currently produced cells as well as any requested cells could be published with varying degrees of accuracy without risk of disclosure.

Table 1 shows an example table of one month employment totals (second month of the quarter) for four quarters of QCEW employment data. The table was produced for one industry comprised of three sub-industries over a given MSA. The original QCEW table (top) was produced using the original data for the given MSA for the three industries and their aggregates. The table represents data for roughly 80, 2, and 58 establishments, respectively, for the three industries each quarter.

The same table (bottom) was produced using synthetic values obtained from the data-smearing method with parameter values of $k = 3$, $n = 3$, $m = 5$. The method provided synthetic data that produced a table with values close to the original (all within 1% of the true values) for cells represented by more than two establishments and for the aggregate series and the annual totals. Unsurprisingly, the cells that differ the most are for the middle sub-series, which of course are composed of the smallest number of establishments. This row would be suppressed under the CS method currently used

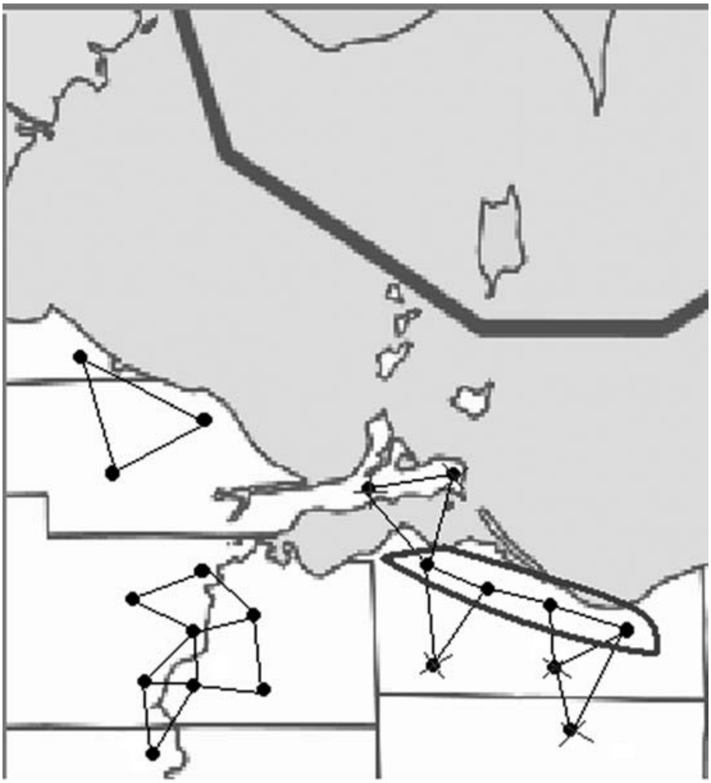

Fig. 4. The selected area is an example of an area that is not closed and unlikely to satisfy the conditions of Property 2.1 because the number of units in the network located outside the selected area is larger than the number of units contained in the area. The values from the establishments in the closed k-network that have been excluded from the selected area are likely to be at least of equal magnitude to the values from the establishments that are included in the selected area. Therefore, the estimated total for the selected area could be biased using the synthetic data.

by the BLS. In addition, another row (probably row three) would be suppressed as the secondary suppression.

The metric used to produce Table 1 used longitude and latitude of each establishment to find the geographical distance $geo(\cdot)$ between establishments and the six-digit industry classification code,

$$d(\mathbf{u}_i, \mathbf{u}_j) = geo(\mathbf{u}_i, \mathbf{u}_j) + \nu \mathbb{1}_{\{ind6_i \neq ind6_j\}}, \tag{6}$$

where $\nu = \infty$, and $ind6_i$ is the six-digit industry code for establishment i. Defining the metric in this way, we are forcing the algorithm to pair establishments with the same industry classification in close geographic proximity to one another. This could also be achieved by applying the algorithm to industries with the same six-digit industry code separately and using only the geographical distance between establishments.

Next we illustrate the method by applying it to one month of QCEW employment data for all (non-government-owned) establishments, over all industries, across the entire country. Again we use parameter values of $k = 3$, $n = 3$, $m = 5$, and the metric given by Equation (6). The weights are defined by (4). As we mentioned earlier, the statistical agency could produce multiple ($m > 1$) synthetic data sets for publication, but the results

Table 1. Example of a 2010 QCEW employment table for one MSA for establishments in a given industry code composed of three sub-series. The first table (Top) was produced using the true values while the second table (Bottom) used the synthetic values. Totals for each of quarter-1 through quarter-4 are displayed for the series and each sub-series along with the annual totals. For this MSA, the table is based on data from roughly 180, 2, and 58 establishments in the three industrial sub-series, sub1, sub2, and sub3, respectively

Industry	qrtr-1	qrtr-2	qrtr-3	qrtr-4	a-total
Series 1	2,600	2,899	3,022	2,599	11,120
Sub1	1,981	2,256	2,382	1,957	8,576
Sub2	32	33	37	33	135
Sub3	587	610	603	609	2,409

Industry	qrtr-1	qrtr-2	qrtr-3	qrtr-4	a-total
Series 1	2,622	2,929	3,062	2,589	11,202
Sub1	1,989	2,271	2,420	1,947	8,627
Sub2	42	38	40	34	154
Sub3	591	620	602	608	2,421

below are focused on one synthetic data set using Equation (2), the average of the five independent draws. A comparison of the true and the synthetic values presented in Figure 5 shows that synthetic values produced are highly correlated to the true values. The very small values are inflated while larger values tend to be decreased by the method.

The data-smearing approach acts like a synthetic approach to disclosure limitation in the sense that it replaces each value at the microlevel with a synthetic value. Unlike many synthetic data approaches to disclosure limitation, this current method does not attempt to match the distribution of the synthetic data to that of the original data. Because we are replacing individual values with the mean value of a surrounding area, extreme values are replaced by values closer to the middle of the distribution. Though the two distributions are similar, this figure illustrates the tendency of the method to shift the true values toward the mean. For instance, the new synthetic distribution has a smaller proportion of units with the smallest value. As an example, Figure 6 displays the distribution of total employment

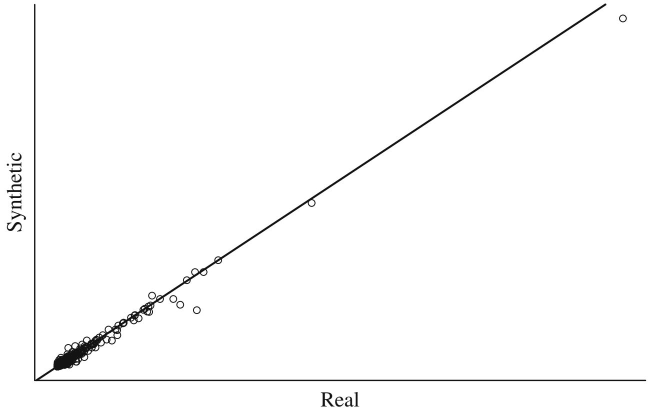

Fig. 5. Relationship between true values (x-axis) and synthetic values (y-axis) with the line $y = x$.

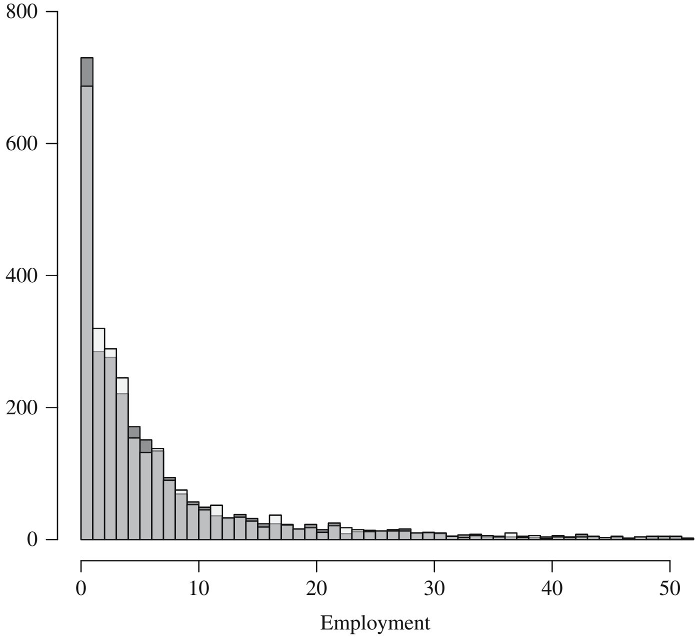

Fig. 6. Histogram of true (dark grey) and synthetic (light grey) employment values for establishments in a given industry and state.

for all establishments in the population in a specific industrial classification. As the figure shows, the synthetic values do not have the same distribution as the real values. Therefore, relying on individual microlevel data for statistical analysis would be very problematic. This is as it should be, since we are protecting the individual values.

The choice of parameters, k, n, and m affects the level of protection as well as how closely the synthetic values represent the real values. Larger k values will smear the value of a given establishment over a wider area. The value of n, and more particularly n/k will affect the variance between each of the m imputed values, with larger values giving smaller variances. By using a large value for m (which we recommend) and the estimator defined by Equation (2), this variance (and the protection derived from the sampling) could be virtually eliminated. The data of individual units would still be protected as long as $n \geq 2$. A value of $n \geq 2$ ensures that the data published for an individual establishment will be the average of at least two other units. In our evaluations of the method, we found that when using even moderate values of m, varying parameter values had a relatively small impact on the overall estimates compared to changing the metric. We proceed by investigating the impact of the metric on the method.

Using the synthetic values obtained from the smearing method, we computed aggregated employment counts e_j for every two, four, and six-digit industry level. The metric given by Equation (6) was designed to give accurate answers for all industrial classifications, so we would expect estimates of total employment aggregated by industry classifications to be close to the true estimates. For each cell estimate produced, we calculate the percent relative difference (PRD) $100 * (\tilde{e}_i - e_i)/e_i$ between the synthetic value $\tilde{e}_i$ and the true value e_i.

Figure 7 displays boxplots of the PRD for each cell estimate over different quantiles of cell sizes, where size is the number of establishments. The top graph gives the results for the two-digit industry level aggregates, the middle graph shows the four-digit level and the bottom graph the result for the six-digit level. As expected, the estimates produced using the synthetic values are all close to the true cell totals. The cells aggregated to the two-digit industry level are within 0.5% of the true value for all 24 cells. This is not surprising

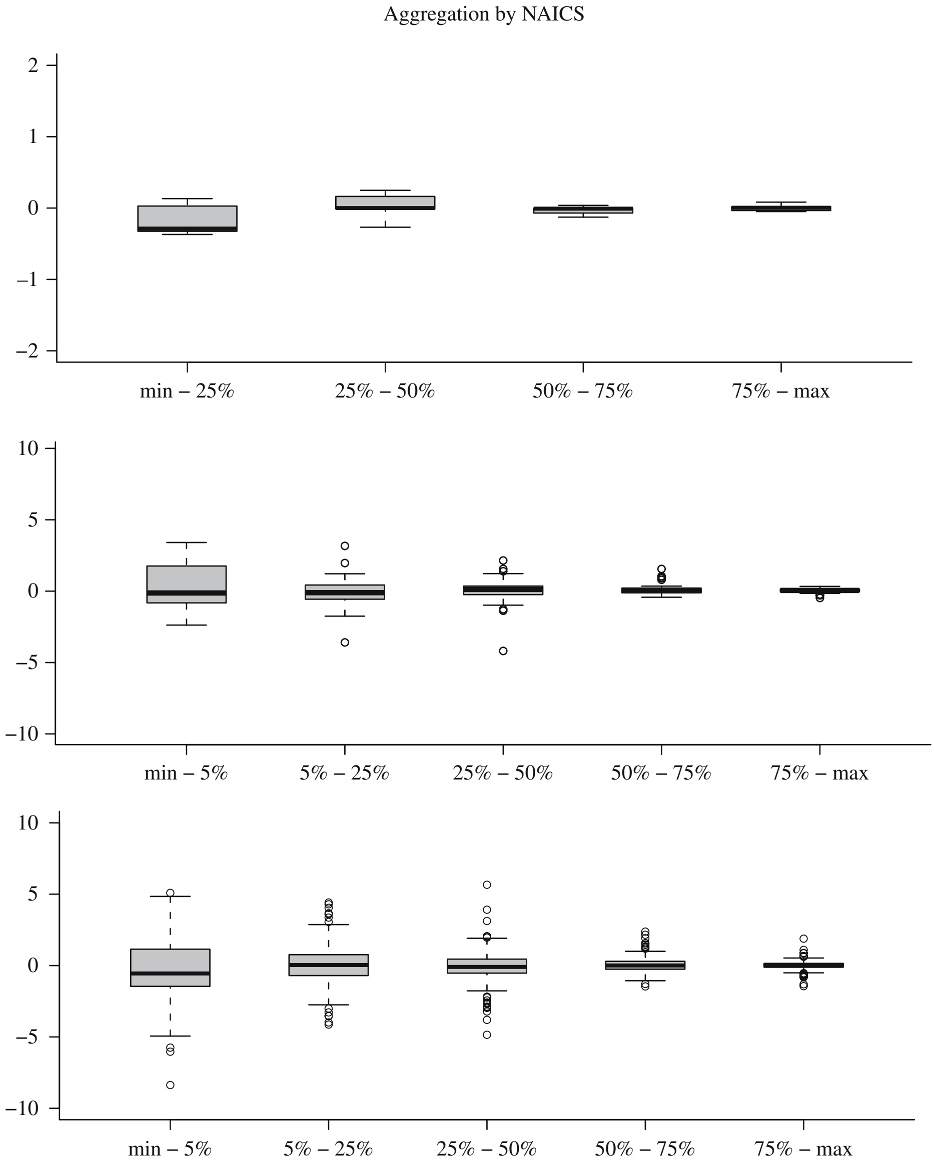

Fig. 7. Boxplots of the percent relative differences for the synthetic values of industry totals compared to the true values. The boxplots are given for different quantiles of area size (number of establishments representing the industry total). The top graph is of quantiles of errors for all the two-digit industry code totals, the middle graph is of quantiles of errors for all four-digit industry code totals, and the bottom graph is for all six-digit totals. All graphs represent percent relative errors for the synthetic values using the metric given by Equation (6).

since the smallest cell size is 14,652 at this level of aggregation. However, the method produced estimates close to the true values, even at the six-digit industry level of aggregation. Indeed, 99% of the cells have less than a 4.5% difference. This is despite the fact that the smallest 1% of the cells have fewer than 22 establishments.

The situation is very different however, when we consider cells aggregated by state and industry. The top graph in Figure 8 shows the same boxplots of the PRD by quantile of cell size for cells aggregated by state to the two-digit industry level. The cell estimates are not nearly as good for state estimates, even at this high level of aggregation. Though a handful of these cells have fewer than five establishments, 99% have more than 18 so the method should be expected to produce reasonable estimates for most of these cells.

Because state was not included in the metric, there is no penalty for choosing neighbors that are across a state border. Therefore, it should be expected that many synthetic data values represent averages of establishments over more than one state, which biases state-level estimates. However, after adding state to the metric (6), the state-level estimates were still not very good, even though the penalty was rather high (100 miles) for being in a different state.

This is because of the hard restriction that the establishments in a neighborhood must all be in the same industry to the six-digit level. At the six-digit level of aggregation more than 18.1% of state industry cells have fewer than four establishments. This means that the method is forced to use at least one establishment from another state to produce the synthetic values for each of these cells (no matter how large the penalty). This biases the estimates of both states.

Instead we replace the metric given by (6) with

$$d(\mathbf{u}_i, \mathbf{u}_j) = geo(\mathbf{u}_i, \mathbf{u}_j) + \nu_1 \mathbb{1}_{\{ind4_i \neq ind4_j\}} + \nu_2 \mathbb{1}_{\{ind5_i \neq ind5_j\}} + \nu_3 \mathbb{1}_{\{ind6_i \neq ind6_j\}} + \nu_4 \mathbb{1}_{\{state_i \neq state_j\}}, \qquad (7)$$

where $(\nu_1, \nu_2, \nu_3, \nu_4) = (\infty, 50, 10, 100)$, and *indt* is the *t*th-digit industry code. This new metric replaces the hard restriction that all industries match to the six-digit level with one at the four-digit level. In addition, there are 50 and ten-mile penalties for not matching at the five and six-digit industry levels respectively. There is a 100-mile penalty for being in a different state.

The percentage of state industry cells with less than 4 establishments drops from 18.1% at the six-digit level to 7.4% at the four-digit level. Therefore, we would expect the values given by the method using the metric (7) to continue to give accurate cell estimates aggregated to the four-digit industry level, while giving improved state-level industry estimates. The bottom graph in Figure 8 shows the same boxplots for cells aggregated by state to the two-digit industry level as the top graph, but instead using this new metric. This shows that the cell estimates for the state two-digit industry level are indeed improved; 95% of all the estimates are within 4% of the true value. There are still a number of estimates that are significantly off, but this is to be expected given that there are a number of small cells for some states even at the two-digit industry level.

Figure 9 gives the results for the same two, four and six-digit industry level aggregates as Figure 7 for the new metric given by Equation (7). The results show that the estimates for the two and four-digit industry level aggregates remain close to the

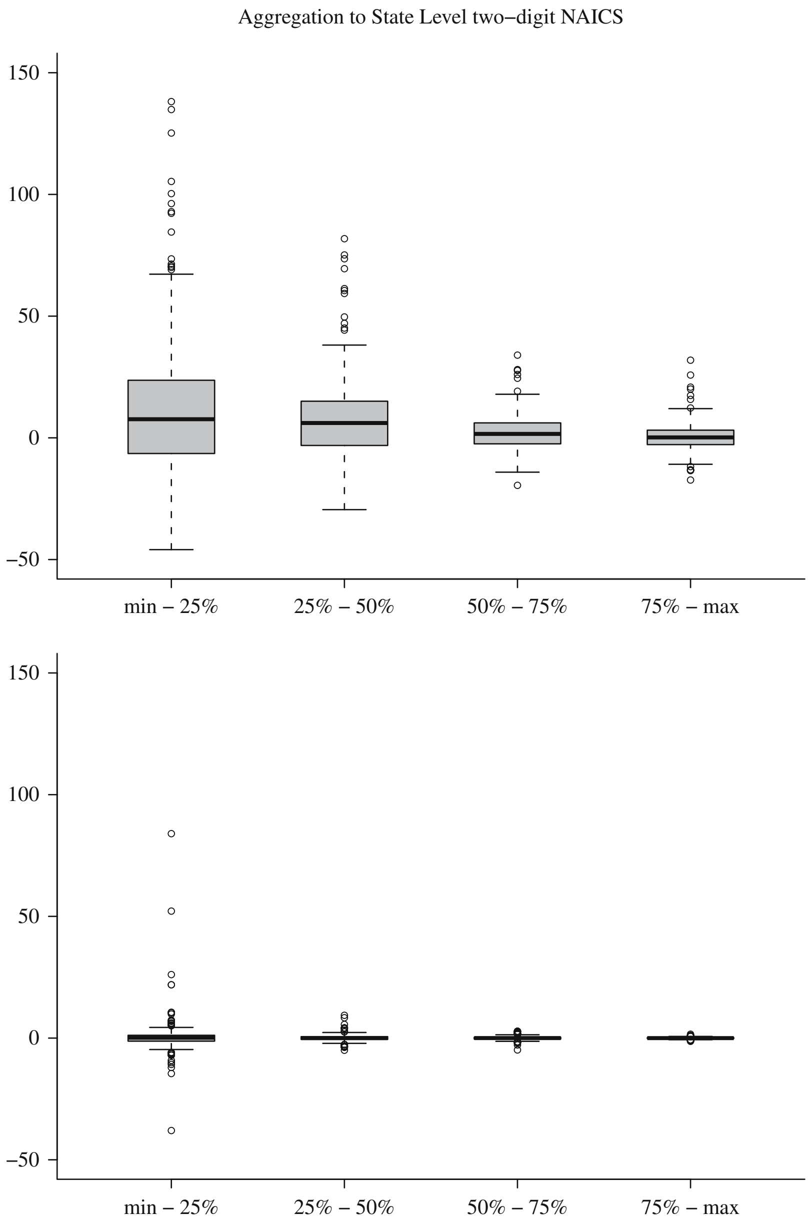

Fig. 8. Boxplots of the percent relative differences for the synthetic values of two-digit industry totals by state compared to the true values. The boxplots are given for different quantiles of area size (number of establishments representing the two-digit industry state total). The top graph is of quantiles of percent relative errors for the synthetic values using the metric given by Equation (6) while the bottom graph uses metric given by Equation (7).

true values under this new measure. However, as we would expect, since the penalty for not matching industry code at the six-digit level is small, many of the estimates at the six-digit industry level are no longer accurate. This demonstrates that the data provider would only be able to give assurances for marginals being controlled for by the metric. However, as long as the interior of the cells being estimated were large

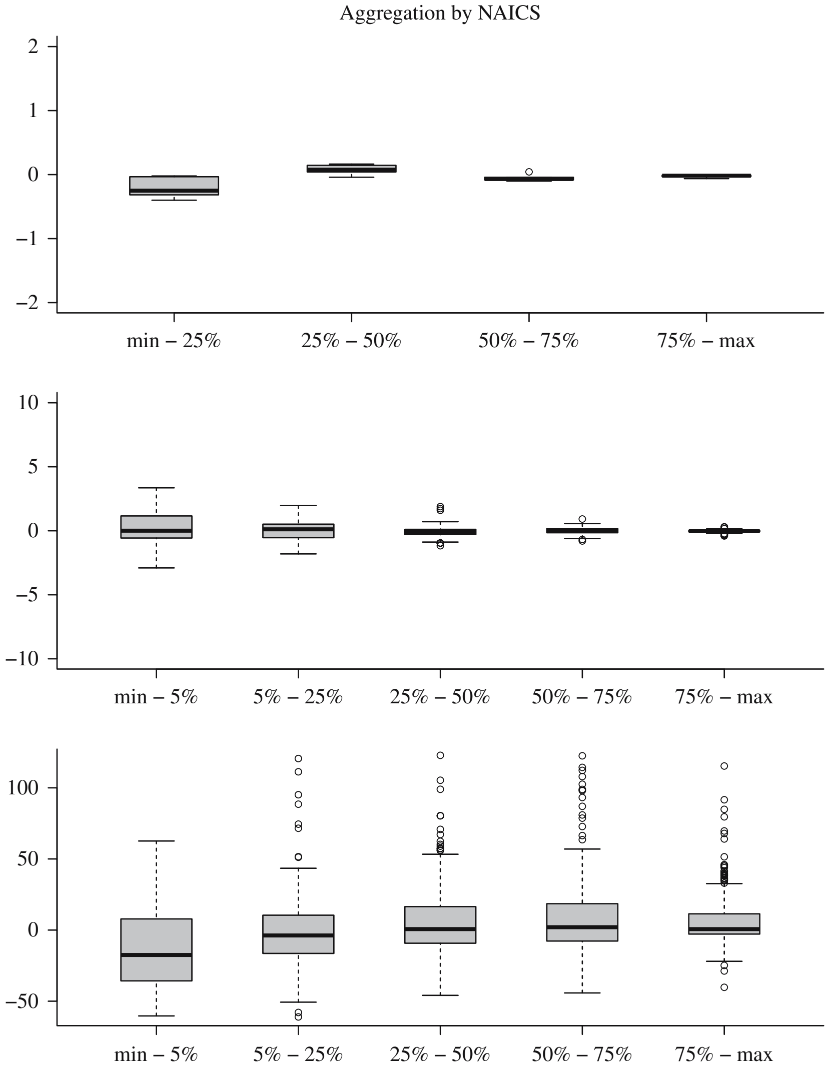

Fig. 9. Boxplots of the percent relative differences for the synthetic values of industry totals compared to the true values. The boxplots are given for different quantiles of area size (number of establishments representing the industry total). The top graph is of quantiles of errors for all the two-digit industry code totals, the middle graph is of quantiles of errors for all four-digit industry code totals, and the bottom graph is for all six-digit totals. All graphs represent percent relative errors for the synthetic values using the metric given by Equation (7).

compared to its boundary, the estimates produced should be increasingly accurate the larger the cell, as stated in Property 2.1.

4. Discussion

We have introduced a new disclosure limitation method, "data smearing." The method is intended to allow the release of a synthetic data set that can be used to produce accurate contingency tables while protecting the data of individual units. Though the method was demonstrated on census data in this article, the method will work equally well for sample data.

Unlike other synthetic data approaches, the method focuses on producing accurate contingency tables rather than trying to match the distribution of the original data. The released data for each unit has the intuitive interpretation of representing the average value for the units in the surrounding neighborhood. Neighborhoods are defined by the metric chosen by the agency releasing the data and can be shared with the data users. Importantly, the tables can be user defined after the data set has been released. Additionally, the data from each unit is guaranteed to be protected in that the value assigned to every unit is the average value of at least $n + 1$ units.

We demonstrate the method using QCEW employment data using two different metrics. One metric is defined to connect units within the same six-digit industrial classification that are in close geographical proximity. The second metric tries to connect units within the same state that are in close geographical proximity and have matching industrial classification codes to at least the first four digits. The relative performance of the two metrics shows that the accuracy of a contingency table produced using the synthetic data from this method is highly dependent on the variables included in the metric.

The proposed DLM has performed well during the initial testing on the QCEW data set. It has been shown to produce accurate aggregated cell estimates on cells for which the metric was designed. However, this article attempts only to introduce the method. There is much further testing to be done and properties of the method yet to be explored as well as a number of possible extensions of the method. These and other questions are sure to be the subject of future research.

5. References

Cox, L. 1995. "Network Models for Complementary Cell Suppression." *Journal of the American Statistical Association* 90: 1453–1462.

Evans, T., L. Zayatz, and J. Slanta. 1998. "Using Noise for Disclosure Limitation of Establishment Tabular Data." *Journal of Official Statistics* 14: 537–551.

Fuller, W. 1993. "Masking Procedures for Microdata Disclosure Limitation." *Journal of Official Statistics* 9: 383–406.

Graham, P., J. Young, and R. Penny. 2009. "Multiply Imputed Synthetic Data: Evaluation of Hierarchical Bayesian Imputation Models." *Journal of Official Statistics* 25: 245–268.

Holan, S., D. Toth, M. Ferreira, and A. Karr. 2010. "Bayesian Multiscale Multiple Imputation With Implications for Data Confidentiality." *Journal of the American Statistical Association* 105: 564–577.

Lambert, D. 1993. "Measures of Disclosure Risk and Harm." *Journal of Official Statistics* 9: 313–331.

Reiter, J. 2002. "Satisfying Disclosure Restrictions with Synthetic Data Sets." *Journal of Official Statistics* 18: 531–544.

Reiter, J. 2004. "New approaches to data dissemination: A glimpse into the future (?)." *Chance* 17: 12–16.

Reiter, J. and T. Raghunathan. 2007. “The Multiple Adaptations of Multiple Imputation.” *Journal of the American Statistical Association* 102: 1462–1471.

Rubin, D. 1993. “Discussion: Statistical disclosure limitation.” *Journal of Official Statistics* 9: 462–468.

Wasserman, L. and S. Zhou. 2010. “A statistical framework for differential privacy.” *Journal of the American Statistical Association* 105: 375–389.

Yang, M., S. Pramanik, A. Mushtaq, F. Scheuren, M. Buso, S. Butani, and D. Hiles. 2012. “Evaluation of Three Disclosure Limitation Models for the QCEW Program.” In Proceedings, Joint Statistical Meeting, American Statistical Association. San Diego, July 28–August 2. 4217–4229.

Received December 2012
Revised September 2014
Accepted September 2014