Outlier Protection in Continuous Microdata Masking*

Josep Maria Mateo-Sanz[1], Francesc Sebé[2], and Josep Domingo-Ferrer[2]

[1] Universitat Rovira i Virgili, Statistics Group
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
jmateo@etseq.urv.es
[2] Universitat Rovira i Virgili
Dept. of Computer Engineering and Mathematics
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
{fsebe,jdomingo}@etse.urv.es

Abstract. Masking methods protect data sets against disclosure by perturbing the original values before publication. Masking causes some information loss (masked data are not exactly the same as original data) and does not completely suppress the risk of disclosure for the individuals behind the data set. Information loss can be measured by observing the differences between original and masked data while disclosure risk can be measured by means of record linkage and confidentiality intervals. Outliers in the original data set are particularly difficult to protect, as they correspond to extreme inviduals who stand out from the rest. The objective of our work is to compare, for different masking methods, the information loss and disclosure risk related to outliers. In this way, the protection level offered by different masking methods to extreme individuals can be evaluated.

Keywords: Statistical database protection, Statistical disclosure control, Outliers, Masking methods.

1 Introduction

Publication of statistical data sets can lead to disclosure of confidential information related to the individual respondents from whom the data have been collected. Several proposals to measure disclosure risk have been proposed in the literature. Measures for disclosure risk are divided into two groups: those designed for categorical data and those designed for continuous data.

Our work focuses on continuous data masked using perturbative methods before publication. Specifically, we concentrate on how statistical disclosure control (SDC) masking methods protect outliers. It must be noted that extreme individuals are particularly easy to identify; thus their disclosure risk is higher.

* This work was partly supported by the European Commission under project "CASC" (IST-2000-25069) and by the Spanish Ministry of Science and Technology and the FEDER fund under project "STREAMOBILE" (TIC-2001-0633-C03-01).

J. Domingo-Ferrer and V. Torra (Eds.): PSD 2004, LNCS 3050, pp. 201–215, 2004.

Our work analyzes, for each masking method, the information loss and disclosure risk of extreme records with respect to the average results for the overall data set. Our objective is not to make an exhaustive study on masking methods, but to gain some understanding on how masking affects outliers.

Disclosure risk measures are based on distance-based record linkage and the construction of confidentiality intervals.

The masking methods compared are: resampling, JPEG lossy compression, multivariate microaggregation, additive noise and rankswapping. Our measures have been obtained from the experimental study of the masked versions of two data sets.

Section 2 lists the masking methods which have been studied in this work. Section 3 specifies the measures used to compute information loss and disclosure risk. The two data sets used to obtain measures are described in Section 4. Section 5 reports on the results obtained for the different masking methods applied to each of the two data sets, *i.e.* the information loss and disclosure risk of outliers with respect to the overall data set. In this way, we determine which masking methods offer the same protection to outliers and average individuals and which masking methods offer a lower protection to outliers. Conclusions are summarized in Section 6.

2 Masking Methods for Continuous Microdata

Several masking methods for protection of statistical data have been presented in the literature. See [10, 5, 11, 1] for a survey of such methods. The masking methods considered in this paper are:

- **JPEG:** This method is based on the JPEG [7] lossy image compression algorithm. The idea is to take original data values (properly scaled) as pixels of a digital image and apply the JPEG algorithm. The JPEG algorithm takes a parameter p (between 1, maximum compression, and 100, minimum compression) which determines the compression level. The masked data set corresponds to the lossy reconstructed data after JPEG compression (properly unscaled). Our experiments have taken p from 10 to 100 in steps of 5.
- **Rank swapping:** This methods perturbs a different variable at each step. It ranks the values of the variable and randomly swaps values that are within a certain maximum rank difference [9]. The maximum rank difference between two swappable values is specified by parameter p, which is expressed as a percent of the total number of records. Our experiments have taken p from 1 to 20 in steps of 1.
- **Additive noise:** This method perturbs a different variable at each step. Each variable is perturbed by the addition of a random Gaussian value with 0 mean and standard deviation $p \cdot s$, where s is the standard deviation of the variable and p is a parameter [8]. Our experiments have taken p from 0.02 to 0.2 in steps of 0.02.

- **Resampling:** For each variable of the masked data set, values are taken through random sampling with replacement from the values in the original data set. This is done n times for each variable and then, after sorting the extracted samples, the average of the n sampled values at each position is computed. At the end, we take each variable in turn and re-order the obtained averages by the order of original data. We have taken n from 1 to 3 in our experiments.
- **Microaggregation:** Taking all variables at a time, groups of k records are built so that each group is formed by records with the maximum similarity between them [3]. The average of each group is published in the masked data set. Our experiments have been performed by taking values for k from 3 to 20 in steps of 1.

3 Measures for Information Loss and Disclosure Risk

Any work aimed at the comparison of different masking methods requires quantitative measures to obtain results. These measures are divided into two groups: those measuring information loss and those measuring disclosure risk.

3.1 Information Loss Measures

In [4] several information loss measures are proposed. Five of these measures were used in that paper to construct a benchmark. Some of those measures compare original and masked data directly and some compare statistics on both data sets. Measures targeted at comparison of statistics are not useful in our work as they refer to the overall data set and do not provide information on individual records, and particularly on outliers.

Let $X = \{x_{ij}\}$ be the original data set and $X' = \{x'_{ij}\}$ a masked version of X. Both data sets consist of n records of d variables each. Our work uses two measures:

- $IL1$: This measure from [4] computes the mean variation between the original and the perturbed version of a record i:

$$IL1 = \frac{100}{d} \sum_{j=1}^{d} \frac{|x_{ij} - x'_{ij}|}{|x_{ij}|}$$

We must take into account that, if the j-th variable in the i-th original record has a value $x_{ij} = 0$, the aforementioned measure will result in a "division by 0" error. In this case, we replace x_{ij} by x'_{ij} in the formula. If both x_{ij} and x'_{ij} are 0, the j-th variable is excluded from the computation for the i-th record (since there are no changes in that variable). In $IL1$, the effect of absolute variation depends on the distance between x_{ij} and 0. $IL1$ is greater for variables near 0. The measure $IL1s$ proposed in [11] overcomes this drawback.

– $IL1s$: Given the original and the perturbed version of a record i

$$IL1s = \frac{1}{d}\sum_{j=1}^{d}\frac{|x_{ij} - x'_{ij}|}{\sqrt{2}S_j}$$

where S_j is the standard deviation of the j-th variable in the original data set.

3.2 Disclosure Risk Measures

The disclosure risk measures used in our work were proposed in [4]. These measures evaluate: i) the risk that an intruder having additional information can link a masked record with the corresponding record in the original data set; ii) the risk that the values of an original record can be accurately estimated from the published masked records. The first kind of risk is evaluated through record linkage and the second kind through the construction of confidentiality intervals.

– **Record linkage:** This measure is based on the assumption that an intruder has additional information (disclosure scenarios) so that she can link the masked record of an individual to its original version. There exist several techniques for record linkage, such as probabilistic [6] and distance-based. Our work uses the distance-based technique. In this technique, linkage proceeds by computing the distances between records in the original and masked data sets. The distances used are standardized to avoid scaling problems. For each record in the masked data set, the distance to every record in the original data set is computed. A record in the masked data set is labeled as correctly linked when the nearest record in the original data set turns out to be the corresponding original record. The percentage DLD of correctly linked records is a measure of disclosure risk.
– **Interval disclosure:** Given the value of a masked variable, we check whether the corresponding original value falls within an interval centered on the masked value. The width of the interval is based on the rank of the variable or on its standard deviation. For data without outliers, using the rank or the standard deviation yields similar results; in the presence of outliers, both ways of determining the interval width are different and complementary.
  - **Rank-based intervals:** For a record in the masked data set, compute rank intervals as follows. Each variable is independently ranked and a rank interval is defined around the value the variable takes for each record. The ranks of values within the interval for a variable around record r should differ less than p percent of the total number of records and the rank in the center of the interval should correspond to the value of the variable in record r. Then, the proportion of original values that fall into the interval centered around their corresponding masked value is a measure RID of disclosure risk. A 100 percent proportion means that an attacker is completely sure that the original value lies in the interval around the masked value (interval disclosure).

- **Standard deviation-based intervals:** Intervals are built around values of the masked variables for each record, but the interval width is not computed in terms of a rank percentage but in terms of a percentage p of the standard deviation of the variable. A measure of risk $SDID$ can be obtained in a way analogous to the way RID is obtained for rank intervals.

In both measures, we have considered 10 different interval widths (from $p = 1\%$ up to $p = 10\%$). The final result is the average of the results obtained for the 10 interval widths.

4 Data Sets

During the Research Meeting of the CASC project[1] held in April 2002 in Plymouth, the need was detected for reference data sets to test and compare microdata SDC methods. As a consequence, three data sets were proposed as reference data for numerical microdata protection. In our work, we have selected two out of those three data sets. The two selected data sets are:

- **"CENSUS" Data Set:** This test data set was obtained using the Data Extraction System of the U.S. Bureau of the Census[2]. Specifically, from the available data sources, the "March Questionnaire Supplement - Person Data Files" from the Current Population Survey of year 1995 was used. 13 quantitative variables were chosen from this data source: AFNLWGT, AGI, EMCONTRB, ERNVAL, FEDTAX, FICA, INTVAL, PEARNVAL, POTHVAL, PTOTVAL, STATETAX, TAXINC, WSALVAL. From the obtained records, a final subset of 1080 records was selected so that there were no repeated values in 7 of the 13 variables. This was done to approximate the behavior of continuous variables (where repetitions are unlikely). A more detailed description of the procedure followed to compile this data set can be found in [2].
- **"EIA" Data Set:** This data set was obtained from the U.S. Energy Information Authority and contains 4092 records[3]. Initially, the data file contained 15 variables from which the first 5 were removed as they corresponded to identifiers. We have worked with the variables: RESREVENUE, RESSALES, COMREVENUE, COMSALES, INDREVENUE, INDSALES, OTHREVENUE, OTHRSALES, TOTREVENUE, TOTSALES.

5 Results

The objective of our empirical work is to compare the measures for outliers with those obtained for the overall data set. We also study the general performance of

[1] `http://neon.vb.cbs.nl/casc`
[2] `http://www.census.gov/DES/www/welcome.html`
[3] `http://www.eia.doe.gov/cneaf/electricity/page/eia826.html`

each measure for each masking method. First of all, we describe the procedure that has been applied to classify a record as an outlier.

The procedure is as follows:

1. To start with, all variables of the original data set are independently standardized. This is done by subtracting from each value the average of the variable and then dividing by the standard deviation of that variable.
2. Next, the average record is considered. Since data have been standardized, the co-ordinates of the average record correspond to the all-zeroes vector.
3. The distance from each record to the average record is now computed.
4. Finally, all records are sorted by their distance to the average record. The 5% farthest records are classified as outliers.

Next, we report on the results obtained for each masking method using the measures described in Section 3. Measures regarding outliers were computed as follows:

- For information loss, $IL1$ and $IL1s$ are computed taking for each outlier record in the original data set its corresponding masked record from the masked data set.
- For disclosure risk, each masked outlier record is compared against all records in the original data set.

These results are presented in several graphics. More specifically, a graphic for each measure and data set is presented. In these graphics, the black line indicates the measure obtained for the overall data set and the gray line indicates the results obtained for the outliers.

Each graphic is divided into five sections corresponding each to a family of masking methods: JPEG (J), rankswapping (R), additive noise (N), resampling (RS) and microaggregation (M). Labels along the abscissae indicate the methods that have been tried: the label for a method consists of a letter indicating the method family and a number indicating a particular parameter choice.

5.1 Information Loss: $IL1$

For the "CENSUS" data set (Figure 1), we observe:

- With JPEG, additive noise and microaggregation, information loss is lower for outliers. This means they receive a lower perturbation.
- With rank swapping, information loss is higher for outliers. As parameter choices change, $IL1$ stays low without significant oscillations.
- For low values of its parameter, JPEG causes the highest information loss. This parameter has a direct influence on the information loss measure.
- The resampling method is the one presenting the lowest information loss.
- In additive noise, the value of the parameter has a great influence on information loss, but less than in the case of JPEG.

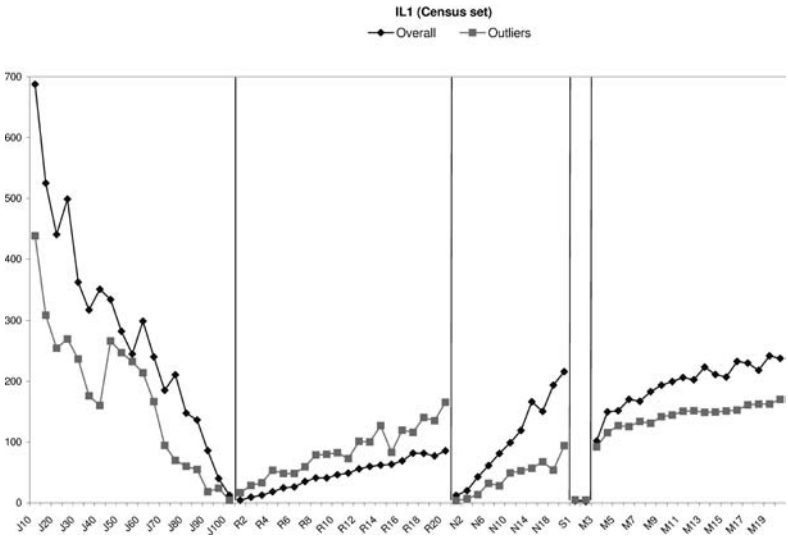

Fig. 1. $IL1$ results for the "CENSUS" data set

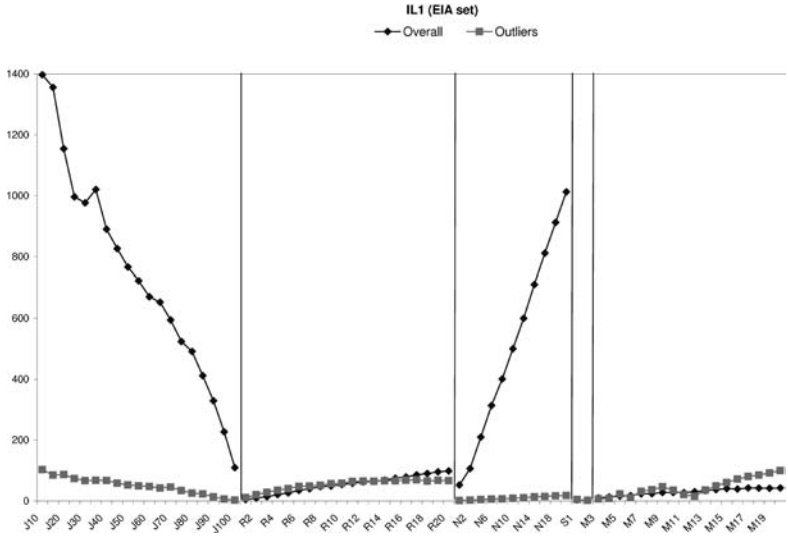

Fig. 2. $IL1$ results for the EIA data set

For the "EIA" data set (Figure 2), we observe:

– Results for the JPEG and additive noise methods are similar to those obtained in the "CENSUS" data set. The differences between $IL1$ for outliers and for the overall data set are very sharp. The overall information loss is much higher than in the "CENSUS" data set.

- In rankswapping, the differences in information loss for outliers and for the overall data set depend on the parameter choice. The overall loss stays close to the one obtained for the "CENSUS" data set.
- If microaggregation is used, the overall loss is much lower than in the "CENSUS" data set. Loss for outliers is similar to the loss for the overall data set until the parameter is close to 15; beyond this value, loss for outliers is a bit higher than for the overall data set.

5.2 Information Loss: $IL1s$

For the "CENSUS" data set (Figure 3), we observe:

- With JPEG, rankswapping, resampling and microaggregation, the information loss for outliers is higher than when measured over the overall data set.
- With additive noise, $IL1s$ is the same for outliers and for the overall data set.
- Generally, additive noise and resampling are those with the lowest $IL1s$. JPEG and rankswapping present a wide range of losses depending on the chosen parameters while microaggregation stays more stable when modifying its parameter.

For the "EIA" data set (Figure 4), we observe:

- In all methods, the behavior of outliers is similar to those observed in the "CENSUS" data set. It is worth mentioning that, with rankswapping, the differences between outliers and the overall data set are magnified.
- In general, $IL1s$ presents the same behavior than in the Census data set. The most evident difference is that, now, resampling and microaggregation are those with the lowest loss.

5.3 Disclosure Risk: Record Linkage DLD

For the "CENSUS" data set (Figure 5), we observe:

- With JPEG, additive noise and resampling, DLD is higher for outliers than for the overall data set; with rankswapping and microaggregation, the contrary happens.
- Globally, resampling presents the highest DLD risk while microaggregation presents the lowest DLD.
- As the parameter choice changes, we can observe that, for JPEG and rankswapping, DLD oscillates in a very wide range; for additive noise, the DLD range is more moderate; for microaggregation, the DLD range is narrowest.

For the "EIA" data set (Figure 6), we observe:

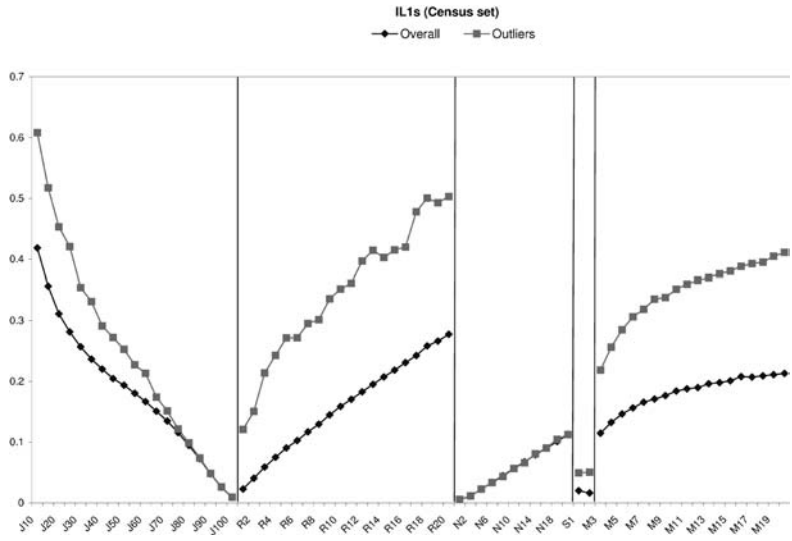

Fig. 3. *IL1s* results for the "CENSUS" data set

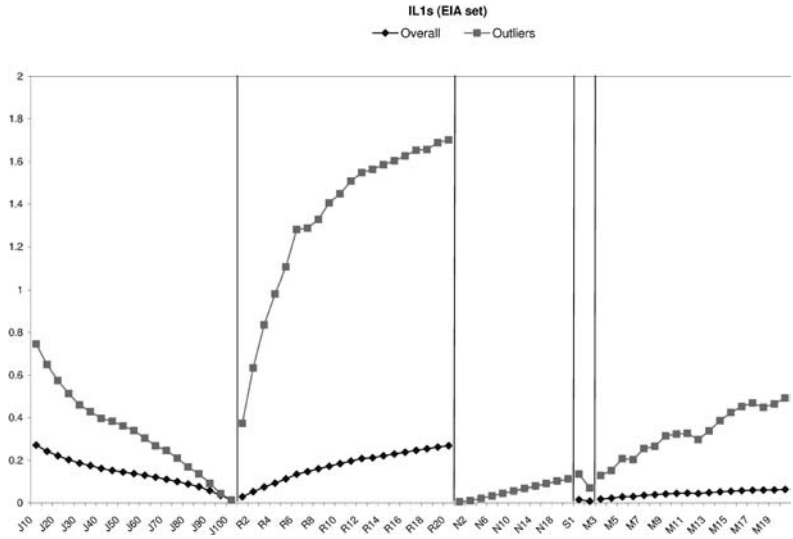

Fig. 4. *IL1s* results for the "EIA" data set

- With JPEG, additive noise, resampling and microaggregation, the *DLD* disclosure risk is higher for outliers than for the overall data set. The difference between the two groups of individuals is very large in JPEG and additive noise, while in microaggregation the difference is very small.
- Rankswapping presents a lower *DLD* for outliers than for the overall data set.

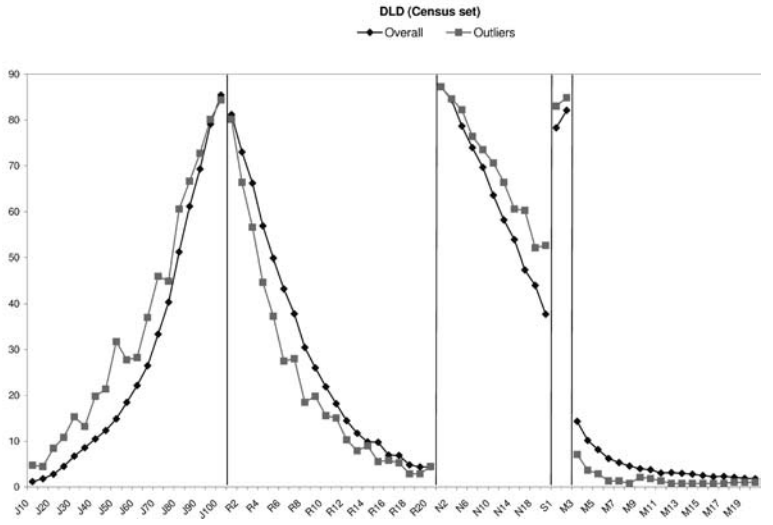

Fig. 5. *DLD* results for the "CENSUS" data set

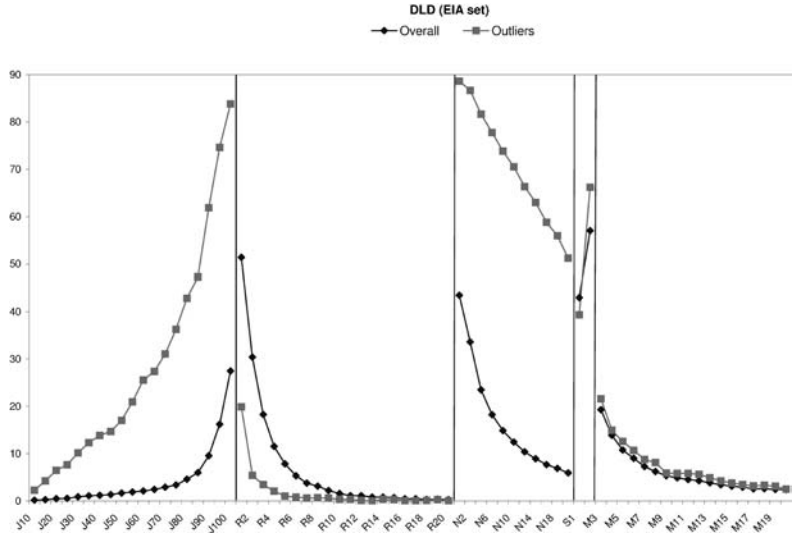

Fig. 6. DLD measures for the "EIA" data set

– Even though the *DLD* value for most methods is heavily dependent on the parameter choice, microaggregation stands out as a method where *DLD* is fairly stable. In rankswapping, *DLD* is very low for parameter values higher than 12.

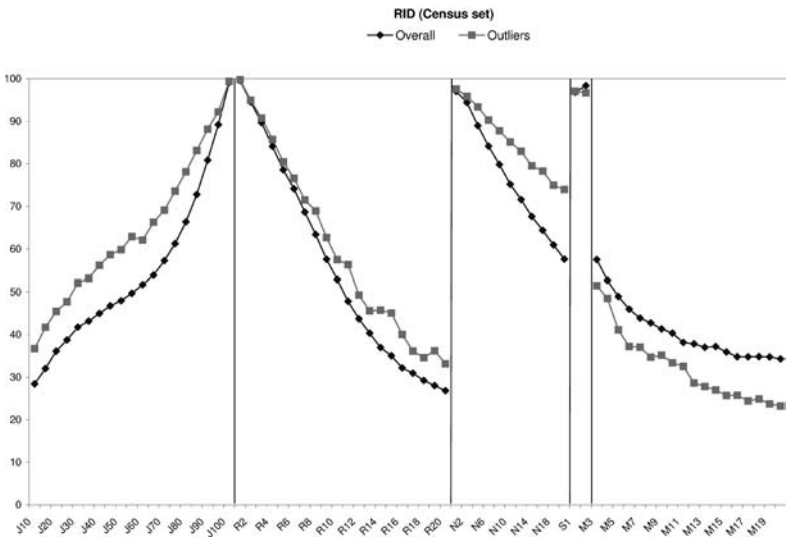

Fig. 7. *RID* measures for the "CENSUS" data set

5.4 Disclosure Risk: Confidence Intervals *RID*

For the "CENSUS" data set (Figure 7), we observe:

- With JPEG, rank swapping and additive noise, *RID* is higher for outliers than for the overall data set. Microaggregation behaves the other way round.
- Resampling has the highest *RID* values, which are very close to the maximum.
- Microaggregation has always lower *RID* values than additive noise.
- As the parameter choice changes, *RID* oscillates in a very wide range for JPEG and rankswapping; for additive noise, the range is only moderately wide; microaggregation displays a narrower range for *RID*.

For the "EIA" data set (Figure 8), we observe:

- With all masking methods, outliers present higher *RID* values than the overall data set. This difference is very large for JPEG and additive noise, moderate with microaggregation and small with rank swapping.
- *RID* values in resampling are extremely high again.
- The width of the range of *RID* values as parameters change exhibits the same behavior, for all methods, than in the "CENSUS" data set.
- Unlike in the previous data set, the behavior of additive noise is better than that of microaggregation.

5.5 Disclosure Risk: Confidence Intervals *SDID*

For the "CENSUS" data set (Figure 9), we observe:

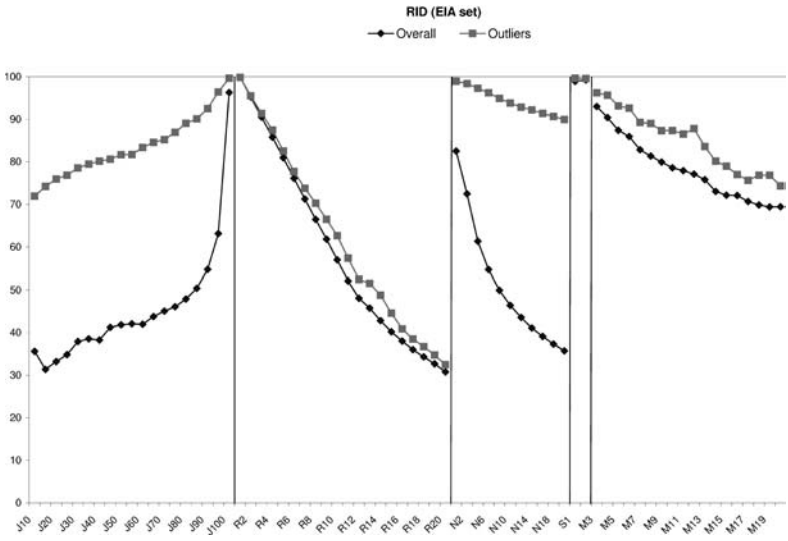

Fig. 8. *RID* measures for the "EIA" data set

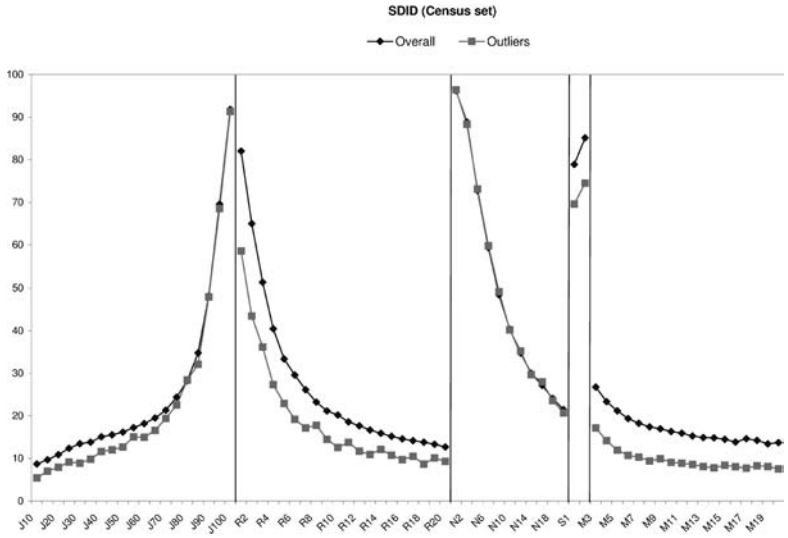

Fig. 9. *SDID* measures for the "CENSUS" data set

- With all methods, except additive noise and some parameterizations of JPEG, *SDID* risk is lower for outliers than for the overall data set. With additive noise and JPEG with high values on the parameter, no difference between the two groups of individuals is perceived.
- Resampling presents the highest *SDID* disclosure risk.
- In general, microaggregation presents better results than noise.

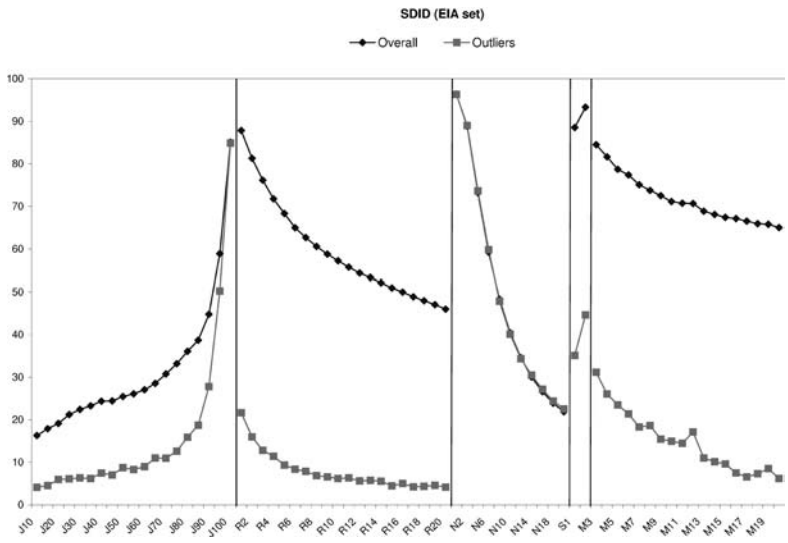

Fig. 10. SDID measures for the EIA data set

- JPEG with low and moderate parameter values, rankswapping with medium and high parameter values and microaggregation with any parameter present similar values on the $SDID$ measure.
- As the parameter choice changes, the range of $SDID$ for additive noise is very wide. Except for high parameter choices of JPEG and low parameter choices of rank swapping, this range for JPEG, rankswapping and microaggregation is small.

For the "EIA" data set (Figure 10), we observe:

- In the various masking methods, $SDID$ values for outliers and for the overall data set compare in a similar way as for the "CENSUS" data set, with the variation that differences are sharper here.
- In general, resampling presents the highest $SDID$ disclosure risk again.
- Also in general, the range of $SDID$ values, depending on the chosen parameter, is very wide for additive noise, moderate for JPEG and rankswapping, and very narrow for microaggregation.

6 Conclusions

From the results outlined in the previous section, we can enumerate some conclusions of the comparison between the masking methods considered in terms of information loss and disclosure risk. Note that our conclusions are based on the results obtained with the two studied data sets. However, a dramatical change of results with other data sets is unlikely. There are two types of conclusions: general or focused on the behavior of outliers with respect to the overall data set.

- Resampling presents the lowest information loss but also the highest disclosure risk. This means that this method is not suitable for SDC.
- Additive random noise has an irregular behavior with respect to other methods when measuring information loss; for instance, it has low $IL1s$ values in the "CENSUS" data set but high $IL1$ values in the "EIA" data set. Regarding disclosure risk measures, additive noise usually yields higher values than other methods; in addition, there is a problem with outliers, for which the risk is higher than for the overall data set. This means that this method should be used very carefully, especially when outliers are present that should be protected.
- Measures on JPEG take a very wide range of values depending on the chosen parameter. When information loss is low, disclosure risk is high, and conversely. Thus, this method is not appropriate for high or low parameter values. It may be usable for medium parameter values. When measuring disclosure risk (DLD and RID), one finds out that the disclosure risk is higher for outliers, especially in the "EIA" data set. Care should be exerted when using this method.
- Rankswapping is a good masking method especially when taking moderately high parameter values. For these values of the parameter, disclosure risk measures stay low while information loss, particularly $IL1$, does not increase significantly. Regarding the behavior of disclosure risk for outliers, we observe that outliers incur a lower risk than the overall data set for the DLD and $SDID$ measures; for the RID measure, the behavior is the opposite but the difference between outliers and the overall data set is not really big. In summary, rankswapping offers more protection to outliers, which is positive.
- Microaggregation is a good masking method as it presents low disclosure risk, especially in the "CENSUS" data set, while information loss stays moderate (lower in the "EIA" data set). The parameter choice is less influential than in rankswapping: the behavior for microaggregation is more robust, even though medium to high values are the best parameter choices. Regarding outliers, this method behaves similarly to rankswapping, except for DLD in the "EIA" data set and RID in the "CENSUS" data set.

We conclude that the best methods, both in general and in regard to outliers, are rankswapping with moderately high parameter values, and microaggregation with medium to high parameter values.

References

1. R. A. Dandekar, M. Cohen and N. Kirkendall, "Sensitive micro data protection using latin hypercube sampling technique", *Lecture Notes in Computer Science*, vol. 2316, pp.245-253, Springer, 2002.
2. J. Domingo-Ferrer and J. M. Mateo-Sanz, project "OTTILIE-R: Optimizing the Tradeoff between Information Loss and Disclosure Risk for continuous microdata", Deliverable D4: "Experiments on test data", U.S.Bureau of the Census (U.S. Department of Commerce), 2001.

3. J. Domingo-Ferrer and J. M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control", *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, pp. 189-201, 2002.
4. J. Domingo-Ferrer, J. M. Mateo-Sanz and V. Torra, "Comparing SDC methods for microdata on the basis of information loss and disclosure risk", *Proceedings of ETK-NTTS 2001*, Luxemburg: Eurostat, pp.807-825, 2001.
5. J. Domingo-Ferrer and V. Torra, "Disclosure protection methods and information loss for microdata", in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, eds. L. Zayatz, P. Doyle, J. Theeuwes and J. Lane, Amsterdam: North-Holland, 2001, pp. 91-110.
6. M. A. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida", *Journal of the American Statistical Association*, vol. 84, pp. 414-420, 1989.
7. Joint Photographic Experts Group, Standard IS 10918-1 (ITU-T T.81). `http://www.jpeg.org`
8. J. J. Kim, "A method for limiting disclosure in microdata based on random noise and transformation", in *Proc. of the ASA Sect. on Survey Res. Meth.*, pp.303-308, 1986.
9. R. Moore, "Controlled data swapping techniques for masking public use microdata sets", U.S. Bureau of the Census, 1996 (unpublished manuscript).
10. L.Willenborg and T. de Waal, *Elements of Statistical Disclosure Control*, New York: Springer-Verlag, 2001.
11. W. E. Yancey, W. E. Winkler and R. H. Creecy, "Disclosure risk assissment in perturbative microdata protection", *Lecture Notes in Computer Science*, vol. 2316, pp.135-152, Springer, 2002.