LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Matthias Schmid & Hans Schneeweiss

The Effect of Microaggregation by Individual Ranking on the Estimation of Moments

Technical Report Number 025, 2008
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

The Effect of Microaggregation by Individual Ranking on the Estimation of Moments

Matthias Schmid[*] Hans Schneeweiss[†]

Abstract

Microaggregation by individual ranking (IR) is an important technique for masking confidential data. While being a successful method for controlling the disclosure risk of observations, IR is also known for its favorable property of having a relatively small effect on the results of statistical analyses. In this paper we conduct a detailed theoretical analysis on the estimation of arbitrary moments from a data set that has been anonymized by means of the IR method. We show that classical moment estimators remain both consistent and asymptotically normal under relatively weak assumptions. This theory provides the justification for applying standard statistical estimation techniques to the anonymized data without having to correct for a possible bias caused by anonymization.

Keywords: consistent estimation, disclosure control, individual ranking, microaggregation, general moments.

1 Introduction

Confidential data that have been collected by a statistical office are usually anonymized before publication. Anonymization is accomplished by making use of statistical disclosure control techniques. These techniques result in a reduction of the information content of the data and thus in a low re-identification risk of the observations in the published data set. A drawback

[*] Institut für Medizininformatik, Biometrie und Epidemiologie, Friedrich-Alexander-Universität Erlangen-Nürnberg, Waldstraße 6, D–91054 Erlangen, Germany. Email: matthias.schmid@imbe.imed.uni-erlangen.de

[†] Institut für Statistik, Ludwig-Maximilians-Universität München, Ludwigstraße 33, D–80539 München, Germany. Email: hans.schneeweiss@stat.uni-muenchen.de

of disclosure control techniques is that the reduction of the information content often leads to an efficiency loss and/or to biased statistical analysis (Willenborg and de Waal 2001, Doyle et al. 2001, Domingo-Ferrer and Torra 2004, Ronning et al. 2005, Aggarwal and Yu 2008). Due to confidentiality requirements, a certain amount of efficiency loss cannot be avoided. However, if the efficiency loss is not too large, data users will still benefit from the published data. In order to control the efficiency loss arising from the anonymization of data sets, the effect of statistical disclosure control techniques on statistical analysis has to be carefully examined.

In this paper the focus is on the effect of microaggregation by individual ranking (IR) on the estimation of general moments and, by implication, on the least squares (LS) estimation of a linear model in transformed variables. IR, which has been introduced by Defays and Anwar (1998), is an important statistical disclosure control technique for continuous microdata. The idea of IR is to anonymize each continuous variable in a data set one after another by forming small groups (usually of size 3 or 5) of "similar" data values and by replacing the original data values with the respective group means. It is thus hoped that the multivariate distribution of a data set is approximately preserved. Although several authors have argued that the re-identification risk resulting from the application of IR remains relatively high (Domingo-Ferrer and Torra 2001, Domingo-Ferrer et al. 2002 , Winkler 2002), IR has nevertheless been shown to be a successful technique for anonymizing confidential data sets (see Ronning et al. 2005, who recommended IR for use in official statistics in Germany).

In two previous papers (Schmid 2006, Schmid and Schneeweiss 2008) we have analyzed the effect of IR on the estimation of linear models. In Schmid (2006) it was shown analytically that a linear model can be consistently estimated from the microaggregated data by standard LS estimation techniques. In addition, if the continuous variables in a data set are assumed to follow a mixed normal distribution each, the efficiency loss due to IR is asymptotically zero. These results provide the justification of the application of least squares techniques to a data set whose continuous variables have been anonymized by means of the IR technique. In Schmid and Schneeweiss (2008) we have extended this theory by considering linear models in transformed variables, where nonlinear variable transformations are applied to the data *after* microaggregation. We have shown that even in this case the LS estimators of a linear model remain consistent under mild regularity assumptions.

It should be pointed out that the consistency results derived for transformed data (Schmid and Schneeweiss 2008) do not automatically follow from the results for untransformed data (Schmid 2006). This is because nonlinear transformations of microaggregated data introduce an additional (finite sample) bias in the LS estimators. For instance, the empirical mean of three logarithmized data values is usually different from the logarithmized mean of the three values.

The purpose of this paper is to provide a generalization of the theory presented in Schmid (2006) and Schmid and Schneeweiss (2008) to the estimation of *arbitrary* moments based on transformed and untransformed microaggregated data. The variables involved need not be continuous variables as in Schmid and Schneeweiss (2008), so the consistency proof has to be adapted to this more general case. In addition, arbitrary multivariate moments are considered and not only product moments as in Schmid and Schneeweiss (2008). We will not only prove the consistency of the empirical moments computed from microaggregated data but will also specify conditions and regularity assumptions under which the moments are asymptotically normal. Arbitrary moments include first, second, and product moments of the transformed and untransformed data as special cases. Thus, the consistency results for linear models presented in Schmid (2006) and Schmid and Schneeweiss (2008) are confirmed. Moreover, since the consistent estimation of arbitrary moments from the microaggregated data is guaranteed, any method-of-moments estimator is in turn consistent if computed from the microaggregated data. It should be noted that these results (obtained for the IR method) are fundamentally different from previous results obtained for other microaggregation techniques, such as multivariate microaggregation with a sorting variable (Mateo-Sanz and Domingo-Ferrer 1998, Domingo-Ferrer and Torra 2001). In the latter case, moment estimators have been shown to be asymptotically biased, see Schmid et al. (2007).

The paper is organized as follows: In Section 2 we give an example of the IR method and illustrate the problems arising from nonlinear transformations of the microaggregated data. In Section 3 the consistency of the empirical moments computed from transformed micraggregated data is proved. Section 4 deals with the asymptotic normality of these estimators. Section 5 contains a simulation study and some examples on the theoretical results derived in Sections 3 and 4. A summary of the results presented in this paper is given in Section 6.

2 Microaggregation by individual ranking

Microaggregation by individual ranking works as follows: First, a fixed group size K is chosen. Next, the data set is sorted by the first continuous variable, and groups of K consecutive observations are formed. The values of the first continuous variable in each group are replaced by their corresponding group means, while the values of the other variables in the data set are left unchanged. Then the same procedure is repeated for the second continuous variable, and so on. If the number of observations n is not a multiple of K, it is common practice to alter the procedure such that the groups around the medians contain $K + \mathrm{mod}\,(n/K)$ adjacent data values (see Domingo-Ferrer et al. 2002). If there are discrete variables in the data set, they are left unchanged during the IR procedure. It is generally considered necessary to form groups of at least $K = 3$ observations, as data attackers can easily identify an observation in a group of less than 3 observations if they have sufficient background knowledge on only *one* of the observations. In practice, it is common to form groups of sizes 3 or 5.

As an example of IR we consider a data set consisting of two vectors x and y, both containing continuous data. In addition, we consider a "dummy" vector z containing the values of a discrete binary variable. Assume that the original data set is given by

x	2	4	7	0	9	5	1	8	3
y	4	2	0	9	1	5	6	11	10
z	1	0	1	0	1	1	1	1	1

.

The first step of IR results in the sorted data set

x	0	1	2	3	4	5	7	8	9
y	9	6	4	10	2	5	0	11	1
z	0	1	1	1	0	1	1	1	1

,

where the rows of the original data set have been ordered according to the values of x. In the second step of IR, with K chosen to be 3, the values of x are microaggregated:

$\tilde{x}$	1	1	1	4	4	4	8	8	8
y	9	6	4	10	2	5	0	11	1
z	0	1	1	1	0	1	1	1	1

.

The third step of IR results in the sorted data set

$\tilde{x}$	8	8	4	1	4	1	1	4	8
y	0	1	2	4	5	6	9	10	11
z	1	1	0	1	1	1	0	1	1

,

where the rows have been ordered according to the values of y. Finally, in the fourth step of IR, again with K chosen to be 3, the values of y are microaggregated:

$\tilde{x}$	8	8	4	1	4	1	1	4	8
$\tilde{y}$	1	1	1	5	5	5	10	10	10
z	1	1	0	1	1	1	0	1	1

.

Now suppose that $\tilde{y}$ is additionally transformed by means of a quadratic transformation. Then

$$\tilde{y}^2 = (1, 1, 1, 25, 25, 25, 100, 100, 100) \ .$$

Obviously, taking the squares of the microaggregated values of y results in a different data set than when the squared values of y are microaggregated. In the latter case, one would have obtained

$$\widetilde{y^2} = (1.67, 1.67, 1.67, 25.67, 25.67, 25.67, 100.67, 100.67, 100.67) \ .$$

Now consider the estimation of a theoretical moment, i.e., the expectation of an arbitrary one-dimensional function of the random variables (X, Y, Z) from the microaggregated data. Since the original data have been altered by IR, the consistent estimation of the theoretical moment by its corresponding ordinary empirical moment is not guaranteed any more. In the next sections we will address this problem.

3 Consistent estimation of moments

Let X be a real random variable and let x_i, $i = 1, \dots, n$, be an i.i.d. sample taken from the distribution of X. The corresponding individually microaggregated data are denoted by $\tilde{x}_i$, $i = 1, \dots, n$, where the group size for the aggregation of X is denoted by K. For simplicity we always assume n to be a multiple of K (this assumption does not affect the asymptotic results derived in the following). We want to prove that the usual consistent estimator of the moments of the distribution of X remains consistent if we replace the original data by their microaggregated data values.

Let us consider very general moments: Suppose that the expectation $\mathrm{E}(h(X))$ for some measurable function h exists. We know that, given an i.i.d. sample $(x_1, \dots, x_n)$, $\mathrm{E}(h(X))$ can be consistently estimated by the empirical mean $\frac{1}{n}\sum_{i=1}^{n} h(x_i)$. Can we estimate $\mathrm{E}(h(X))$ also via the microaggregated sample $(\tilde{x}_1, \dots, \tilde{x}_n)$ in the same way? The next theorem gives an answer.

Theorem 1. *Let X be a real-valued random variable and $(x_1, \dots, x_n)$ an i.i.d. sample from the distribution of X. Let $(\tilde{x}_1, \dots, \tilde{x}_n)$ be the corresponding microaggregated sample with fixed aggregation group size K, assuming (w.l.o.g.) n to be a multiple of K. Let h be a continuously differentiable function with domain $\mathcal{D} = (d_l, d_u)$ (which is a finite or infinite open interval in $\mathbb{R}$). The support of X is assumed to be contained in $\mathcal{D}$. Suppose $|h(X)|$ is monotone (increasing or decreasing) for $d_l < x < c_l$ and for $c_u < x < d_u$ with some $c_l < c_u$. If $\mathrm{E}(h(X))$ exists, then a. s.*

$$\mathrm{lim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} h(\tilde{x}_i) = \mathrm{E}(h(X)) \ . \tag{1}$$

Proof. Let $\mathcal{B} = [b_l, b_u] \subset \mathcal{D}$ be a closed finite interval such that $(c_l, c_u) \subset \mathcal{B}$ and $\mathrm{E}[|h(X)|\mathrm{I}_{\bar{\mathcal{B}}}(X)] < \epsilon$ for some preassigned $\epsilon > 0$. This is possible because $\mathrm{E}(h(X))$ exists. Let $B = \{i : x_i \in \mathcal{B}\}$ and let $G_i = \{j :$ x_j and x_i belong to the same microaggregation group$\}$. We will prove that

$$\lim_{n\to\infty}\left|\frac{1}{n}\sum_{i=1}^{n} h(\tilde{x}_i) - \frac{1}{n}\sum_{i=1}^{n} h(x_i)\right| = 0 \ \text{ a. s.} \tag{2}$$

Obviously,

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n} h(\tilde{x}_i) - \frac{1}{n}\sum_{i=1}^{n} h(x_i)\right| \quad &\leq \quad \left|\frac{1}{n}\sum_{i:G_i\subset B} h(\tilde{x}_i) - \frac{1}{n}\sum_{i:G_i\subset B} h(x_i)\right| \qquad (3)\\ &+\left|\frac{1}{n}\sum_{i:G_i\subset \bar{B}} h(\tilde{x}_i) - \frac{1}{n}\sum_{i:G_i\subset \bar{B}} h(x_i)\right|\\ &+\left|\frac{1}{n}\sum_{\substack{i:G_i\not\subset B\\ G_i\not\subset \bar{B}}} h(\tilde{x}_i) - \frac{1}{n}\sum_{\substack{i:G_i\not\subset B\\ G_i\not\subset \bar{B}}} h(x_i)\right| \ .\end{aligned}$$

Each of the three terms on the r. h. s. of (3) can be made less than a fixed multiple of ϵ for sufficiently large n.

First consider the first term. By a Taylor series expansion,

$$h(x_i) = h(\tilde{x}_i) + h'(x_i^*)(x_i - \tilde{x}_i) \ , \qquad (4)$$

where $x_i^* = t_i x_i + (1-t_i)\tilde{x}_i$, $t_i \in (0,1)$. Note that $x_i^* \in \mathcal{B}$ because $G_i \subset B$. Therefore, with $H' = \max_{x\in\mathcal{B}} |h'(x)|$ (which exists because $\mathcal{B}$ is closed and h' is continuous),

$$\begin{aligned}\left|\frac{1}{n}\sum_{i:G_i\subset B} h(\tilde{x}_i) - \frac{1}{n}\sum_{i:G_i\subset B} h(x_i)\right| \quad &\leq \quad \frac{1}{n}\sum_{i:G_i\subset B} |h'(x_i^*)||x_i - \tilde{x}_i|\\ &\leq \quad \frac{1}{n}H' \sum_{i:G_i\subset B} |x_i - \tilde{x}_i|\\ &\leq \quad \frac{1}{n}H' \sum_{i:G_i\subset B} ||G_i||\\ &\leq \quad \frac{1}{n}H'K||\mathcal{B}|| \ , \qquad (5)\end{aligned}$$

where $||G_i||$ is the range of the x_j belonging to group G_i and $||\mathcal{B}|| = b_u - b_l$ is the length of the interval $\mathcal{B}$. The last inequality follows because there are K elements in each G_i. The last term converges to 0 as $n \to \infty$, so that with probability 1

$$\left|\frac{1}{n}\sum_{i:G_i\subset B} h(\tilde{x}_i) - \frac{1}{n}\sum_{i:G_i\subset B} h(x_i)\right| < \epsilon \qquad (6)$$

for sufficiently large n.

Next consider the second term on the r. h. s. of (3). First note that

$$\left| \frac{1}{n} \sum_{i:G_i \subset \bar{B}} h(\tilde{x}_i) - \frac{1}{n} \sum_{i:G_i \subset \bar{B}} h(x_i) \right| \leq \frac{1}{n} \sum_{i:G_i \subset \bar{B}} |h(\tilde{x}_i)| + \frac{1}{n} \sum_{i:G_i \subset \bar{B}} |h(x_i)| \ . \tag{7}$$

As $\tilde{x}_i$ is the arithmetic mean of the x_j with $j \in G_i$, clearly $\min_{j \in G_i}\{x_j\} \leq \tilde{x}_i \leq \max_{j \in G_i}\{x_j\}$. Because of the monotonicity of $|h(x)|$ in $\bar{\mathcal{B}}$, it follows that

$$|h(\tilde{x}_i)| \leq \max_{j \in G_i} |h(x_j)| \leq \sum_{j \in G_i} |h(x_j)| \ . \tag{8}$$

Therefore, since each G_i has K elements,

$$\sum_{i:G_i \subset \bar{B}} |h(\tilde{x}_i)| \leq K \sum_{i:G_i \subset \bar{B}} |h(x_i)| \leq K \sum_{i \in \bar{B}} |h(x_i)| \ . \tag{9}$$

Also,

$$\sum_{i:G_i \subset \bar{B}} |h(x_i)| \leq \sum_{i \in \bar{B}} |h(x_i)| \tag{10}$$

and hence

$$\frac{1}{n} \sum_{i:G_i \subset \bar{B}} |h(\tilde{x}_i)| + \frac{1}{n} \sum_{i:G_i \subset \bar{B}} |h(x_i)| \ \leq \ (K+1) \frac{1}{n} \sum_{i \in \bar{B}} |h(x_i)| \ . \tag{11}$$

By the Strong Law of Large Numbers, the last term converges a. s. to $(K+1)\mathrm{E}[|h(X)\mathrm{I}_{\bar{\mathcal{B}}}(X)|]$, which, by assumption, is less than $(K+1)\epsilon$. Thus, for sufficiently large n, a. s.

$$\left| \frac{1}{n} \sum_{i:G_i \subset \bar{B}} h(\tilde{x}_i) - \frac{1}{n} \sum_{i:G_i \subset \bar{B}} h(x_i) \right| < (K+1)\, \epsilon \tag{12}$$

with probability 1.

Finally consider the last term on the r. h. s. of (3). This term comprises the sample points of only two microaggregation groups, one on each side of the interval $\mathcal{B}$ (on rare occasions, there may be only one group such that some x_i in this group lie to the left of b_l and some to the right of b_u, but such a group can be treated in a similar way as the other ones). These two groups (which could also be empty) are in effect negligible. Nevertheless, we will treat them in detail. It suffices to consider the group surrounding b_u. Let G_u be the corresponding set of indices such that some x_i with $i \in G_u$ lie in $\mathcal{B}$, while some other x_i with $i \in G_u$ lie to the right of b_u. Since at most $K-1$ sample points x_i lie in $\mathcal{B}$, we have

$$\frac{1}{n}\sum_{i \in G_u} |h(x_i)| \leq \frac{K-1}{n} H + \frac{1}{n}\sum_{i \in \bar{B}} |h(x_i)| < \epsilon \tag{13}$$

a. s. for sufficiently large n, where $H = \max_{x \in \mathcal{B}} |h(x)|$. Denote the mean of the x_i, $i \in G_u$, by $\tilde{x}_u$. There are three cases now: *Case 1:* $\tilde{x}_u \in \mathcal{B}$. Then $|h(\tilde{x}_u)| \leq H$. *Case 2:* $\tilde{x}_u > b_u$ and $|h(x)|$ monotone decreasing for $x > b_u$. Then $h(\tilde{x}_u) \leq h(b_u)$. *Case 3:* $\tilde{x}_u > b_u$ and $|h(x)|$ monotone increasing for $x > b_u$. Then $|h(\tilde{x}_u)| \leq \sum_{i \in \bar{B}} |h(x_i)|$. In all three cases,

$$\frac{1}{n}\sum_{i \in G_u} |h(\tilde{x}_i)| = \frac{K}{n} |h(\tilde{x}_u)| < \epsilon \tag{14}$$

a. s. for n sufficiently large. The same arguments hold true for the group on the left side of $\mathcal{B}$, and so

$$\left| \frac{1}{n} \sum_{\substack{i:G_i \not\subset B \\ G_i \not\subset \bar{B}}} h(\tilde{x}_i) - \frac{1}{n} \sum_{\substack{i:G_i \not\subset B \\ G_i \not\subset \bar{B}}} h(x_i) \right| < 4\,\epsilon\ . \tag{15}$$

Inequalities (3), (6), (12), and (15) imply that

$$\left| \frac{1}{n}\sum_{i=1}^{n} h(\tilde{x}_i) - \frac{1}{n}\sum_{i=1}^{n} h(x_i) \right| < (5 + (K+1))\,\epsilon \tag{16}$$

a. s. for n sufficiently large. □

Example 1: Theorem 1 applies to ordinary moments of X and moments of transformed variables $h(X)$, such as $\log(X)$ or X^λ, $\lambda \in \mathbb{R}^+$.

Remark 1: Theorem 1 also holds if we only suppose that $|h(x)| \leq h^*(x)$, where $h^*(x)$ has the monotonicity property required in Theorem 1, and that $\mathrm{E}(h^*(X))$ exists. Thus, equation 1 also holds for functions such as

$h(x) = \sin(x)$, which do not satisfy the monotonicity condition required in Theorem 1.

We next study a pair of random variables (X, Y) and a general bivariate moment given by some (measurable) function $h(X, Y)$. We assume that $\mathrm{E}(h(X, Y))$ exists. We give conditions under which $\mathrm{E}(h(X, Y))$ can be consistently estimated by the empirical mean $\frac{1}{n}\sum_{i=1}^{n} h(\tilde{x}_i, \tilde{y}_i)$ constructed from the individually microaggregated data $(\tilde{x}_i, \tilde{y}_i)$, $i = 1, \ldots, n$. The following conditions will be sufficient for our purpose.

- $\mathcal{H}1$: $h(x, y)$ is defined on an open rectangle $\mathcal{D} = \mathcal{D}_x \times \mathcal{D}_y$, where $\mathcal{D}_x$ and $\mathcal{D}_y$ are finite or infinite open intervals on the real line. The support of (X, Y) is contained in $\mathcal{D}$.
- $\mathcal{H}2$: $h(x, y)$ has continuous partial derivatives on $\mathcal{D}$.
- $\mathcal{H}3$: There exist non-negative continuously differentiable functions h_x and h_y defined on $\mathcal{D}_x$ and $\mathcal{D}_y$, respectively, such that $|h(x, y)| \leq h_x(x) + h_y(y)$ for all $(x, y) \in \mathcal{D} \cap \bar{\mathcal{C}}$, where $\mathcal{C} = \mathcal{C}_x \times \mathcal{C}_y$ is a closed finite rectangle contained in $\mathcal{D}$.
- $\mathcal{H}4$: $\mathrm{E}(h_x^2(X)) < \infty$ and $\mathrm{E}(h_y^2(Y)) < \infty$.
- $\mathcal{H}5$: h_x is monotone on each side of $\bar{\mathcal{C}}_x$ and h_y is monotone on each side of $\bar{\mathcal{C}}_y$.

Theorem 2. *Let (X, Y) be a pair of real random variables and $(x_i, y_i), i = 1, \ldots, n$, an i.i.d. sample from the distribution of (X, Y). Let $(\tilde{x}_i, \tilde{y}_i), i = 1, \ldots, n$, be the corresponding individually microaggregated sample with fixed group size K for aggregating x and fixed group size L for aggregating y. Let $h(x, y)$ satisfy the conditions $\mathcal{H}1$ to $\mathcal{H}5$. Then a. s.*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h(\tilde{x}_i, \tilde{y}_i) = \mathrm{E}(h(X, Y)) \ . \tag{17}$$

Proof. Let $\mathcal{B}_x$ and $\mathcal{B}_y$ be closed finite intervals such that $\mathcal{C}_x \subset \mathcal{B}_x \subset \mathcal{D}_x$ and $\mathcal{C}_y \subset \mathcal{B}_y \subset \mathcal{D}_y$, and let $\mathcal{B} = \mathcal{B}_x \times \mathcal{B}_y$. Choose $\mathcal{B}_x$ such that

$$\mathrm{E}[h_x(X)\mathrm{I}_{\bar{\mathcal{B}}_x}(X)] < \epsilon \ , \tag{18}$$

$$\mathrm{E}[h_x^2(X)\mathrm{I}_{\bar{\mathcal{B}}_x}(X)] < \epsilon \ , \tag{19}$$

$$\mathrm{P}(X \in \bar{\mathcal{B}}_x) < \epsilon \tag{20}$$

for some preassigned $\epsilon > 0$. Choose $\mathcal{B}_y$ analogously. Let $B_x = \{i : x_i \in \mathcal{B}_x\}$, $B_y = \{i : y_i \in \mathcal{B}_y\}$, $B = B_x \cap B_y = \{i : (x_i, y_i) \in \mathcal{B}\}$, and $G_i = \{j : x_j$ and x_i belong to the same microaggregation group for $x\}$. Similarly, let $H_i = \{j : y_j$ and y_i belong to the same microaggregation group for $y\}$. We will prove that

$$\lim_{n\to\infty} \left| \frac{1}{n}\sum_{i=1}^{n} h(\tilde{x}_i, \tilde{y}_i) - \frac{1}{n}\sum_{i=1}^{n} h(x_i, y_i) \right| = 0 \quad \text{a. s.} \tag{21}$$

Now

$$\left| \frac{1}{n}\sum_{i=1}^{n} h(\tilde{x}_i, \tilde{y}_i) - \frac{1}{n}\sum_{i=1}^{n} h(x_i, y_i) \right| \leq S_1 + S_2 + S_3 + S_4 + S_5 \ , \tag{22}$$

where

$$\begin{align}
S_1 &= \frac{1}{n} \sum_{\substack{i:G_i \subset B_x \\ H_i \subset B_y}} |h(\tilde{x}_i, \tilde{y}_i) - h(x_i, y_i)| \ , \tag{23}\\
S_2 &= \frac{1}{n} \sum_{i:G_i \subset \bar{B}_x} |h(\tilde{x}_i, \tilde{y}_i) - h(x_i, y_i)| \ , \tag{24}\\
S_3 &= \frac{1}{n} \sum_{i:H_i \subset \bar{B}_y} |h(\tilde{x}_i, \tilde{y}_i) - h(x_i, y_i)| \ , \tag{25}\\
S_4 &= \frac{1}{n} \sum_{\substack{i:G_i \not\subset B_x \\ G_i \not\subset \bar{B}_x}} |h(\tilde{x}_i, \tilde{y}_i) - h(x_i, y_i)| \ , \tag{26}\\
S_5 &= \frac{1}{n} \sum_{\substack{i:H_i \not\subset B_y \\ H_i \not\subset \bar{B}_y}} |h(\tilde{x}_i, \tilde{y}_i) - h(x_i, y_i)| \ . \tag{27}\\
& \tag{28}
\end{align}$$

We start with S_1. By a Taylor series expansion,

$$h(x_i, y_i) = h(\tilde{x}_i, \tilde{y}_i) + \frac{\partial}{\partial x} h(x_i^*, y_i^*)(x_i - \tilde{x}_i) + \frac{\partial}{\partial y} h(x_i^*, y_i^*)(y_i - \tilde{y}_i) \ , \tag{29}$$

where $(x_i^*, y_i^*) = t_i \cdot (x_i, y_i) + (1 - t_i) \cdot (\tilde{x}_i, \tilde{y}_i)$, $t_i \in (0, 1)$, and thus $(x_i^*, y_i^*) \in \mathcal{B}$.

Let H' be an upper bound for $\frac{\partial}{\partial x}h(x,y)$ and $\frac{\partial}{\partial y}h(x,y)$, $(x,y) \in \mathcal{B}$. Then

$$\begin{aligned} S_1 &\leq \frac{1}{n}H' \sum_{\substack{i:G_i \subset B_x \\ H_i \subset B_y}} \Big(|x_i - \tilde{x}_i| + |y_i - \tilde{y}_i| \Big) \\ &\leq \frac{1}{n}H' \left(\sum_{i:G_i \subset B_x} |x_i - \tilde{x}_i| + \sum_{i:H_i \subset B_y} |y_i - \tilde{y}_i| \right) \\ &\leq \frac{1}{n}H' \left(\sum_{i:G_i \subset B_x} ||G_i|| + \sum_{i:H_i \subset B_y} ||H_i|| \right) \\ &\leq \frac{1}{n}H' \Big(K||\mathcal{B}_x|| + L||\mathcal{B}_y|| \Big) \ , \end{aligned} \tag{30}$$

where $||G_i||$ and $||\mathcal{B}_x||$ are defined as in the proof of Theorem 1 and $||H_i||$ and $||\mathcal{B}_y||$ are defined analogously. It follows that

$$S_1 < \epsilon \tag{31}$$

for sufficiently large n. Next consider S_2. First,

$$\begin{aligned} S_2 &\leq \frac{1}{n} \sum_{i:G_i \subset \bar{B}_x} |h(\tilde{x}_i, \tilde{y}_i)| + \frac{1}{n} \sum_{i:G_i \subset \bar{B}_x} |h(x_i, y_i)| \\ &\leq \frac{1}{n} \sum_{i:G_i \subset \bar{B}_x} (h_x(\tilde{x}_i) + h_y(\tilde{y}_i)) + \frac{1}{n} \sum_{i:G_i \subset \bar{B}_x} (h_x(x_i) + h_y(y_i)) \ . \end{aligned} \tag{32}$$

Now, because of the monotonicity of h_x on $\bar{\mathcal{B}}_x$, just as in the proof of Theorem 1,

$$\frac{1}{n} \sum_{i:G_i \subset \bar{B}_x} h_x(\tilde{x}_i) \leq \frac{K}{n} \sum_{i \in \bar{B}_x} h_x(x_i) < K\epsilon \tag{33}$$

a. s., because $\frac{1}{n}\sum_{i \in \bar{B}_x} h_x(x_i)$ converges a. s. to $\mathrm{E}[h_x(X)\mathrm{I}_{\bar{\mathcal{B}}_x}(X)]$, which is smaller than ϵ. Next,

$$\begin{aligned} \frac{1}{n} \sum_{i:G_i \subset \bar{B}_x} h_y(\tilde{y}_i) &\leq \frac{1}{n} \sum_{i \in \bar{B}_x} h_y(\tilde{y}_i) = \frac{1}{n} \sum_{i=1}^{n} h_y(\tilde{y}_i)\mathrm{I}_{\bar{\mathcal{B}}_x}(x_i) \\ &\leq \sqrt{\frac{1}{n} \sum_{i=1}^{n} h_y^2(\tilde{y}_i)} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \mathrm{I}_{\bar{\mathcal{B}}_x}(x_i)} < \sqrt{\mathrm{E}(h_y^2(Y))}\,\epsilon \end{aligned} \tag{34}$$

a. s. for n sufficiently large. The last inequality of (34) holds because a. s., due to Theorem 1 (with h_y^2 in place of h),

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h_y^2(\tilde{y}_i) = \mathrm{E}(h_y^2(Y)) \tag{35}$$

and $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}_{\bar{\mathcal{B}}_x}(x_i) = \mathrm{P}(X \in \bar{\mathcal{B}}_x)$, which is smaller than ϵ by assumption. Thus, by (33) and (34),

$$\frac{1}{n} \sum_{i:G_i \subset \bar{B}_x} (h_x(\tilde{x}_i) + h_y(\tilde{y}_i)) < \left(K + \sqrt{\mathrm{E}(h_y^2(Y))}\right) \epsilon \ . \tag{36}$$

Setting $K = 1$ it follows that

$$\frac{1}{n} \sum_{i:G_i \subset \bar{B}_x} (h_x(x_i) + h_y(y_i)) \leq \left(1 + \sqrt{\mathrm{E}(h_y^2(Y))}\right) \epsilon \ . \tag{37}$$

Now, (32), (36), and (37) imply

$$S_2 < C_2\, \epsilon \tag{38}$$

a. s. with some constant C_2 for sufficiently large n. The sum S_3 can be treated in a similar way, and we get

$$S_3 < C_3\, \epsilon \tag{39}$$

a. s. for sufficiently large n. The sums S_4 and S_5 are border line cases and can be neglected. Nevertheless, let us study S_4 in some detail. Let G_u be defined as in the proof of Theorem 1 with $\mathcal{B}_x$ in place of $\mathcal{B}$. There are K indices in G_u. If $(x_i, y_i) \in \mathcal{B}$ then $|h(x_i, y_i)| \leq H$, where $H = \max_{(x,y)\in\mathcal{B}} |h(x, y)|$. Similarly, $|h(\tilde{x}_i, \tilde{y}_i)| \leq H$ for $(\tilde{x}_i, \tilde{y}_i) \in \mathcal{B}$. For the other sample points, $|h(x_i, y_i)| \leq h_x(x_i) + h_y(y_i)$ if $(x_i, y_i) \in \bar{\mathcal{B}}$, and $h(\tilde{x}_i, \tilde{y}_i) \leq h_x(\tilde{x}_i) + h_y(\tilde{y}_i)$ if $(\tilde{x}_i, \tilde{y}_i) \in \bar{\mathcal{B}}$. $\sum_{i\in G_u} h_x(x_i)$ and $\sum_{i\in G_u} h_x(\tilde{x}_i)$ can be bounded in the same way as in the proof of Theorem 1. As to the sums with y_i and $\tilde{y}_i$, we have a. s. for sufficiently large n

$$\frac{1}{n} \sum_{i\in G_u} h_y(y_i) \leq \frac{1}{n} K H_y + \frac{1}{n} \sum_{i\in \bar{B}_y} h_y(y_i) < \epsilon \ , \tag{40}$$

where $H_y = \max_{y\in\mathcal{B}_y} h_y(y)$, and

$$\frac{1}{n} \sum_{i\in G_u} h_y(\tilde{y}_i) \leq \frac{1}{n} K H_y + \frac{1}{n} \sum_{\tilde{y}_i\in \bar{\mathcal{B}}_y} h_y(\tilde{y}_i) < \epsilon \ . \tag{41}$$

Thus $S_4 < C_4\epsilon$ and similarly $S_5 < C_5\epsilon$ a. s. for sufficiently large n. Summing up, we have

$$\left|\frac{1}{n}\sum_{i=1}^{n} h(\tilde{x}_i, \tilde{y}_i) - \frac{1}{n}\sum_{i=1}^{n} h(x_i, y_i)\right| < C\epsilon \ . \tag{42}$$

with some constant C. □

Example 2: Let $h(x, y) = h_1(x)h_2(y)$ and suppose that the second moments of $h_1(X)$ and $h_2(Y)$ both exist. Then $\mathrm{E}(h(X, Y))$ is the mixed moment of $h_1(X)$ and $h_2(Y)$. As

$$|h(x, y)| \leq h_1^2(x) + h_2^2(y) \ , \tag{43}$$

the conditions of Theorem 2 are satisfied if h_1 and h_2 are continuously differentiable and $|h_1|$ and $|h_2|$ are monotone for x and y outside a finite interval, respectively. In Schmid and Schneeweiss (2008), Theorem 2 was proved for this special case.

4 Asymptotics

Theorem 1 states that we can consistently estimate any moment $\mathrm{E}(h(X))$ of a random variable X by the arithmetic mean of the transformed microaggregated data $h(\tilde{x}_i)$, in the same way as we would estimate $\mathrm{E}(h(X))$ from the transformed non-microaggregated data $h(x_i)$. We now give conditions under which the estimator constructed from the microaggregated data is as efficient as the corresponding estimator constructed from the non-aggregated data. We can prove even more: The two estimators are asymptotically equivalent under certain conditions, in the sense that $\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} h(\tilde{x}_i) - \frac{1}{n}\sum_{i=1}^{n} h(x_i))$ tends to 0 in probability with $n \to \infty$. There are two conditions that we need, one concerning the transformation function h, the other one concerning the distribution of X.

$\mathcal{H}$ (Condition on h):

a) $h(x)$ is a continuously differentiable function on its domain $\mathcal{D} = (d_l, d_u)$.
b) There is a fixed point $b_0 \in \mathcal{D}$ with the following property: For any closed interval $\mathcal{B} = [b_l, b_u] \subset \mathcal{D}$ such that b_0 is in the interior of $\mathcal{B}$, let $H_l(b_l) =$

$\max_{b_l \le x \le b_0} |h'(x)|$ and $H_u(b_u) = \max_{b_0 \le x \le b_u} |h'(x)|$. Then there exist positive numbers a_l, a_u and m_l, m_u such that for some bounds $\gamma > 0$ and $\delta > 0$,

$$\begin{aligned}
&\text{if } d_l = -\infty, \text{ then } H_l(b_l) \le a_l(b_0 - b_l)^{m_l} \quad \text{for } b_0 - b_l > \gamma \ , \\
&\text{if } d_l > -\infty, \text{ then } H_l(b_l) \le a_l(b_l - d_l)^{-m_l} \quad \text{for } b_l - d_l < \delta \ , \\
&\text{if } d_u = \infty, \text{ then } H_u(b_u) \le a_u(b_u - b_0)^{m_u} \quad \text{for } b_u - b_0 > \gamma \ , \\
&\text{if } d_u < \infty, \text{ then } H_u(b_u) \le a_u(d_u - b_u)^{-m_u} \quad \text{for } d_u - b_u < \delta \ .
\end{aligned}$$

Remark 2: If condition $\mathcal{H}$ holds for some $b_0 \in \mathcal{D}$, then it holds for any $b_0 \in \mathcal{D}$. In particular, b_0 can always be chosen such that $\mathrm{P}(X < b_0) > 0$ and $\mathrm{P}(X > b_0) > 0$.

$\mathcal{F}$ **(Condition on the distribution of** X**):**

a) The support of X is inside $\mathcal{D}$.
b) Let F be the distribution function of X. Then

$$\begin{aligned}
&\text{if } d_l = -\infty, \text{ then } \lim_{n\to\infty} \left[1 - F(b_0 - n^{\frac{1}{4(m_l+1)}})\right]^n = 1 \ , \\
&\text{if } d_l > -\infty, \text{ then } \lim_{n\to\infty} \left[1 - F(d_l + n^{-\frac{1}{4m_l}})\right]^n = 1 \ , \\
&\text{if } d_u = \infty, \text{ then } \lim_{n\to\infty} \left[F(b_0 + n^{\frac{1}{4(m_u+1)}})\right]^n = 1 \ , \\
&\text{if } d_u < \infty, \text{ then } \lim_{n\to\infty} \left[F(d_u - n^{-\frac{1}{4m_u}})\right]^n = 1 \ .
\end{aligned}$$

Remark 3: If condition $\mathcal{F}$ holds for some b_0 it holds for every b_0. Therefore the b_0 of condition $\mathcal{F}$ need not necessarily be the same as the b_0 of condition $\mathcal{H}$ (although it will be so in this paper). In particular, one could set $b_0 = 0$.

While the conditions on h bound the growth of $h'(x)$ when x approaches the boundaries of $\mathcal{D}$, the conditions on the distribution of X describe how fast the distribution function $F(x)$ has to tend to 0 or 1 when x approaches the boundaries of $\mathcal{D}$. The stronger $h'(x)$ grows, i.e., the larger the numbers m_l and m_u are, the faster $F(x)$ has to go to its limits 0 or 1.

Theorem 3. *Suppose an i.i.d. sample $x_1, \ldots, x_n$ of a random variable X has been microaggregated. If the transformation function h and the distribution* *function of X satisfy conditions $\mathcal{H}$ and $\mathcal{F}$, respectively, then*

$$\mathrm{plim}_{n\to\infty}\frac{1}{\sqrt{n}}\left[\sum_{i=1}^{n}h(\tilde{x}_i)-\sum_{i=1}^{n}h(x_i)\right]=0\ . \tag{44}$$

Proof. We need some preliminary definitions. Let $\mathcal{B}_n = [b_l(n), b_u(n)]$ be a closed finite interval depending on n with b_0 lying in its interior. Let $b_l = b_l(n)$ and $b_u = b_u(n)$ be the following functions depending on whether the boundaries of $\mathcal{D}$ are finite or infinite:

$$\begin{aligned}
&\text{if } d_l = -\infty,\ \ b_0 - b_l = n^{\frac{1}{4(m_l+1)}}\ ,\\
&\text{if } d_l > -\infty,\ \ b_l - d_l = n^{-\frac{1}{4m_l}}\ ,\\
&\text{if } d_u = \infty,\ \ b_u - b_0 = n^{\frac{1}{4(m_u+1)}}\ ,\\
&\text{if } d_u < \infty,\ \ d_u - b_u = n^{-\frac{1}{4m_u}}\ .
\end{aligned}$$

Define $B_{l0} := \{i : x_i \in [b_l, b_0]\}$ and $B_{0u} := \{i : x_i \in [b_0, b_u]\}$. According to Remark 1, we can choose b_0 and a fixed closed interval $\mathcal{B}_0 = [b_{0l}, b_{0u}]$ such that $b_{0l} < b_0 < b_{0u}$, $\mathrm{P}(b_{0l} < X < b_0) > 0$, and $\mathrm{P}(b_0 < X < b_{0u}) > 0$. Let G_0 be the the set of indices of the aggregation group for which some x_i lie to the left and some to the right of b_0. (There is at most one such group).

The proof has four variants depending on whether the bounds of $\mathcal{D}$ are finite or infinite. Let us consider only the case, where $d_u = \infty$ and $d_l > -\infty$ (i.e., one bound is finite, the other one infinite). The other three cases can be treated similarly. For any $\epsilon > 0$ let A_n be the event that $\frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}h(\tilde{x}_i) - \sum_{i=1}^{n}h(x_i)\right| > \epsilon$ and let B_n be the event that $x_i \in \mathcal{B}_n$ for all $i = 1, \ldots, n$. Finally let C_n be the event that the sample points with indices belonging to G_0 lie all inside $\mathcal{B}_0$. We have to prove that $\lim_{n\to\infty}\mathrm{P}(A_n) = 0$ for all $\epsilon > 0$. Now

$$\mathrm{P}(A_n)\ \le\ \mathrm{P}(A_n\cap B_n\cap C_n)+\mathrm{P}(\bar{B}_n)+\mathrm{P}(\bar{C}_n)\ . \tag{45}$$

We want to prove that $\mathrm{P}(A_n \cap B_n \cap C_n) \to 0$ as well as $\mathrm{P}(\bar{B}_n) \to 0$ and $\mathrm{P}(\bar{C}_n) \to 0$. First consider the event $A_n \cap B_n \cap C_n$. Under this event (with the notations of the proof of Theorem 1,

$$\begin{aligned}
\epsilon &< \frac{1}{\sqrt{n}}\sum_{i=1}^{n}|h(\tilde{x}_i)-h(x_i)| \;\le\; \frac{1}{\sqrt{n}}\sum_{i=1}^{n}|h'(x_i^*)||x_i-\tilde{x}_i| \\
&\le \frac{1}{\sqrt{n}}\Bigg[H_l(b_l)\sum_{i:G_i\subset B_{l0}}|x_i-\tilde{x}_i| + H_u(b_u)\sum_{i:G_i\subset B_{0u}}|x_i-\tilde{x}_i| \\
&\qquad + \sum_{i\in G_0}|h'(x_i^*)||x_i-\tilde{x}_i|\Bigg] \\
&\le \frac{K}{\sqrt{n}}\Big[H_l(b_l)(b_0-b_l)+H_u(b_u)(b_u-b_0)+h_0\Big]\ , \qquad (46)
\end{aligned}$$

where $h_0 = H_0|\mathcal{B}_0|$, $H_0 = \max_{x\in\mathcal{B}_0}|h'(x)|$, and $|\mathcal{B}_0|$ is the length of $\mathcal{B}_0$. We assume n large enough so that $b_u - b_0 > \gamma$ and $b_l - d_l < \delta$. With $b_0 - b_l = b_0 - d_l - (b_l - d_l)$ and using $\mathcal{H}$, we have

$$\begin{aligned}
\epsilon &< \frac{K}{\sqrt{n}}\Big[a_l(b_0-d_l)(b_l-d_l)^{-m_l}+a_l(b_l-d_l)^{-m_l+1}+a_u(b_u-b_0)^{m_u+1}+h_0\Big] \\
&\le \frac{K}{\sqrt{n}}\Big[a_l(b_0-d_l)n^{\frac{1}{4}}+a_l n^{\frac{1}{4}}+a_u n^{\frac{1}{4}}+h_0\Big] \\
&= K\Big[a_l(b_0-d_l+1)+a_u\Big]n^{-1/4}+Kh_0n^{-1/2}\ , \qquad (47)
\end{aligned}$$

which goes to 0 if $n\to\infty$. Thus $\mathrm{P}(A_n\cap B_n\cap C_n)\to 0$. Next consider $\bar{B}_n$. We have

$$\begin{aligned}
\mathrm{P}(\bar{B}_n) &\le \mathrm{P}\Big(\max_{i=1,\dots,n}x_i > b_u\Big)+\mathrm{P}\Big(\min_{i=1,\dots,n}x_i < b_l\Big) \\
&= 1-[F(b_u)]^n+1-[1-F(b_l)]^n\ . \qquad (48)
\end{aligned}$$

Now, by condition $\mathcal{F}$,

$$[F(b_u)]^n = \Big[F(b_0+n^{\frac{1}{4(m_u+1)}})\Big]^n \to 1 \qquad (49)$$

and

$$[1-F(b_l)]^n = \Big[1-F(d_l+n^{-\frac{1}{4m_l}})\Big]^n \to 1\ . \qquad (50)$$

Thus $\mathrm{P}(\bar{B}_n)\to 0$.

Finally, $\mathrm{P}(\bar{C}_n)\to 0$ because $\mathrm{P}(b_{0l} < X < b_0) > 0$ and $\mathrm{P}(b_0 < X < b_{0u})$ > 0. □

Assuming that the estimator $\frac{1}{n}\sum_{i=1}^{n} h(x_i)$ of $\mathrm{E}(h(X))$ is asymptotically normal with asymptotic variance σ_h^2/n, the asymptotic equivalence of $\frac{1}{n}\sum_{i=1}^{n} h(\tilde{x}_i)$ and $\frac{1}{n}\sum_{i=1}^{n} h(x_i)$ implies that $\frac{1}{n}\sum_{i=1}^{n} h(\tilde{x}_i)$ is also asymptotically normal with the same asymptotic variance:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} h(\tilde{x}_i) - \mathrm{E}(h(X))\right) \to \mathrm{N}(0, \sigma_h^2) \ . \tag{51}$$

Thus the estimator with microaggregated data is (asymptotically) just as efficient as the estimator with the original data.

Example 3: Let $X \sim \mathrm{N}(\mu, \sigma^2)$ and $h(x) = x^k$, $k \in \mathbb{Z}^+$. The estimator $\frac{1}{n}\sum_{i=1}^{n} h(x_i)$ then estimates the k-th moment of X. With microaggregated data the estimator is $\frac{1}{n}\sum_{i=1}^{n} h(\tilde{x}_i)$. We show that the conditions of Theorems 1 and 3 are satisfied. First note that $\mathcal{D} = (-\infty, \infty)$. Obviously, $h(x)$ is continuously differentiable and $|h(x)|$ is monotone for $x > 0$ as well as for $x < 0$. Also $\mathrm{E}(X^k)$ exists. By Theorem 1, $\frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i^k$ is a consistent estimator of $\mathrm{E}(X^k)$.

To show that the conditions of Theorem 3 are satisfied, we choose $b_0 = 0$ and let $b_l = -b_u$, where $b_u > 0$. Obviously $\mathcal{H}$ is satisfied with $m_u = m_l = k - 1$. $\mathcal{F}$ is also satisfied because (assuming w.l.o.g. $\mu = 0$ and $\sigma^2 = 1$)

$$\lim \left[\Phi(n^{\frac{1}{4k}})\right]^n = 1 \ , \tag{52}$$

see Schmid et al. (2007). Similarly,

$$\left[1 - \Phi(-n^{\frac{1}{4k}})\right]^n = \left[\Phi(n^{\frac{1}{4k}})\right]^n \to 1 \ . \tag{53}$$

Thus by Theorem 3, $\frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i^k$ is an asymptotically normal estimator of $\mathrm{E}(X^k)$ with the same asymptotic variance as $\frac{1}{n}\sum_{i=1}^{n} x_i^k$.

Example 4: Let X be lognormally distributed, i.e., $\log X \sim \mathrm{N}(\mu, \sigma^2)$ and let $h(x) = (\log X)^k$, $k \in \mathbb{Z}^+$. The estimator $\frac{1}{n}\sum_{i=1}^{n} h(x_i)$ then estimates the k-th moment of $\log X$. With microaggregated data the estimator is $\frac{1}{n}\sum_{i=1}^{n} h(\tilde{x}_i)$. We show that the conditions of Theorems 1 and 3 are satisfied.

First note that the domain of $\log X$ is $\mathcal{D} = (0, \infty)$ and the support of X coincides with $\mathcal{D}$. Obviously h is continuously differentiable. In addition, $|h(x)| = |\log X|^k$ is monotone for $0 < x < 1$ and for $x > 1$. Also, $\mathrm{E}(h(X))$

exists. Thus Theorem 1 can be applied showing that $\frac{1}{n}\sum_{i=1}^{n}(\log\tilde{x}_i)^k$ is a consistent estimator of $\mathrm{E}[(\log X)^k]$.

To verify the conditions for Theorem 3, first note that $d_l = 0$ and $d_u = \infty$. Let $b_0 = 1$ and let $0 < b_l < 1 < b_u$. As $|h'(x)| = k|\log x|^{k-1}x^{-1}$,

$$H_u(b_u) = \max_{1\le x\le b_u} |h'(x)| = k(k-1)^{k-1}e^{1-k} =: a \tag{54}$$

for b_u large enough and

$$H_l(b_l) = \max_{b_l\le x\le 1} |h'(x)| = k|\log b_l|^{k-1}b_l^{-1} \ . \tag{55}$$

Now $|\log b_l|^{k-1}b_l^{-1} < b_l^{-k}$ because $|\log x| < \frac{1}{x}$ for $0 < x < 1$. Thus, condition $\mathcal{H}$ is satisfied with $m_u = 0$ and $m_l = k$. Without loss of generality we may assume X to be *standard* log-normally distributed, i.e., $\log X \sim \mathrm{N}(0,1)$. Then (with $b_0 = 1$ and $m_u = 0$)

$$F\big(b_0 + n^{\frac{1}{4(m_u+1)}}\big) = \Phi\{\log(1+n^{\frac{1}{4}})\} =: \Phi(b_n) \ . \tag{56}$$

Now $\Phi(b_n) > 1 - \frac{1}{\sqrt{2\pi}}\frac{1}{b_n}e^{-\frac{1}{2}b_n^2}$ for $b_n > 0$, see Durrett (1991), Theorem (1.3). Since $a_n := \frac{n}{\sqrt{2\pi}b_n}e^{-\frac{1}{2}b_n^2} = \frac{n}{\sqrt{2\pi}\log(1+n^{1/4})}e^{-\frac{1}{2}\log^2(1+n^{1/4})} \to 0$ for $n \to \infty$, we have $(1-\frac{a_n}{n})^n \to 1$ and

$$[\Phi(b_n)]^n \to 1 \ . \tag{57}$$

Thus

$$[F(b_0 + n^{\frac{1}{4(m_u+1)}})]^n \to 1 \tag{58}$$

for $n \to \infty$. Similarly, with $d_l = 0$ and $m_l = k$,

$$\begin{aligned} & [1 - F(d_l + n^{-\frac{1}{4m_l}})]^n = [1 - F(n^{-\frac{1}{4k}})]^n \\ = \ & [1 - \Phi(-\frac{1}{4k}\log n)]^n = [\Phi(\frac{1}{4k}\log n)]^n \ \to \ 1 \ . \end{aligned} \tag{59}$$

This shows that condition $\mathcal{F}$ is satisfied as well. Thus Theorem 3 can be applied showing that $\frac{1}{n}\sum_{i=1}^{n}(\log\tilde{x}_i)^k$ is an asymptotically normal estimator of $\mathrm{E}[(\log X)^k]$.

The multivariate case

Theorem 3 carries over to the multivariate case under similar conditions. Here we only consider the bivariate case.

Theorem 4. *Let (X,Y) be a pair of random variables as in Theorem 2 and let $h(x,y)$ be a function satisfying coditions $\mathcal{H}1$ and $\mathcal{H}2$. Let $b_0 = (b_{0x}, b_{0y})$ be a point in $\mathcal{D}$ and let $\mathcal{B} = [b_{lx}, b_{ux}] \times [b_{ly}, b_{uy}]$ be a closed finite rectangle in $\mathcal{D}$. Consider the four subrectangles of $\mathcal{B}$:*

$$\begin{aligned}
\mathcal{B}_{uu} &= [b_{0x}, b_{ux}] \times [b_{0y}, b_{uy}] \ , && (60)\\
\mathcal{B}_{lu} &= [b_{lx}, b_{0x}] \times [b_{0y}, b_{uy}] \ , && (61)\\
\mathcal{B}_{ul} &= [b_{0x}, b_{ux}] \times [b_{ly}, b_{0y}] \ , && (62)\\
\mathcal{B}_{ll} &= [b_{lx}, b_{0x}] \times [b_{ly}, b_{0y}] \ . && (63)
\end{aligned}$$

Let $H_{uu} = H_{uu}(b_{ux}, b_{uy}) = \max_{(x,y)\in\mathcal{B}_{uu}} \max\left(\frac{\partial}{\partial x}h(x,y), \frac{\partial}{\partial y}h(x,y)\right)$ and similarly H_{lu}, H_{ul}, and H_{ll}. Assume that there exist functions $H_{ux} = H_{ux}(b_{ux})$, $H_{uy} = H_{uy}(b_{uy})$, $H_{lx} = H_{lx}(b_{lx})$, and $H_{ly} = H_{ly}(b_{ly})$ such that $H_{uu} \leq H_{ux} + H_{uy}$, $H_{ul} \leq H_{ux} + H_{ly}$, $H_{lu} \leq H_{lx} + H_{uy}$, and $H_{ll} \leq H_{lx} + H_{ly}$, where

$$\begin{aligned}
H_{ux}(b_{ux}) &= \begin{cases} a_u(b_{ux} - b_{0x})^{m_{ux}} & \text{if } d_{ux} = \infty \\ a_u(d_{ux} - b_{ux})^{-m_{ux}} & \text{if } d_{ux} < \infty \end{cases} , && (64)\\
H_{uy}(b_{uy}) &= \begin{cases} a_u(b_{uy} - b_{0y})^{m_{uy}} & \text{if } d_{uy} = \infty \\ a_u(d_{uy} - b_{uy})^{-m_{uy}} & \text{if } d_{uy} < \infty \end{cases} , && (65)\\
H_{lx}(b_{lx}) &= \begin{cases} a_l(b_{0x} - b_{lx})^{m_{lx}} & \text{if } d_{lx} = -\infty \\ a_l(b_{lx} - d_{lx})^{-m_{lx}} & \text{if } d_{lx} > -\infty \end{cases} , && (66)\\
H_{ly}(b_{ly}) &= \begin{cases} a_l(b_{0y} - b_{ly})^{m_{ly}} & \text{if } d_{ly} = -\infty \\ a_l(b_{ly} - d_{ly})^{-m_{ly}} & \text{if } d_{ly} > -\infty \end{cases} . && (67)
\end{aligned}$$

The marginal distributions of X and Y are supposed to satisfy conditions analogous to $\mathcal{F}$ (with $\mathcal{D}$, d_u, d_l b_0, m_u, m_l replaced by $\mathcal{D}_x$, d_{ux}, d_{lx}, b_{0x}, m_{ux}, m_{lx}, and $\mathcal{D}_y$, d_{uy}, d_{ly}, b_{0y}, m_{uy}, m_{ly}, respectively). Under these conditions,

$$\mathrm{plim}_{n\to\infty} \frac{1}{\sqrt{n}} \left[\sum_{i=1}^{n} h(\tilde{x}_i, \tilde{y}_i) - \sum_{i=1}^{n} h(x_i, y_i) \right] = 0 \ . \tag{68}$$

Proof. In principle, the proof of Theorem 4 is analogous to the proof of Theorem 3. Preliminaries are as follows:

Choose $\mathcal{B}_n = [b_{lx}(n), b_{ux}(n)] \times [b_{ly}(n), b_{uy}(n)]$ such that

$$\begin{array}{rcll} b_{ux} - b_{0x} & = & n^{\frac{1}{4(m_{ux}+1)}} & \text{if } \ d_{ux} = \infty \ , \\ d_{ux} - b_{ux} & = & n^{-\frac{1}{4m_{ux}}} & \text{if } \ d_{ux} < \infty \ , \\ b_{0x} - b_{lx} & = & n^{\frac{1}{4(m_{lx}+1)}} & \text{if } \ d_{lx} = -\infty \ , \\ b_{lx} - d_{lx} & = & n^{-\frac{1}{4m_{lx}}} & \text{if } \ d_{lx} > \infty \ , \end{array}$$

and similarly for b_{uy} and b_{ly}.

Define $\mathcal{B}_{0x}$ as $\mathcal{B}_0$ in the proof of Theorem 3 and similarly $\mathcal{B}_{0y}$ with y in place of x. Define G_{0x} as G_0 in the proof of Theorem 3 and similarly G_{0y} with y in place of x. Let $B_{0ux} = \{i : x_i \in [b_{0x}, b_{ux}]\}$, $B_{l0x} = \{i : x_i \in [b_{lx}, b_{0x}]\}$, $B_{0uy} = \{i : y_i \in [b_{0y}, b_{uy}]\}$, and $B_{l0y} = \{i : y_i \in [b_{ly}, b_{0y}]\}$. Now let

- A_n be the event that $\frac{1}{\sqrt{n}} \left|\sum_{i=1}^n h(\tilde{x}_i, \tilde{y}_i) - h(x_i, y_i)\right| > \epsilon$,
- B_n be the event that $(x_i, y_i) \in \mathcal{B}_n$ for all $i = 1, \ldots, n$,
- C_n be the event that $x_i \in \mathcal{B}_{0x}$ if $i \in G_{0x}$ and $y_i \in \mathcal{B}_{0y}$ if $i \in G_{0y}$.

Then $\mathrm{P}(A_n) \leq \mathrm{P}(A_n \cap B_n \cap C_n) + \mathrm{P}(\bar{B}_n) + \mathrm{P}(\bar{C}_n)$.

We prove that all these probabilities go to zero. Consider $A_n \cap B_n \cap C_n$. Under this event $\epsilon < \frac{1}{\sqrt{n}} \sum_{i=1}^n |h(\tilde{x}_i, \tilde{y}_i) - h(x_i, y_i)|$. This sum splits into the five partial sums

$$\begin{array}{ll} \displaystyle\sum_{\substack{G_i \subset B_{0ux} \\ H_i \subset G_{0uy}}} |h(\tilde{x}_i, \tilde{y}_i) - h(x_i, y_i)| \ , & \displaystyle\sum_{\substack{G_i \subset B_{0ux} \\ H_i \subset B_{l0y}}} |h(\tilde{x}_i, \tilde{y}_i) - h(x_i, y_i)| \ , \\ \displaystyle\sum_{\substack{G_i \subset B_{l0x} \\ H_i \subset B_{0uy}}} |h(\tilde{x}_i, \tilde{y}_i) - h(x_i, y_i)| \ , & \displaystyle\sum_{\substack{G_i \subset B_{l0x} \\ H_i \subset B_{l0y}}} |h(\tilde{x}_i, \tilde{y}_i) - h(x_i, y_i)| \ , \\ \displaystyle\sum_{i \in G_{0x} \cup G_{0y}} |h(\tilde{x}_i, \tilde{y}_i) - h(x_i, y_i)| \ . & \end{array}$$

Each partial sum, divided by $\sqrt{n}$, goes to zero as $n \to \infty$. For the last sum this follows from arguments similar to those used in the proof of Theorem 3. As to the other four sums we consider only the first one. The other three can be treated in a similar way.

In the first sum the index i is such that $G_i \subset B_{0ux}$ and $H_i \subset B_{0uy}$. We have

$$\begin{aligned}
& \frac{1}{\sqrt{n}} \sum_i |h(\tilde{x}_i, \tilde{y}_i) - h(x_i, y_i)| \\
= \; & \frac{1}{\sqrt{n}} \sum_i \left| \frac{\partial}{\partial x} h(x_i^*, y_i^*)(x_i - \tilde{x}_i) + \frac{\partial}{\partial y} h(x_i^*, y_i^*)(y_i - \tilde{y}_i) \right| \\
\leq \; & \frac{1}{\sqrt{n}} H_{uu}(b_{ux}, b_{uy}) \left(\sum_i |x_i - \tilde{x}_i| + \sum_i |y_i - \tilde{y}_i| \right) \\
\leq \; & \frac{K}{\sqrt{n}} \left[H_{ux}(b_{ux}) + H_{uy}(b_{uy}) \right] \left[b_{ux} - b_{0x} + b_{uy} - b_{0y} \right] \; .
\end{aligned} \tag{69}$$

Case 1: $d_{ux} = \infty$, $d_{uy} = \infty$. In this case

$$\begin{aligned}
& \frac{K}{\sqrt{n}} \left[H_{ux}(b_{ux}) + H_{uy}(b_{uy}) \right] \left[b_{ux} - b_{0x} + b_{uy} - b_{0y} \right] \\
= \; & \frac{K a_u}{\sqrt{n}} \left[(b_{ux} - b_{0x})^{m_{ux}} + (b_{uy} - b_{0y})^{m_{uy}} \right] \left[b_{ux} - b_{0x} + b_{uy} - b_{0y} \right] \\
= \; & \frac{K a_u}{\sqrt{n}} \left[n^{\frac{m_{ux}}{4(m_{ux}+1)}} + n^{\frac{m_{uy}}{4(m_{uy}+1)}} \right] \left(n^{\frac{1}{4(m_{ux}+1)}} + n^{\frac{1}{4(m_{ux}+1)}} \right) \\
\leq \; & K a_u 2 n^{-1/4} \left(n^{\frac{1}{4(m_{ux}+1)}} + n^{\frac{1}{4(m_{ux}+1)}} \right) \; \to 0
\end{aligned} \tag{70}$$

because $n^{-\frac{1}{4} + \frac{1}{4(m+1)}} = n^{-\frac{m}{4(m+1)}} \to 0$.

Case 2: $d_{ux} < \infty$, $d_{uy} < \infty$. In this case

$$\begin{aligned}
& \frac{K}{\sqrt{n}} \left[H_{ux}(b_{ux}) + H_{uy}(b_{uy}) \right] \left[b_{ux} - b_{0x} + b_{uy} - b_{0y} \right] \\
= \; & \frac{K a_u}{\sqrt{n}} \left[(d_{ux} - b_{ux})^{-m_{ux}} + (d_{uy} - b_{uy})^{-m_{uy}} \right] \\
& \cdot \left[(d_{ux} - b_{0x}) - (d_{ux} - b_{ux}) + (d_{uy} - b_{0y}) - (d_{uy} - b_{uy}) \right] \\
= \; & K a_u 2 n^{-1/4} \left(c_1 - n^{-\frac{1}{4 m_{ux}}} - n^{-\frac{1}{4 m_{uy}}} \right) \to 0 \; ,
\end{aligned} \tag{71}$$

where $c_1 := d_{ux} - b_{0x} + d_{uy} - b_{0y}$.

Case 3: $d_{ux} < \infty$, $d_{uy} = \infty$. In this case

$$\begin{aligned} & \frac{K}{\sqrt{n}} \left[H_{ux}(b_{ux}) + H_{uy}(b_{uy})\right] \left[b_{ux} - b_{0x} + b_{uy} - b_{0y}\right] \\ = \; & \frac{Ka_u}{\sqrt{n}} \left[(d_{ux} - b_{ux})^{-m_{ux}} + (b_{uy} - b_{0y})^{m_{uy}}\right] \\ & \cdot \left[(d_{ux} - b_{0x}) - (d_{ux} - b_{ux}) + (b_{uy} - b_{0y})\right] \\ = \; & \frac{Ka_u}{\sqrt{n}} \left[n^{1/4} + n^{\frac{m_{uy}}{4(m_{uy}+1)}}\right] \left[c_2 - n^{-\frac{1}{4m_{ux}}} + n^{\frac{1}{4(m_{uy}+1)}}\right] \\ \leq \; & Ka_u 2n^{-1/4} \left(c_2 - n^{-\frac{1}{4m_{ux}}} + n^{\frac{1}{4(m_{uy}+1)}}\right) \to 0 \end{aligned} \tag{72}$$

as in Case 1, where $c_2 := d_{ux} - b_{0x}$.

Case 4: $d_{ux} = \infty$, $d_{uy} < \infty$. This case can be treated as Case 3.

Concerning the events $\bar{B}_n$ and $\bar{C}_n$, we have

$$\begin{aligned} \mathrm{P}(\bar{B}_n) \; \leq \; & \mathrm{P}(\max_i x_i > b_{ux}) + \mathrm{P}(\min_i x_i < b_{lx}) \\ & + \mathrm{P}(\max_i y_i > b_{uy}) + \mathrm{P}(\min_i y_i < b_{ly}) \\ = \; & 1 - [F_x(b_{ux})]^n + \; 1 - [1 - F_x(b_{lx})]^n \\ & + 1 - [F_y(b_{uy})]^n + \; 1 - [1 - F_y(b_{ly})]^n \to 0 \end{aligned} \tag{73}$$

and $\mathrm{P}(\bar{C}_n) \to 0$ as in the proof of Theorem 3. □

5 Simulations and data example

We start with a simulation study on the quadratic regression

$$Y = 5 \cdot X^2 + \epsilon \; , \tag{74}$$

where X and ϵ are independent and standard normally distributed each. The slope parameter $\beta = 5$ can be expressed as a continuously differentiable function of the moments $\mathrm{E}(Y \cdot X^2)$ and $\mathrm{E}((X^2)^2)$. Now suppose that Y and X have both been microaggregated with group size $K = 3$, and that the quadratic transformation has to be applied to the data values of X after microaggregation. Then Theorems 1 and 2 guarantee the consistent estimation

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Original data	4.904	4.978	5.009	5.008	5.032	5.157
Microaggregated data	4.902	4.987	5.011	5.017	5.049	5.155

Table 1: Simulation study on quadratic regression - summary statistics of the 100 least squares estimates of β in model (74). The standard deviations of the least squares estimates, multiplied with $\sqrt{n} = \sqrt{300}$, were 0.792 (original data) and 0.855 (microaggregated data).

of $\mathrm{E}(Y \cdot X^2)$ and $\mathrm{E}((X^2)^2)$ from the data (see Examples 1 and 2), and thus also the consistent least squares estimation of β. Moreover, due to Theorem 4, application of the delta method guarantees the asymptotic normality and efficiency of the least squares estimator of β computed from the microaggregated data. It is straightforward to extend these results to the case of a multiple polynomial regression. Table 1 shows the estimation results for $n = 300$ and 100 simulation runs. The similarities between the least squares estimator based on the non-aggregated data and the least squares estimator based on the transformed microaggregated data are obvious.

Our next example is the method-of-moments estimator of the shape and scale parameters of a Gamma distributed random variable X. Denote the shape parameter by α and the scale parameter by β. It is well known that the method-of-moments estimators computed from an i.i.d. sample $x_1, \ldots, x_n$ are $\hat{\alpha} = m_1^2/(m_2 - m_1^2)$ and $\hat{\beta} = (m_2 - m_1^2)/m_1$, where $m_1 := \sum_{i=1}^n x_i/n$ and $m_2 := \sum_{i=1}^n x_i^2/n$ are the first and second empirical moments of X. Since we have shown in Theorem 1 that the corresponding empirical moments computed from a microaggregated data set $\tilde{x}_1, \ldots, \tilde{x}_n$ converge a. s. to m_1 and m_2 as $n \to \infty$, estimation of α and β based on the microaggregated data yields asymptotically the same values as estimation based on the original data. Table 2, where the estimation results of a simulation study with 100 simulation runs are shown, confirms this result ($n = 300$, $K = 3$, $\alpha = 0.5$, $\beta = 2$).

Our third example concerns the maximum likelihood estimation of the scale parameter c of a Levy distribution with density function

$$f(x) = \sqrt{\frac{c}{2\pi}}\,\frac{e^{-c/(2x)}}{x^{3/2}} \ . \tag{75}$$

The score function of a Levy distributed i.i.d. data sample $x_1, \ldots, x_n$ is given

Estimates of α

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Original data	0.372	0.468	0.510	0.514	0.561	0.729
Microaggregated data	0.394	0.479	0.517	0.520	0.564	0.729

Estimates of β

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Original data	1.324	1.757	1.972	1.966	2.167	2.676
Microaggregated data	1.320	1.736	1.945	1.938	2.105	2.622

Table 2: Simulation study on the shape and scale parameter estimation of a gamma distribution - summary statistics of the 100 method-of-moments estimates ($\alpha = 0.5$, $\beta = 2$).

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Original data	1.710	1.900	2.002	2.012	2.124	2.375
Microaggregated data	1.712	1.905	2.012	2.018	2.134	2.377

Table 3: Simulation study on maximum likelihood estimation of the scale parameter c of a Levy distribution - summary statistics of the 100 maximum likelihood estimates ($c = 2$).

by

$$\frac{\partial l}{\partial c}(x_1, \ldots, x_n) = \frac{n}{2c} - \sum_{i=1}^{n} \frac{1}{2x_i} \; . \tag{76}$$

As the maximum likelihood estimator $\hat{c} := \left[\sum_{i=1}^{n}(1/x_i)/n\right]^{-1}$ is a consistent estimator of c, $\left[\sum_{i=1}^{n}(1/\tilde{x}_i)/n\right]^{-1}$ is also consistent (which is guaranteed by Theorem 1 and the monotonicity of $h(x) = 1/x$). Table 3, where the estimation results of a simulation study with 100 simulation runs are shown, confirms this result ($n = 300$, $K = 3$, $c = 2$).

Our final example is an analysis based on the data of the 2004 cost structure survey of enterprises of the mining and manufacturing industry in Germany (KSE). This survey is carried out regularly by the German Federal Statistical Office. As the data obtained from this survey contain comprehensive information on the German industry, they form an important basis for the national accounts of Germany. Also, they are a typical example of an officially collected data set that has to be anonymized before dissemination. The 2004 KSE data has been obtained from $n = 16\,099$ companies with 20 or

	β_1	β_2	β_3	β_4	β_5
non-aggregated data	0.426	0.313	0.065	0.131	0.046
IR (log trafo after MA)	0.426	0.313	0.065	0.131	0.046

Table 4: Least squares estimates of model (77) obtained from the 2004 KSE data. The abbreviation "MA" stands for "microaggregation".

	σ_{β_1}	σ_{β_2}	σ_{β_3}	σ_{β_4}	σ_{β_5}
non-aggregated data	0.0021	0.0036	0.0015	0.0023	0.0027
IR (log trafo after MA)	0.0021	0.0036	0.0015	0.0023	0.0027

Table 5: Estimated standard deviations of the least squares estimates of model (77) obtained from the 2004 KSE data. The abbreviation "MA" stands for "microaggregation".

more employees. Following the approach of Fritsch and Stephan (2003) and Ronning et al. (2005) , we estimate a linear model of the form

$$\log(Y) = \gamma_0 + \sum_{j=1}^{5} \beta_j \log(X_j) + \epsilon \ , \tag{77}$$

where Y is an adjusted gross output of the companies and the regressors $X_1, \ldots, X_5$ are various cost factors. Model (77) corresponds to a logarithmized Cobb-Douglas production function whose production elasticities are equal to the coefficients $\beta_1, \ldots, \beta_5$. As the least squares estimator of model (77) from a microaggregated data set with variables $Y, X_1, \ldots, X_5$ is based on the first and second moments of $\log(Y), \log(X_1), \ldots, \log(X_5)$, Theorems 1 to 4 apply (see Examples 1 and 3), assuming that the regressor variables are (at least approximately) lognormally distributed.

Tables 4 and 5 show the estimation results obtained from the transformed original data and from the transformed microaggregated data (IR with group size $K = 3$). As expected, we see that IR has virtually no effect on the coefficient estimates of model (77) and their estimated standard deviations.[1]

[1] Note that the coefficient estimates in Tables 4 and 5 have been rounded to two decimal places, so the results obtained from the non-aggregated data are not *exactly* the same as the results obtained from the microaggregated data.

6 Summary and conclusion

Microaggregation by individual ranking (IR) is a disclosure control technique which is generally considered to have a relatively small impact on the analytic potential of an anonymized data set. In this paper we have shown analytically that IR has indeed favorable properties with respect to the estimation of statistical models: Any arbitrary moment which is defined as the expectation of a continuously differentiable function h of a set of random variables can be consistently estimated from the microaggregated data by using the standard empirical moment estimators. Moreover, we did not assume the variables under consideration to be continuous. Thus, mixed moments between a microaggregated continuous and a non-microaggregated discrete variable can be estimated, as well as moments purely based on microaggregated continuous variables.

A further important result is the proof of asymptotic normality of the moment estimators based on the microaggregated data. This follows from the fact that the moment estimators are asymptotically equivalent to the corresponding moment estimators computed from the non-aggregated data. Moment estimators with microaggregated and with the original data are thus equally efficient asymptotically. These results have been derived under suitable regularity conditions concerning the behavior of the transformation function h and of the distribution at the border of the domain of h. The simulation studies and data examples presented in Section 5 show that the asymptotic theory derived in this paper is already applicable when sample sizes are relatively small, i.e., when $n \geq 300$.

It should finally be pointed out that the favorable properties of the IR method go hand in hand with a relatively weak protection effect of IR (there is generally a trade-off between analytic potential and protection effect of a disclosure control technique). The protection effect, however, can be enhanced if the group size K is taken sufficiently large. Our asymptotic results do not depend on K. In addition, experiments conducted by Ronning et al. (2005) have shown that the application of IR to a set of continuous variables is sufficiently protective if the discrete variables (which serve as the main identifiers for an attacker) are suitably anonymized by means of appropriate disclosure control techniques for discrete data (an overview of such methods is given in Willenborg and de Waal 2001). Other microaggregation methods, such as distance-based microaggregation techniques (Domingo-Ferrer and Mateo-Sanz 2002, Laszlo and Mukherjee 2005, Domingo-Ferrer et al. 2006) or microaggregation by a sorting variable (Mateo-Sanz and Domingo-Ferrer 1998, Domingo-Ferrer and Torra 2001), are generally considered to be more effective in protecting confidential data than the IR method. However, the analytic potential of these methods seems to be limited (though not useless, see Schmid et al. 2007 for the case of microaggregation by a sorting variable).

While analyzing the protection effect of IR clearly is beyond the scope of this paper, we suggest that in those cases where the application of IR to a data set sufficiently reduces the disclosure risk, IR *should* be applied, since the method guarantees that many standard estimation techniques result in valid findings.

Acknowledgements:

We thank Nadine Bartke and Mario Walter of the Bavarian State Office for Statistics and Data Processing for their support with the analysis of the 2004 KSE data. We also thank Daniel Rost for valuable discussions.

References

Aggarwal, C. C. and P. S. Yu (2008). *Privacy-Preserving Data Mining: Models and Algorithms*. New York: Springer. To appear.

Defays, D. and M. N. Anwar (1998). Masking microdata using microaggregation. *Journal of Official Statistics 14*(4), 449–461.

Domingo-Ferrer, J., A. Martinez-Balleste, J. M. Mateo-Sanz, and F. Sebe (2006). Efficient multivariate data-oriented microaggregation. *International Journal on Very Large Data Bases 15*(4), 355–369.

Domingo-Ferrer, J. and J. M. Mateo-Sanz (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering 14*(1), 189–201.

Domingo-Ferrer, J., A. Oganian, A. Torres, and J. M. Mateo-Sanz (2002). On the security of microaggregation with individual ranking: Analytical attacks. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10*(5), 477–491.

Domingo-Ferrer, J. and V. Torra (2001). A quantitative comparison of disclosure control methods for microdata. In P. Doyle, J. Lane,

J. Theeuwes, and L. Zayatz (Eds.), *Confidentiality, Disclosure, and Data Access*, pp. 111–133. Amsterdam: North-Holland.

Domingo-Ferrer, J. and V. Torra (2004). *Privacy in Statistical Databases.* Berlin: Springer.

Doyle, P., J. Lane, J. Theeuwes, and L. Zayatz (2001). *Confidentiality, Disclosure, and Data Access.* Amsterdam: North-Holland.

Durrett, R. (1991). *Probability: Theory and Examples.* Pacific Grove: Wadsworth & Brooks/Cole.

Fritsch, M. and A. Stephan (2003). Die Heterogenität der technischen Effizienz innerhalb von Wirtschaftszweigen - Auswertungen auf Grundlage der Kostenstrukturstatistik des Statistischen Bundesamtes. In R. Pohl, J. Fischer, U. Rockmann, and K. Semlinger (Eds.), *Analysen zur regionalen Industrieentwicklung - Sonderauswertungen einzelbetrieblicher Daten der amtlichen Statistik*, pp. 143–156. Berlin: Statistisches Landesamt.

Laszlo, M. and S. Mukherjee (2005). Minimum spanning tree partitioning algorithm for microaggregation. *IEEE Transactions on Knowledge and Data Engineering 17*(7), 902–911.

Mateo-Sanz, J. M. and J. Domingo-Ferrer (1998). A comparative study of microaggregation methods. *Questiio 22*(3), 511–526.

Ronning, G., R. Sturm, J. Höhne, R. Lenz, M. Rosemann, M. Scheffler, and D. Vorgrimler (2005). *Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten.* Statistik und Wissenschaft 4. Wiesbaden: Statistisches Bundesamt. In German.

Schmid, M. (2006). Estimation of a linear model under microaggregation by individual ranking. *Allgemeines Statistisches Archiv 90*(3), 419–438.

Schmid, M. and H. Schneeweiss (2008). Estimation of a linear model in transformed variables under microaggregation by individual ranking. Submitted manuscript, University of Munich.

Schmid, M., H. Schneeweiss, and H. Küchenhoff (2007). Estimation of a linear regression under microaggregation with the response variable as a sorting variable. *Statistica Neerlandica 61*(4), 407–431.

Willenborg, L. and T. de Waal (2001). *Elements of Statistical Disclosure Control.* New York: Springer.

Winkler, W. E. (2002). Single-ranking micro-aggregation and reidentification. Statistical Research Division Report RR 2002/08, U.S. Bureau of the Census, Washington.