Journal of Official Statistics, Vol. 14, No. 4, 1998, pp. 399–410

Synthetic and Combined Estimators in Statistical Disclosure Control

Jeroen Pannekoek and Ton de Waal[1]

An often applied procedure in the statistical disclosure control of microdata sets is to prescribe a minimum number of population elements for each category of a combination of identifying variables and to take measures to ensure that there are no categories with a population frequency less than the prescribed minimum. In many cases the population frequencies will be unknown and the disclosure protection procedure can only be applied if a reasonable estimator for these frequencies is available. The usual unbiased direct estimator cannot be applied because it is based on too few sample observations. Since one of the identifying variables is almost always a regional indicator, it seems natural to consider small area estimators for this problem. In this article a synthetic and a combined estimator are proposed and studied, and expressions for their expected mean squared errors are derived. The proposed estimators are compared by means of an example based on data from the Dutch Labour Force Survey.

Key words: Statistical disclosure control; synthetic estimators; combined estimators; small area estimation.

1. Introduction

A common concern of statistical offices that release microdata for use by external researchers is to diminish the risk of disclosure of confidential information on individuals. It is generally accepted that it is insufficient to discard directly identifying variables like names, addresses, etc. because individuals may also be recognized on the basis of their values on other, indirectly, identifying variables such as a geographical indicator, profession, age and sex. If certain combinations of values of identifying variables, or keys, occur only once in the population, the associated individuals score uniquely on these variables. If a researcher, i.e., a potential discloser, knows the values of the identifying variables for certain unique individuals he or she can establish a link between the record and the individual it belongs to. Such a link (identification) leads to disclosure of the remaining information in the record, which was not known beforehand.

To guard the confidentiality of the information provided by the respondents, statistical

[1] Division Research and Development, Department of Statistical Methods, P.O. Box 4000, 2270 JM VOORBURG, The Netherlands.
Acknowledgments: The authors thank the associate editor and a referee for comments that led to considerable improvements of the article.
The views expressed in this article are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

offices apply disclosure limitation techniques (see e.g., Greenberg (1990), Fienberg (1994), Marsh et al. (1994) and the 1993 special issue of the Journal of Official Statistics on confidentiality and data access). An often applied technique for categorical identifiers is to ensure that in the data set combinations of, say, up to three identifying variables do not lead to rare value combinations, i.e., combinations for which the population frequency is less than a prescribed minimum number. This is attained by combining values of identifying variables (recoding) or setting the value of identifiers in some records at ''unknown'' (suppression). The motivation behind such procedures is that unusual value combinations can attract the attention of a researcher who then may be tempted to try and single out unique individuals by using additional identifying variables. Therefore, it is better to try to avoid the occurrence of combinations of scores in the data set that are rare in the population instead of trying to avoid population-uniques only.

Statistics Netherlands applies the above idea for the two kinds of microdata sets it disseminates. The first kind of microdata sets are so-called public use files. A public use file can be obtained by everybody. These data sets are protected rather severely, e.g., variables referring directly to a region of residence are not included to prevent identification. Moreover, very sensitive variables such as variables on sexual behaviour or criminal activities are also not included, to limit the consequences in the unlikely event that despite all precautions an identification still happens.

The second kind of microdata sets are so-called microdata sets for research. A microdata set for research can only be obtained by well-respected (statistical) research offices. The information content of a microdata set for research is much more extensive than that of a public use file. For instance, geographical indicators with much regional detail, such as ''Place of residence,'' may be included in a microdata set for research. Because of the extensive information content of a microdata set for research, researchers have to sign a declaration stating that they will protect any information about an individual respondent that might be disclosed by them. In the remainder of this article we will only consider microdata sets for research.

The disclosure avoidance policy of Statistics Netherlands prescribes that the keys that have to be examined for a microdata set for research consist of three identifying variables, one of which is always a geographical indicator. The (estimated) population frequency of these trivariate combinations should be at least d, where d is a certain well-chosen threshold parameter. When a certain combination does not occur frequently enough in the population, disclosure limitation techniques (see e.g., Greenberg (1990), Marsh et al. (1994)) are applied. Examples of such techniques are combining values of identifying variables (recoding) or setting the value of identifiers in some records at ''unknown'' (suppression).

This rule, including an appropriate value for the threshold parameter d, has been found after a time-consuming trial-and-error process. Many different kinds of combinations have been checked, using many values for the threshold parameter. The final result, the above-mentioned rule, seems to be satisfactory: On the one hand the microdata sets resulting from application of this rule are considered sufficiently protected against disclosure, and on the other hand their information content is still rich enough to suit many statistical analyses. For more information on the kinds of microdata sets released by Statistics Netherlands and their rules we refer to Keller and Willenborg (1993).

Application of the above procedure is trivial if the number of population elements (the population frequency) is known for each category for which a minimum population frequency is required. Often this will not be the case, however, and in such situations one can consider estimating these population frequencies from the sample data. If the sampling fraction is sufficiently large the usual direct estimator for the population frequency (the sample frequency divided by the sampling fraction) can be applied to estimate the population frequencies accurately. If the sampling fraction is not large enough, however, the direct estimator will be too imprecise to be useful. For instance, suppose that the minimum population frequency of a certain category were set at, say, 100, then with a sampling fraction a little less than 1:100, the direct estimator would be zero for zero sample frequencies and more than the minimum of 100 for sample frequencies of 1 or larger. This would imply that no disclosure protection measures were necessary for small samples, a highly implausible result. Of course, no one would consider estimating a population frequency on the basis of only a single sample observation.

As an alternative, we will describe in this article the application of small area estimators, such as synthetic and combined (or compromise) estimators for the required population frequencies. Small area estimators are based upon the sample data as well as upon a model for the population proportions rather than, as is the case with direct estimators, upon the sample only. The model is of vital importance for the quality of the synthetic estimator. If an appropriate model is used then the resulting estimator will usually be quite good, but when an inappropriate model is used the estimator can be severely biased. The combined estimator is hampered less by the use of an inappropriate model because the bias of the synthetic component of the combination is, to some extent, compensated for by the zero bias of the direct component.

For more detailed discussions on the disclosure problem in general we refer to Duncan and Lambert (1989), Bethlehem et al. (1990), Mokken et al. (1992), and Skinner et al. (1994). For a discussion on the disclosure problem in general, and a discussion on the approach based on protecting the individuals with value combinations that occur rarely in the population we refer to De Waal and Willenborg (1996), and Willenborg and De Waal (1996).

The remainder of this article is organized as follows. In Section 2 the synthetic and combined estimators are described and estimators for the expected mean squared errors are derived. In Section 3 the proposed estimators are compared by means of an example based on data from the Dutch Labour Force Survey (LFS). A summary of our conclusions is given in Section 4.

2. Synthetic and Combined Estimators for Proportions in Small Areas

2.1. The synthetic estimator

The proportion μ_{ij} of population elements in an area i that belong to category j is equal to Y_{ij}/N_i, where N_i is the number of inhabitants of area i and Y_{ij} is the number of inhabitants of area i belonging to category j.

Assuming simple random sampling with replacement, the sample proportion Z_{ij} is an

unbiased estimator for μ_{ij}, given by

$$Z_{ij} = \frac{y_{ij}}{n_i} \tag{1}$$

where n_i is the sample size in area i and y_{ij} is the corresponding number of units in the sample in area i that belong to category j. An unbiased estimator for the number of units Y_{ij} in the population in area i that belong to category j is $N_i Z_{ij}$. As is well-known, the variance and the mean squared error (MSE) of Z_{ij} with respect to the sample distribution is given by

$$Var_s(Z_{ij}) = MSE_s(Z_{ij}) = \frac{1}{n_i}\mu_{ij}(1-\mu_{ij}) \tag{2}$$

We can use the overall sample proportion, S_{ij}, as a synthetic estimator for μ_{ij}. So, we define S_{ij} by

$$S_{ij} = \frac{y_{+j}}{n} \tag{3}$$

where $y_{+j} = \Sigma_i y_{ij}$ and $n = \Sigma_i n_i$. The synthetic estimator S_{ij} will, in general, be a biased estimator for μ_{ij}. Only if the μ_{ij} are equal for all areas i will S_{ij} be unbiased for μ_{ij}. A corresponding synthetic estimator for the number of units Y_{ij} in the population in area i that belong to category j is $N_i S_{ij}$.

The variance of S_{ij} with respect to the sample distribution is given by

$$Var_s(S_{ij}) = \frac{1}{n}\mu_{+j}(1-\mu_{+j}) \tag{4}$$

where $\mu_{+j} = \Sigma_i N_i \mu_{ij}/N = \Sigma_i Y_{ij}/N$

The variance of Z_{ij} is at least equal to the variance of S_{ij} because $n_i \leq n$. On the other hand, the synthetic estimator S_{ij} is biased whereas the direct estimator Z_{ij} is not. The bias of S_{ij} is given by

$$b_{ij} = E_s S_{ij} - \mu_{ij} = \mu_{+j} - \mu_{ij} \tag{5}$$

where E_s denotes the expectation with respect to the sample distribution. The mean squared error of S_{ij} is given by

$$MSE_s(S_{ij}) = Var_s(S_{ij}) + b_{ij}^2 \tag{6}$$

2.2. *Estimators for the EMSE of Z_{ij} and S_{ij}*

The MSE (variance) of Z_{ij} depends on μ_{ij} (see, (2)) and the MSE of S_{ij} depends on b_{ij}, which in turn depends also on μ_{ij} (see (5) and (6)). The dependence on μ_{ij} causes difficulty when it comes to estimating these MSE's since there is no satisfactory unbiased estimator for μ_{ij} available (this was the reason for using synthetic estimation in the first place). The usual approach to solving this problem is to assume that b_{ij} is a random variable with expectation $E_b b_{ij}$ equal to zero and variance $Var_b(b_{ij})$ equal to, say, σ_j^2. Here E_b and Var_b denote the expectation and the variance with respect to the distribution of b_{ij}. With these assumptions we can use, instead of the MSE, the expected value with respect to the distribution of b_{ij} of the MSE (EMSE) as a measure of the precision of both Z_{ij} and

S_{ij}. These EMSE's do not depend on the area specific μ_{ij}'s but on both μ_{+j} and σ_j^2 which do not depend on the area i but only on the category j which makes it possible to estimate the EMSE's.

The expected mean squared error (EMSE) of Z_{ij} is given by

$$\begin{aligned} EMSE(Z_{ij}) &= E_b E_s (Z_{ij} - \mu_{ij})^2 = E_b \mu_{ij}(1 - \mu_{ij})/n_i \\ &= \mu_{+j}(1 - \mu_{+j})/n_i - \sigma_j^2/n_i \end{aligned} \tag{7}$$

The expected mean squared error of S_{ij} is given by

$$\begin{aligned} EMSE(S_{ij}) &= E_b E_s (S_{ij} - \mu_{ij})^2 = Var_s(S_{ij}) + E_b b_{ij}^2 \\ &= \mu_{+j}(1 - \mu_{+j})/n + \sigma_j^2 \end{aligned} \tag{8}$$

In order to evaluate $EMSE(Z_{ij})$ and $EMSE(S_{ij})$ it is necessary to estimate μ_{+j} and σ_j^2. An estimator for μ_{+j} is S_{ij}, i.e., y_{+j}/n. An estimator for σ_j^2 can be obtained by means of the sum of the squared differences between the estimated numbers $n_i S_{ij}$ and $n_i Z_{ij}$. The expectation of this squared difference is equal to

$$E_b E_s n_i^2 (S_{ij} - Z_{ij})^2 = \mu_{+j}(1 - \mu_{+j}) n_i (1 + n_i/n) + \sigma_j^2 n_i (n_i - 1) \tag{9}$$

if $E_b E_s S_{ij} Z_{ij} = \mu_{+j}^2$. This latter assumption is justified if the number of different areas is sufficiently large. By setting the sum of all squared differences equal to the expectation of this sum, we obtain the following moment estimate for σ_j^2:

$$\hat{\sigma}_j^2 = \frac{\Sigma_i n_i^2 (S_{ij} - Z_{ij})^2 - S_{ij}(1 - S_{ij}) \Sigma_i n_i \left(\frac{n_i}{n} + 1\right)}{\Sigma_i n_i (n_i - 1)} \tag{10}$$

Spjøtvoll and Thomsen (1987) apply a simpler estimator instead of (10). Their estimator is equal to

$$\hat{\sigma}_j^2 = \frac{\Sigma_i (S_{ij} - Z_{ij})^2 - m S_{ij}(1 - S_{ij})}{(I - m)} \tag{11}$$

where $m = \Sigma_i 1/n_i$ and $I = \Sigma_i 1$, i.e., I is equal to the number of different areas in the sample. When the variance of the synthetic estimator S_{ij} is negligible and all n_i's are equal, the estimator given by (10) is the same as the estimator given by (11).

2.3. The combined estimator

It is well known that it is possible to construct an estimator with a smaller EMSE than both the direct estimator and the synthetic estimator by using a convex combination of these two estimators. This combined estimator, C_{ij}, is given by

$$C_{ij} = W_{ij} Z_{ij} + (1 - W_{ij}) S_{ij} \tag{12}$$

where

$$W_{ij} = \frac{EMSE(S_{ij})}{EMSE(Z_{ij}) + EMSE(S_{ij})} \tag{13}$$

The weight W_{ij} is chosen such that the expected mean squared error of C_{ij} is minimal. Formula (13) shows that the weight approaches 1 if the EMSE of S_{ij} is large compared to

the EMSE of Z_{ij}. Since the variance of S_{ij} is small this will happen when the ''working assumption'' of homogeneous proportions, i.e., that the μ_{ij} are equal for all i, does not at all hold and, consequently, the bias of S_{ij} will be large. In this case, the combined estimator C_{ij} will be close to the unbiased estimator Z_{ij}. At the other extreme, if the bias of S_{ij} is small (resulting in a small EMSE for S_{ij}) or if the variance of Z_{ij} is large (resulting in a large EMSE for Z_{ij}), the weight will approach 0 and the combined estimator will be close to the synthetic estimator S_{ij}.

The expected mean squared error of C_{ij} is approximately given by

$$EMSE(C_{ij}) = W_{ij}^2 EMSE(Z_{ij}) + (1 - W_{ij})^2 EMSE(S_{ij}) \tag{14}$$

The expected mean squared error of C_{ij} is thus at most equal to the minimum of the expected mean squared error of Z_{ij} and S_{ij}.

An estimator of the form (12) can also be obtained by an empirical Bayes argument (Bishop, Fienberg, and Holland (1975), Ch. 12, Albert and Gupta (1983), Gelman et al. (1995), Ch. 2.) In this approach, for each j, the parameters μ_{ij} are viewed as realizations of a random variable, M_j say, with expectation μ_{+j} and variance σ_j^2. If we take the distribution of M_j (the prior distribution) to be the beta (a_j, β_j) distribution we have $\mu_{+j} = \alpha_j / (\alpha_j + \beta_j)$ and $\sigma_j^2 = \mu_{+j}(1 - \mu_{+j})/(\alpha_j + \beta_j + 1)$. Furthermore, if it is assumed that the conditional distribution of y_{ij} given $M_j = \mu_{ij}$ is binomial with parameters n_i and μ_{ij} then the posterior distribution (the conditional distribution of M_j given y_{ij}) is a beta $(y_{ij} + \alpha_j, n_i - y_{ij} + \beta_j)$ distribution. This beta-binomial model is a special case of the more general Dirichlet-multinomial model discussed in Bishop et al. (1975).

The posterior expectation is a Bayesian estimator for μ_{ij}, given by

$$C_{ij}^B = \frac{y_{ij} + \alpha_j}{n_i + \alpha_i + \beta_j} = W_{ij}^B Z_{ij} + (1 - W_{ij}^B)\mu_{+j} \tag{15}$$

with

$$W_{ij}^B = \frac{n_i}{n_i + \alpha_i + \beta_j} = \frac{\sigma_j^2}{\sigma_j^2 + \mu_{+j}(1 - \mu_{+j})/n_i - \sigma_j^2/n_i} = \frac{\sigma_j^2}{\sigma_j^2 + EMSE(Z_{ij})}$$

The Bayesian estimator can only be calculated if the parameters of the prior distribution are known. Alternatively, an *empirical* Bayesian estimator can be used in which the parameters σ_j^2 and μ_{+j} are replaced by estimates obtained from the data values y_{ij} (see e.g., Carlin and Louis 1996). If μ_{+j} is estimated by S_{ij} the empirical Bayes estimator will be a linear combination of Z_{ij} and S_{ij} just like the combined estimator (12); it will use different weights, however, because the sampling variance of S_{ij} is not taken into account. If in the combined estimator the sampling variance of S_{ij} is ignored then $EMSE(S_{ij}) = \sigma_j^2$ (see, (8)) and the combined estimator used in this article will be an empirical Bayes estimator. If the variance of S_{ij} is ignored and the n_i are all equal or σ_j^2 is estimated by (11) instead of (10) then our estimator is equal to the estimator used by Spjøtvoll and Thomsen (1987).

2.4. *Stratified estimators*

The synthetic estimator S_{ij} is based on the ''working assumption'' of homogeneity of the

population proportions μ_{ij}, i.e., the μ_{ij} are equal for all areas i. Although this assumption does not have to be satisfied exactly for the synthetic estimator to perform well, since the bias that is introduced by deviations from the assumption may be compensated for by the small variance of the synthetic estimator, it is worthwhile to investigate if this homogeneity assumption can be relaxed. A straightforward way to proceed is to divide the areas into a small (compared to the number of areas) number of groups of areas and to assume that the μ_{ij} are equal within groups of areas but allow them to vary between groups. This requires that an auxiliary variable is available that indicates to which group each area belongs. For instance, in the application of this article, the areas are municipalities and the auxiliary variable is ''degree of urbanization'' in five categories. Using this auxiliary information allows for a stratified synthetic estimator T_{ij} based on five different values for μ_{ij} (one for each category) instead of only one value for the synthetic estimator S_{ij}. Using T_{ij} we can also construct a stratified combined estimator D_{ij}.

3. Example

3.1. Introduction

The data we have used for this example have been obtained from the Dutch LFS 1991. The microdata set from this survey consists of 84,796 records. Of these records 42,248 have been obtained from male respondents and 42,548 from female respondents. These respondents ranged in age from 15 to 75. From this data set we have used as identifying variables: sex, an area indicator consisting of 646 municipalities and the variable ''profession'' with 90 different categories. For this example it is supposed that there is a disclosure avoidance rule requiring the population frequency for the combination of profession and sex within each municipality to be above a certain value. So the problem is to estimate these frequencies. For convenience we will use in this example the records for males only.

For each of the 646 municipalities and for each of the 90 categories of ''profession,'' the population proportions and frequencies have been estimated using the estimators described in the previous sections of this article. To describe the performance of these estimators in a concise manner, averages of the EMSE's of the estimated proportions were calculated: an average over all 90 categories, an average over categories that did not conform to the homogeneity assumption underlying the synthetic estimator and an average over categories that did conform to this assumption (in the next subsection (3.2) it is explained how ''conforming to the model'' is defined). As is apparent from the discussion in Section 2.3, the synthetic and combined estimates are similar for categories for which the homogeneity model is a good approximation to reality. In Sections 3.3 and 3.5 it is illustrated how these estimates can diverge for categories that are considered outliers with respect to the homogeneity model. The use of auxiliary information to improve the synthetic and combined estimators is studied in Section 3.4. As auxiliary variable we have made use of ''degree of urbanization,'' a categorical variable with five categories that is available for each municipality. In Section 3.6 we show the consequences of the various estimators for the statistical disclosure control problem.

3.2. Definition of outliers

The combined estimator C_{ij} will be approximately equal to the synthetic estimator S_{ij} for

areas that are in accordance with the homogeneity assumption, i.e., for which the proportions for a certain category are close to the mean proportion $\mu_{+j} = \Sigma_i \mu_{ij}/n$ for that category (see, Section 2.3). It will be of interest to see how these estimators compare for areas that deviate from the homogeneity assumption (outliers). For this, we need to determine whether or not a number y_{ij} of units in the sample in an area i have to be considered to be outliers with respect to the homogeneity assumption for category j.

A simple test for outliers is based on the distribution under homogeneity of the number of units in the sample per area. These numbers y_{ij} are approximately independently Poisson distributed with parameter $n_i\mu_{+j}$. If the probability that a Poisson $(n_i\mu_{+j})$ distributed random variable is at least y_{ij} is less than a certain threshold value t, then area i is considered to be an outlier for category j. Likewise if the probability that a Poisson $(n_i\,\mu_{+j})$ distributed variable is at most y_{ij} is less than t, then area i is also considered to be an outlier for category j. The probabilities can be used to order the outliers. In this way the twenty most severe outliers can be listed. This list will be used in Section 3.4 to illustrate the results for the estimators considered.

3.3. *Comparison of expected mean squared errors*

The average expected mean squared errors over three groups of categories of variable ''occupation'' of the direct estimator Z_{ij}, the synthetic estimator S_{ij} and the combined estimator C_{ij} are given in Table 1. The first group of categories considered consists of all 90 categories of ''occupation,'' the second group of the categories with many outliers and the third group of the categories with a few outliers. A category is considered to have many outliers when the number of outlying municipalities is at least 7, otherwise the category is considered to have few outliers.

For all three groups of categories listed in Table 1 we see that the average expected mean squared error of the synthetic estimator S_{ij} is clearly less than the average expected mean squared error of the direct estimator Z_{ij}.

Table 1 moreover clearly demonstrates that the differences between the results for the synthetic estimator S_{ij} and the combined estimator C_{ij} are to be found for categories with many outliers. In Section 3.5 we will therefore examine the results of Z_{ij}, S_{ij}, and C_{ij} for the twenty most severe outliers.

3.4. *Using auxiliary information*

As auxiliary information we have made use of ''the degree of urbanization.'' This ''degree of urbanization'' consists of five categories, ranging from very urbanized municipalities (Category 1), to rural municipalities (Category 5). The average expected mean squared

Table 1. Comparison between the average expected mean squared errors ($\times 10^{-4}$) of Z_{ij}, S_{ij}, and C_{ij} for all categories and for two groups of categories

	Z_{ij}	S_{ij}	C_{ij}
All categories	1.06	0.17	0.11
Many outliers	2.93	0.69	0.42
Few outliers	0.52	0.02	0.02

errors of the stratified synthetic and combined estimators (T_{ij} and D_{ij}) were 0.08×10^{-4} and 0.06×10^{-4}, respectively.

The use of more auxiliary information will generally lead to a decrease of the bias of the resulting estimator. However, the variance of this estimator usually increases as a result of using more auxiliary information. This increase of the variance may be so large that the expected mean squared error also increases. In the extreme case an auxiliary variable could be so detailed that each category describes only one municipality. The resulting estimator would then be the direct estimator, without bias but with a large variance. In this case only five categories are used and the results for the stratified synthetic estimator do not indicate an increase in expected mean squared error relative to the synthetic estimator. However, if a number of auxiliary variables would be available, it would become important to consider alternatives to using a stratification based on the full crossing of all auxiliary variables. In such cases an estimator could be based, for instance, on a parsimonious logit model for the cell proportions with the auxiliary variables as explanatory variables.

3.5. Results for categories with many outliers

The results of Z_{ij}, S_{ij}, C_{ij}, T_{ij}, and D_{ij} for the twenty most severe outliers are given in Table 2. These results indicate that if there are many respondents the combined estimators C_{ij} are almost the same as the direct estimator Z_{ij}. This is a positive feature of C_{ij} and D_{ij}, because when the number of respondents is large the variance of Z_{ij} is relatively small, i.e., Z_{ij} is rather reliable. When the number of respondents is small the differences between the combined estimates and the direct estimates can be rather large because the combined estimator is then drawn strongly towards the synthetic estimator. This is desirable, because when the number of respondents is small, Z_{ij} will be quite unreliable and one would be willing to accept a ''more synthetic'' estimate. The difference between Z_{ij} and the synthetic estimators S_{ij} and T_{ij} is of course rather large in all cases of Table 2 because only outliers are considered here.

3.6. Consequences of the estimators for statistical disclosure control

In this subsection we illustrate the results of the various estimators for the statistical disclosure control problem. As we have indicated in Section 1, we apply a disclosure control rule of the following kind:

> A combination of values of identifying variables is considered safe, i.e., may be published without further protection, if this combination occurs at least d times in the population, where d is a suitable chosen threshold.

Because the population frequency of a combination of values of identifying variables is generally unknown, the above rule is in practice replaced by the rule that the *estimated* population frequency should be at least equal to threshold d in order for a combination of values to be safe.

Table 3 gives the percentage of unsafe combinations of municipality and profession among the combinations that occur at least once in the survey for various values of the threshold d. The total number of combinations equals $646 \times 90 = 58{,}140$, and the number

Table 2. The results of Z_{ij}, S_{ij}, C_{ij}, T_{ij}, and D_{ij} for the twenty most severe outliers

Category	Municipality	n_i	n_iZ_{ij}	n_iS_{ij}	n_iC_{ij}	n_iT_{ij}	n_iD_{ij}
1. Members of the armed forces	Den Helder	326	28	1.0	20.9	1.5	17.9
2. Farmers	Amsterdam	2,694	0	33.0	1.0	1.7	0.0
3. Fishermen, hunters, etc.	Urk	58	6	<0.1	1.4	<0.1	1.4
4. Farmers	Rotterdam	2,202	0	27.0	1.0	1.4	0.0
5. Fishermen, hunters, etc.	Wieringen	45	5	<0.1	1.0	<0.1	0.4
6. Biologists, biochemists, etc.	Wageningen	203	8	0.3	2.5	0.4	2.2
7. Plumbers, welders, etc.	Amsterdam	2,694	4	28.1	6.5	17.5	5.5
8. Bricklayers, carpenters, etc.	Edam-Volendam	187	18	3.4	10.3	2.9	6.9
9. Professional workers n.e.c.[1)]	Amsterdam	2,694	41	15.7	38.6	27.4	36.8
10. Not reporting any occupation	Enschede	749	170	110.5	161.1	115.4	154.9
11. Chemical processors	Terneuzen	243	8	0.6	3.1	0.6	2.8
12. Farmers	Naaldwijk	166	13	2.0	9.3	2.8	5.1
13. Farmers	Nistelrode	70	9	0.9	4.5	2.1	2.5
14. Bricklayers, carpenters, etc.	Amsterdam	2,694	18	48.7	20.2	24.5	18.7
15. Legal professionals	Amsterdam	2,694	19	4.7	17.6	9.8	16.0
16. Farmers	The Hague	1,966	4	24.1	4.8	1.2	4.0
17. Not reporting any occupation	Haarlemmermeer	655	53	96.6	60.3	94.9	65.3
18. Musicians, actors, etc.	Amsterdam	2,694	15	3.1	13.8	7.0	12.3
19. Sculptors, painters, etc.	Amsterdam	2,694	20	5.4	18.4	9.4	17.4
20. Wood preparation workers, etc.	Pekela	110	4	0.1	1.0	0.1	0.7

1) n.e.c. = not elsewhere classified.

of combinations that occur at least once in the survey equals 14,432. So the percentage of unsafe combinations of municipality and profession among all possible combinations can be obtained from Table 3 by multiplying the numbers in Table 3 by 14,432/58,140.

The direct estimator reveals no unsafe combinations for threshold values less than 100. This is not surprising, because in this example the sampling fraction is not the same for each municipality but varies around an average of 1/130, with a maximum of 1/50. This leads to direct estimates that are 50 or larger for combinations with one sample observation. These are implausible estimates but we cannot expect a reasonable estimate for a population frequency on the basis of only one sample observation.

The synthetic and combined estimators produce more plausible estimates, in the sense that it is possible to conclude that combinations are unsafe for threshold values less than 100. The large percentages of unsafe cells for the larger threshold values are also plausible

Table 3. Percentage of unsafe combinations among the combinations that occur at least once in the survey for various values of d

Estimator		Threshold				
		$d = 10$	20	50	100	200
Z_{ij}	(Direct estimator)	0	0	0	1.5	43.0
S_{ij}	(Synthetic estimator)	3.3	7.7	21.8	39.3	62.0
C_{ij}	(Combined estimator)	1.8	4.9	17.2	35.0	58.8
T_{ij}	(Stratified synthetic estimator)	3.2	7.9	22.0	38.3	59.5
D_{ij}	(Stratified combined estimator)	1.9	5.5	18.5	35.9	57.9

because the population cell frequencies are 94 on average and will vary considerably around this average since there are small and large municipalities and the distribution over the professions is also far from uniform. The combined estimators of course give results that are between those of the direct estimator and the synthetic estimators.

As one of the referees pointed out, a more formal comparison of the estimators could be made by investigating the probability of publishing a cell whose population count is below the threshold. This would involve the evaluation of the probability that the estimated population cell count is equal to or larger than the threshold given that the true population count is below the threshold. Such probability statements, however, require a number of assumptions and some further investigation of these assumptions seems necessary. For instance, if we want to calculate the probability that the estimated population count is below the threshold value given that the true population count Y_{ij} has a value Y_0, say, which is smaller than the threshold, we need the distribution of the estimator under the hypothesis $H_0 : Y_{ij} = Y_0$. For the direct estimator one could simply assume a binomial distribution with the probability Y_0/N_i (see Pannekoek 1999), but appropriate assumptions for the synthetic and combined estimators are less obvious. For the combined estimator C_{ij}, for example, we could assume, in line with the assumptions used in the Bayesian approach outlined in Section 2.3, that the distribution under H_0 is a beta distribution with expectation Y_0/N_i. To fully specify this distribution we also need an estimate of the variance under H_0 and further assumptions are needed to obtain such an estimate.

4. Conclusions

An often applied procedure for disclosure protection of microdata sets is to prescribe a minimum number of population elements for each category of a combination of identifying variables and to take measures to ensure that there are no categories with a population frequency less than the prescribed minimum. In many cases the population frequencies will be unknown and the disclosure protection procedure can only be applied if a reasonable estimator for these frequencies is available. The usual unbiased direct estimator cannot be applied because it is based on too few sample observations.

Small area estimators are proposed in this article as an alternative: a synthetic estimator and a combined estimator. Both kinds of estimators were applied with and without using an auxiliary variable with respect to the municipalities (degree of urbanization). With respect to the disclosure problem, the results for all small area estimators were similar. Of the 58,140 estimated population frequencies (90 professions times 646 municipalities) slightly fewer than 40% were below a minimum of 100, so according to these estimators substantial disclosure protection measures will be necessary. This is in line with practices at statistical offices which will almost always use much less detailed area indicators than ''municipality'' and also use classifications of ''profession'' with much less than 90 categories (a one-digit classification is common).

5. References

Albert, J.H. and Gupta, A.K. (1983). Estimation in Contingency Tables Using Prior Information. Journal of the Royal Statistical Society, Series (B), 45, 60–69.

Bethlehem, J.G., Keller, W.J., and Pannekoek, J. (1990). Disclosure Protection for Microdata. Journal of the American Statistical Association, 85, 38–45.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). Discrete Multivariate Analysis; Theory and Practice. MIT Press, Cambridge Mass.

Carlin, B.P. and Lewis, T.A. (1996). Bayes and Empirical Bayes Methods for Data Analysis. Chapman and Hall, New York.

De Waal, A.G. and Willenborg, L.C.R.J. (1996). A View on Statistical Disclosure Control for Microdata. Survey Methodology, 22, 95–103.

Duncan, G. and Lambert, D. (1989). The Risk of Disclosure for Microdata. Journal of Business and Economic Statistics, 7, 207–217.

Fienberg, S.E. (1994). Conflicts Between the Needs for Access to Statistical Information and Demands for Confidentiality. Journal of Official Statistics, 10, 115–132.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). Bayesian Data Analysis. Chapman and Hall, New York.

Greenberg, B. (1990). Disclosure Avoidance Research at the U.S. Census Bureau. Proceedings of the Annual Research Conference, U.S. Bureau of the Census, 144–166.

Keller, W.J. and Willenborg, L.C.R.J. (1993). Microdata Release Policy at the Netherlands CBS Proceedings of the International Seminar on Statistical Confidentiality, Dublin, 439–444.

Marsh, C., Dale, A., and Skinner, C. (1994). Safe Data versus Safe Settings: Access to Microdata from the British Census. International Statistical Review, 62, 35–53.

Mokken, R.J., Kooiman, P., Pannekoek, J., and Willenborg, L.C.R.J. (1992). Assessing Disclosure Risks for Microdata. Statistica Neerlandica, 46, 49–67.

Pannekoek, J. (1999). Statistical Methods for Some Simple Disclosure Limitation Rules. Statistica Neerlandica, forthcoming.

Skinner, C.J., Marsh, C., Openshaw, S., and Wymer, C. (1994). Disclosure Control for Census Microdata. Journal of Official Statistics, 10, 31–51.

Special Issue on Confidentiality and Data Access (1993). Journal of Official Statistics, 9.

Spjøtvoll, E. and Thomsen, I. (1987). Application of Some Empirical Bayes Methods to Small Area Estimation. Proceedings of the 46th Session of the ISI (Book 4), 435–448.

Willenborg, L. and De Waal, T. (1996). Statistical Disclosure Control in Practice. Springer-Verlag, New York.

Received October 1995
Revised January 1998