Journal of Official Statistics, Vol. 28, No. 1, 2012, pp. 107–120

A Multiplicative Masking Method for Preserving the Skewness of the Original Micro-records

Nicolas Ruiz[1]

Masking methods for the safe dissemination of microdata consist of distorting the original data while preserving a predefined set of statistical properties in the microdata. For continuous variables, available methodologies rely essentially on matrix masking and in particular on adding noise to the original values, using more or less refined procedures depending on the extent of information that one seeks to preserve. Almost all of these methods make use of the critical assumption that the original datasets follow a normal distribution and/or that the noise has such a distribution. This assumption is, however, restrictive in the sense that few variables follow empirically a Gaussian pattern: the distribution of household income, for example, is positively skewed, and this skewness is an essential amount of information that has to be considered and preserved. This article addresses these issues by presenting a simple multiplicative masking method that preserves skewness of the original data while offering a sufficient level of disclosure risk control. Numerical examples are provided, leading to the suggestion that this method could be well-suited for the dissemination of a broad range of microdata, including those based on administrative and business records.

Key words: Disclosure; microdata perturbation; sufficient statistics; skewness; log normal distribution.

1. Introduction

Microdata are individual records coming from surveys or administrative registers. Due to their nature and the incredible amount of details that they contain, they generally meet high quality standards. However, this wealth of information is often untapped due to the legal obligations that National Statistical Offices (NSOs) and other governmental institutions face to protect the confidentiality of their respondents, and conservative behaviours that seek to keep the risk of confidentiality breach as low as possible. Such requirements shape the dissemination policy of microdata at national and international levels. The question then is how to ensure a sufficient level of data protection to meet data producer's concerns in terms of legal and ethical requirements while offering to users a reasonable richness of information. To resolve this tension, several solutions are available. These include providing access to microdata through a controlled environment such as data centres, safe remote access, providing interval data or tabulations rather than data

[1] OECD, Statistics Directorate, 2 rue André Pascal, 75775 Paris Cedex 16, France. Email: nicolas.ruiz@oecd.org
Acknowledgments: The author would like to thank Luka A. Ruiz for his superb research assistance. This article has benefited from comments on earlier drafts from David Brackfield, Marco Mira D'Ercole, Martine Durand, and Conal Smith. The views expressed in this article are those of the author and do not necessarily reflect the policies of OECD.

points, or modifying the individual records before public release by using statistical disclosure control techniques.

Statistical Disclosure Control (SDC) consists of the set of numerical tools that enhances the level of confidentiality of any given micro-record while preserving to a lesser or greater extent its level of information (see Hundepool et al. 2010 for an authoritative survey). The increasing number of available techniques makes this field of statistics a potential candidate to support the growing demand for micro-records (OECD 2010). While standards are still missing for the use of SDC in an integrated and coherent framework both at the national and international levels, some techniques are worth looking at due to their tractability and their performance regarding the trade-off between confidentiality and information. Among them, data perturbation has gained considerable attention in the literature.

Data perturbation involves distortions of the original datasets such that unique combinations of original values disappear and new ones are created. This perturbation is made to preserve statistical confidentiality. At the same time, statistical properties (more specifically a selected subset of them) of the original data are preserved, or do not differ significantly. The selection of suitable perturbation methods requires choosing those that will maximize statistical information while minimizing disclosure risk. No dominant method exists, in the sense that the type of statistical information preserved differs among the different techniques available and for different associated levels of disclosure risk.

The general approach to data perturbation consists of the matching of the original data with random noise terms in a nonreversible way, i.e., the data user cannot recover the original values from the perturbed ones. This can be performed in various ways, from a simple additive structure to nonlinear transformations, applicable to both categorical and numerical variables. However, most of the perturbation techniques focus on continuous variables and so will the methodology presented in this article.

In practice, popular perturbation techniques (Brand 2002; Burridge 2003; Muralidhar and Sarathy 2005) use an additive structure for noise application, where error terms are randomly drawn from a normal distribution, the latter being data-dependently parameterized in such a way that the resulting distributions of the perturbed values have the same first and second order moments as those in the original data. As information on these two moments is sufficient to fully identify a normal distribution, this implies that if the original values follow themselves a normal law then the original and the perturbed values will have exactly the same distribution. The loss of statistical information is thus low, in the sense that only the values of the data points of the underlying distribution are altered but not their overall shape. Such a high degree of preservation is made possible by the use of the Gaussian framework. Apart from its peculiar properties, the choice of additive noise methods is motivated by the fact that normality underlies many statistical and econometric tools, extending thus the usefulness of, and audience for, these techniques.

Additive noise methods have, nonetheless, some drawbacks. The most obvious and crucial is the amount of information that is lost when the original data do not follow a normal law. In this case, analyses performed on perturbed data could produce quite different results from those performed on the original set. In particular, the Gaussian framework implies a strong assumption of symmetry in the original distribution. Clearly for numerous economic variables, this assumption is too strong to be tenable.

In fact, microdata often exhibit positively skewed distributions, as in the case of household income and wealth. In particular, lognormal distributions appear to display a reasonable approximation for a large range of economic variables (Kleber and Kotz 2003; Lydall 1966). As such, Gaussian perturbation methods would be of limited utility when applied to such distributions for at least two reasons:

- The first one is that the sum of skewed and nonskewed distributions provides an identifiable distribution only in very rare cases (Gao et al. 2009; Krishnamoorthy 2006). Thus perturbed datasets will, in most cases, follow unknown and unidentifiable distributions.
- The second reason is linked to protection of confidentiality. As the presence of observations far from the mean leads to a skewed distribution, it follows that adding noise drawn from a normal distribution to those observations will only weakly perturb them. As an example, very large firms in business surveys will be resubmitted to high disclosure risk after perturbation.

At this stage, one can ask why we should care about departing from the Gaussian framework. Is the assumption of normality not providing a good enough approximation for the distribution of most economic (and social) variables? The fact is that observations away from the mean in surveys and, more importantly, in administrative records – whether household, individual or firm based – provide crucial statistical information that could contribute greatly to analysis performed on the data. Some recent studies relying on a growing stream of research on income inequality (such as Piketty and Saez 2003; Atkinson et al. 2010) have pointed out the fact that in most developed countries top incomes contribute disproportionately to the overall level of income inequality in a country. As a result, skewness matters, and perturbation methodologies preserving it are of central interest for statistical disclosure control, despite its relative lack of treatment in the literature (see in particular Muralidhar et al. 1995 for an attempt).

This article presents a new multiplicative masking method that preserves positive skewness of the original data based on the lognormal distributions. This method allows users to generate perturbed data that are similar to the original data to a degree that is selected by the user. The methodology preserves confidentiality constraints in particular for observations away from the mean, by changing their ranks in the sample during the perturbation process. The methodology will be presented in the next section, after having first described the features of a general additive Gaussian method based on Muralidhar and Sarathy 2008. The third section proposes numerical validation. The last section concludes.

2. Methodology

Described in this section is the proposed methodology for preservation of asymmetric distributions based on the identification of sufficiency conditions for lognormal distributions. To fully appraise the departure from additive Gaussian methods, we first describe the latter using the recent methodology of Muralidhar and Sarathy (2008), showing how it is possible to generate perturbed data that preserves the distribution of the original dataset but where data points have a selectable degree of similarity.

2.1. The Muralidhar-Sarathy Hybrid Generator

This methodology inoculates noise to a confidential variable in a very general way and as such encompasses the classical additive noise model where no parametric assumption is made as to the type of noise used (Brand 2002) and the Burridge's data perturbation model as a particular case (Burridge 2003).

Let us assume that X is a confidential variable that we want to perturb, and that S is a nonconfidential or a key variable with a low level of identification risk. (Confidential variables are variables containing sensitive information that has to be protected from disclosure risk, as opposed to nonconfidential variables where their disclosure does not raise any confidentiality issue.) Without loss of generality, it is assumed that the means of X and S are equal to zero. Let σ^2_{XX}, σ^2_{SS} and σ^2_{SX} be, respectively, the variance of X, S and the covariance between X and S. We will denote by Y the perturbed value of X generated by the following equation (where $y_i, x_i, s_i \, \forall i = 1, \ldots, \mathrm{n}$ are the values of Y, X and S variables for the ith respondent in the dataset):

$$y_i = \left[(1-\alpha)\frac{1}{n}\sum_{i=1}^{n} x_i - \beta\frac{1}{n}\sum_{i=1}^{n} s_i\right] + \alpha x_i + \beta s_i + u_i \, \forall i = 1, \ldots, \mathrm{n}$$

α and β are coefficients and u_i is a random term generated from a normal distribution $\mathrm{N}(0, \sigma^2_{uu})$, satisfying

$$\frac{1}{n}\sum_{i=1}^{n} x_i u_i = \frac{1}{n}\sum_{i=1}^{n} s_i u_i = 0 \, (x_i \text{ and } s_i \text{ are orthogonal to } u_i)$$

This equation shows that α is a similarity parameter between Y and X. When $\alpha = 0$, X and Y are completely dissimilar. For $\alpha = 1$, Y equals X and no perturbation is added. Thus, the choice of α allows the user (e.g., NSOs in the case of records from official sources) to control for the degree of similarity between the original and the perturbed variable that will be disseminated.

The conversion of X into Y through the preceding equation adds "noise" to the original X variable. In fact, it is easy to verify that

$$E(y_i) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and thus that X and Y will have the same expectation: the first moment of X's distribution is then preserved. To preserve the second moment, the following condition must be satisfied:

$$\sigma^2_{XX} = \sigma^2_{YY} = E[(\alpha x_i + \beta s_i + u_i)(\alpha x_i + \beta s_i + u_i)] = \alpha^2\sigma^2_{XX} + \beta\sigma^2_{SS} + \sigma^2_{uu} + 2\alpha\beta\sigma^2_{SX}$$

Finally, in order to preserve the covariance between the confidential and nonconfidential variables, the following equation must also hold:

$$\sigma^2_{SX} = \sigma^2_{SY} = \alpha\sigma^2_{SX} + \beta\sigma^2_{SS}\left(\text{as}\frac{1}{n}\sum_{i=1}^{n} s_i u_i = 0\right) \Leftrightarrow \beta = (1-\alpha)\frac{\sigma^2_{SX}}{\sigma^2_{SS}}$$

Combining the two preceding equations, we obtain the following restriction for σ^2_{uu}:

$$\sigma^2_{uu} = (1 - \alpha^2)\left[\sigma^2_{XX} - \frac{\left(\sigma^2_{SX}\right)^2}{\sigma^2_{SS}}\right]$$

The term

$$\left[\sigma^2_{XX} - \frac{\left(\sigma^2_{SX}\right)^2}{\sigma^2_{SS}}\right]$$

is always larger than or equal to zero. Thus the necessary and sufficient condition for $\sigma^2_{uu} > 0$ is that $-1 \leq \alpha \leq 1$. As a negative α induces a negative correlation between the original and the perturbed value, this case is ignored in the following, i.e., we will focus only on $0 \leq \alpha \leq 1$ to fulfill the above restrictions.

When α is set to 1, X = Y and no perturbation is added; when $\alpha = 0$, Y is not a function of the (confidential) value X but only of the nonconfidential variable S and of an error term. The intermediary cases where $0 < \alpha < 1$ create therefore a hybrid dataset, as the released variable is a combination of its original value, of the nonconfidential variable S and of a noise term. Through this method, users can choose to which extent they want to protect their initial release. This procedure is perfectly secure in the sense that no reverse engineering is possible as the hybridation is performed using a random draw for u_i. A direct consequence of this algorithm is that users can choose to communicate transparently their chosen degree of "dissimilarity": in other words, knowledge of α provides access to the value of σ^2_{uu} but not to the u_i values themselves (although maintaining the confidentiality of α does provide an additional security gate).

While it can be argued that this method implies significant information loss, in fact statistical information is preserved to a greater extent than with other approaches (such as those described by Fuller 1993). In particular, the Muralidhar-Sarathy method preserves the first two moments of the X's distribution, these moments being the necessary and sufficient conditions for the identification of a normal distribution; it follows that if the distribution of X is normal, then Y will have exactly the same distribution as the original, undisclosed variable. Moreover, by using a nonconfidential variable in the perturbation process, this method allows preserving the covariance between the confidential variable X and the nonconfidential variable S.

As appealing as this framework appears, it relies nonetheless on the pivotal normality assumption. Normality underlies many statistical analyses commonly used (such as regression and hypothesis testing), and assures that analysis based on the masked data will lead to the same results that one would have obtained with the original data – but with the advantage that the secure environment avoids disclosure risks. But this approach could be problematic for other uses. First, if a user, rather than being interested in performing econometrics and inference, chooses instead to focus on the intrinsic features of the distribution and wants to perform analysis on subdomains, then this methodology while ensuring the preservation of parameters at a general level, does not guarantee it at some finer level of disaggregation. In this case, a recent improvement for hybridation render it

possible to preserve some properties for some selectable level of disaggregation (Domingo-Ferrer and Gonzalez-Nicolas 2010).

Second, this approach could potentially bias the computation of some measures of dispersion as in order to properly perform this task, additional features of the original distribution are required, in particular skewness which conveys substantial and relevant information on the dispersion of a given distribution. This last consideration leads to a new multiplicative method presented hereafter.

2.2. *A Sufficient Multiplicative Masking Method for Lognormal Distributions*

Using the same notation as before, we let X follow a lognormal distribution with parameters $\mu_X > 0$ and σ^2_{XX}

$$X \mapsto LN\left(\mu_X; \sigma^2_{XX}\right)$$

where, by the definition of a lognormal distribution,

$$\mu_X = \frac{1}{n}\sum_{i=1}^{n} \ln x_i \text{ and } \sigma^2_{XX} = \frac{1}{n}\sum_{i=1}^{n} (\ln x_i - \mu_X)^2$$

The first and second order moments of X are thus respectively

$$E(X) = \exp\left(\mu_X + \frac{\sigma^2_{XX}}{2}\right)$$

$$\text{and } V(X) = \left[\exp\left(\sigma^2_{XX}\right) - 1\right]\exp\left(2\mu_X + \sigma^2_{XX}\right)$$

The same assumptions apply for the perturbation U, assumed to be independent of X and with parameters

$$\mu_U = \frac{1}{n}\sum_{i=1}^{n} \ln u_i > 0 \text{ and } \sigma^2_{UU} = \frac{1}{n}\sum_{i=1}^{n} (\ln u_i - \mu_u)^2$$

$$U \mapsto LN\left(\mu_U; \sigma^2_{UU}\right)$$

$$\text{with } E(U) = \exp\left(\mu_U + \frac{\sigma^2_{UU}}{2}\right) \text{ and } V(U) = \left[\exp\left(\sigma^2_{UU}\right) - 1\right]\exp\left(2\mu_U + \sigma^2_{UU}\right)$$

The perturbed value of X, Y is generated through the following equation, a homothetic Cobb-Douglas function:

$$Y = X^{\alpha}U^{1-\alpha} \text{ with } 0 \leq \alpha \leq 1$$

As for the Muralidhar-Sarathy hybrid generator, α is also a similarity parameter: when α is set to 1, $X = Y$ and no perturbation is generated; when $\alpha = 0$, Y is not a function of the confidential value X but only of the lognormal noise. The intermediary cases $0 < \alpha < 1$ create convex combinations of confidential values and noise.

By the properties of a lognormal distribution (Krishnamoorthy 2006), the α power distribution of X also follows a lognormal law

$$X^{\alpha} \mapsto LN\left(\alpha\mu_X; \alpha^2\sigma_{XX}^2\right)$$

and the same applies for the $1 - \alpha$ power of U

$$U^{1-\alpha} \mapsto LN\left((1-\alpha)\mu_U; (1-\alpha)^2\sigma_{UU}^2\right)$$

By independence of U and X, Y has thus the following distribution

$$Y \mapsto LN\left(\alpha\mu_X + (1-\alpha)\mu_U; \alpha^2\sigma_{XX}^2 + (1-\alpha)^2\sigma_{UU}^2\right)$$

with the associated two first moments being:

$$E(Y) = \exp\left(\alpha\mu_X + (1-\alpha)\mu_U + \frac{\alpha^2\sigma_{XX}^2 + (1-\alpha)^2\sigma_{UU}^2}{2}\right) \text{ and}$$

$$V(Y) = \left[\exp\left(\alpha^2\sigma_{XX}^2 + (1-\alpha)^2\sigma_{UU}^2\right) - 1\right] \times \exp\left[2\left(\alpha\mu_X + (1-\alpha)\mu_U\right) + \alpha^2\sigma_{XX}^2 + (1-\alpha)^2\sigma_{UU}^2\right]$$

We can now derive the necessary and sufficient conditions that will ensure that Y has the same distribution as X. Unlike the additive framework, we cannot proceed by preserving the first two moments of Y. More generally any set of *k*-order moments with $k \geq 1$ is not isomorphic to any set of lognormal laws: we can in fact always find other laws (lognormal or not) that have the same moments. To achieve sufficiency we have to consider the logarithmic transformation of Y

$$\ln Y \mapsto N\left(\alpha\mu_X + (1-\alpha)\mu_U; \alpha^2\sigma_{XX}^2 + (1-\alpha)^2\sigma_{UU}^2\right)$$

Being now in a Gaussian case, we can derive conditions for the first two moments

$$\alpha\mu_X + (1-\alpha)\mu_U = \mu_X \Leftrightarrow \mu_X = \mu_U$$

$$\alpha^2\sigma_{XX}^2 + (1-\alpha)^2\sigma_{UU}^2 = \sigma_{XX}^2 \Leftrightarrow \sigma_{UU}^2 = \frac{1-\alpha^2}{(1-\alpha)^2}\sigma_{XX}^2$$

As $\sigma_{UU}^2 > 0$, we also have $1 - \alpha^2 \geq 0$ and thus $0 \leq \alpha \leq 1$, confirming α as a well-defined similarity parameter. Using the sufficiency conditions at the logarithmic level and exponentiating $\ln Y$, we find that U must have the following lognormal distribution:

$$U \mapsto LN\left(\mu_X; \frac{1-\alpha^2}{(1-\alpha)^2}\sigma_{XX}^2\right)$$

As exponentiation establishes a one to one correspondence (i.e., it is a bijective mapping), the sufficiency conditions at the logarithmic scale ensure sufficiency at the original variable scale. Thus this perturbation method preserves the features of the original distribution including its skewness, but allows the similarity of data points to be selected.

This methodology constitutes a natural extension of several previous multiplicative noise protocols that aims to preserve the skewness of micro-records, at least approximately

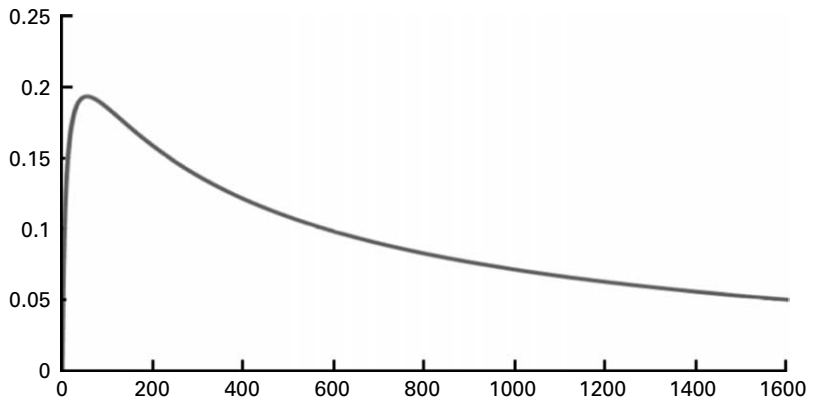

Fig. 1. Density of original data

(see, for example Höhne 2004; Kim and Winkler 2001; Oganian and Karr 2011) or asymptotically (see Sarathy et al. 2002 for a copula-based approach). The present methodology preserves exactly the property of the underlying data at finite distance and asymptotically but to the condition of log-normality. It is thus strongly rooted to a parametric assumption while the other methodologies are more flexible, at the cost of some approximations at least at finite distance. In that sense, our methodology is probably best suited for small samples.

In terms of generalization of the outlined approach, this multiplicative method has been derived in the case of a univariate confidential variable. Its multivariate counterpart (such as is to be found in Muralidhar and Sarathy 2008 for the hybrid generator) is a natural extension to consider. But this generalization to any finite set of confidential variables suffers potentially from several pitfalls, a major one being that working with multivariate log-normal distribution is computationally cumbersome as numeric problems are likely to occur (see Mostafa and Mahmoud 1964 for the bivariate case). Moreover, the availability of more than one skewed confidential variable is highly unlikely in a large variety of micro-records (for example, income and wealth are rarely measured together in a single

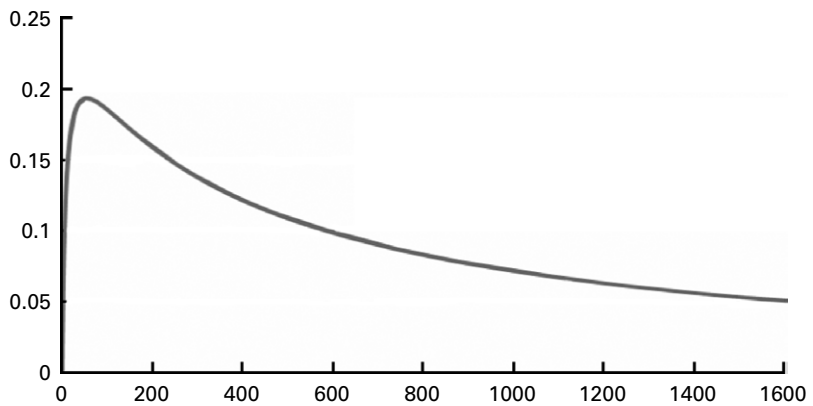

Fig. 2. Density of perturbed data with $\alpha = 0.9$

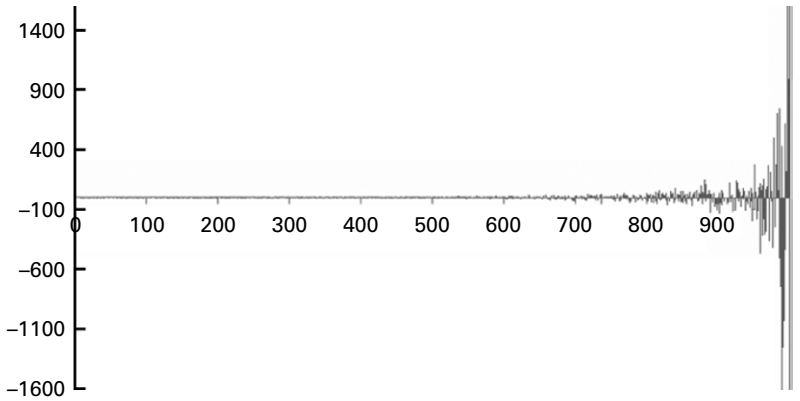

Fig. 3. Differences between original and perturbed data for $\alpha = 0.999$

survey). These two factors lead us to think that, regarding the simplicity and easy implementation of the univariate case and the potential lack of application for the multivariate one, there are no clear advantages that could arise from such an extension.

In term of disclosure risk, and as shown in the following section, this method is also confidentiality efficient in the sense that the risk remains relatively low, in particular for observations far from the mean.

3. Numerical Validation

Methods for statistical disclosure control cannot be fully appraised without experimental validation (programs and the dataset are available from the author upon request). We simulated a vector consisting of one thousand data points drawn from a lognormal distribution with parameters 4 and 2, i.e., a deliberately highly skewed distribution. Figure 1 shows the density of the original distribution.

When $\alpha = 0.9$, the distribution of the perturbed data matches exactly that of the original data: as one can see in Figure 2, the density of the former is strictly identical to the latter.

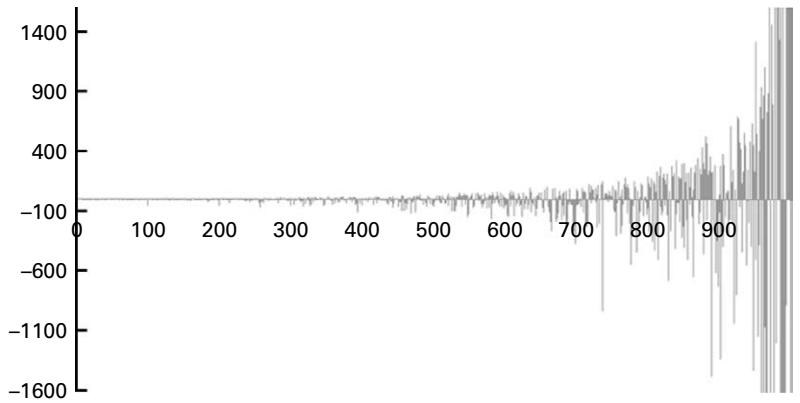

Fig. 4. Differences between original and perturbed data for $\alpha = 0.95$

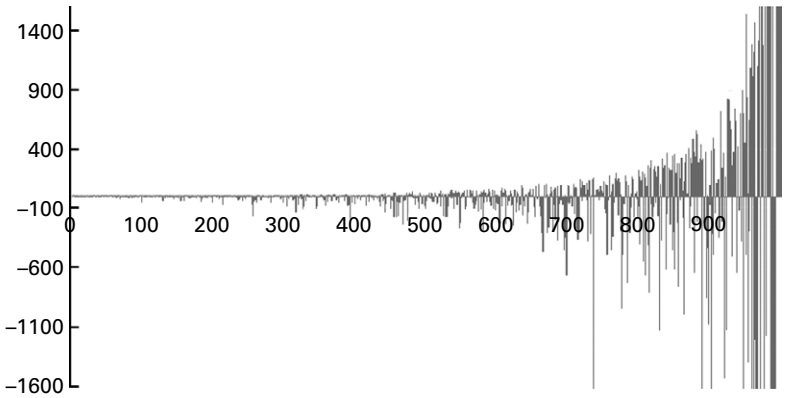

Fig. 5. Differences between original and perturbed data for $\alpha = 0.9$

As derived in the previous section, perturbed distributions will remain the same as the original ones for $0 \leq \alpha \leq 1$. Thus, the multiplicative masking method preserves the initial data structure. Nonetheless, data points are altered in an interesting way, in particular for confidentiality purposes. Figure 3 depicts the changes that occur in the absolute values for each point (ranked in ascending order on the x-axis according to their original values).

One immediately sees that, for a small value of the dissimilarity parameter, most of the data points that are close to the mean are very close to the original values while, due to the multiplicative structure used, values that are far away from the population mean are substantially altered. And as these high values are those where disclosure risk is higher, this pattern of perturbation is the one most appropriate. For lower values of α, and thus greater dissimilarity, perturbations start to spread along the distribution, from the upper to the lower tail as can be seen in Figures 4, 5 and 6.

As perturbations can both reduce and increase values of different data points, the ranking of data points is likely to be changed during the process, thus increasing data protection against disclosure risk (in particular observations away from the mean could now become close to it and conversely). As shown in Figures 7 and 8, the more dissimilarity is introduced, the more changes occur in the data ranking.

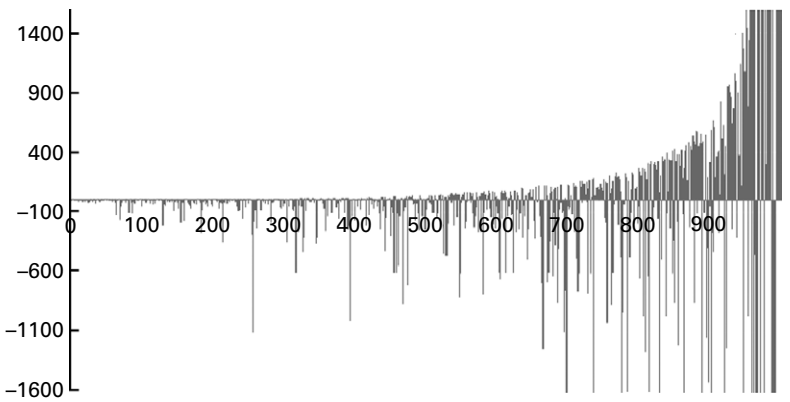

Fig. 6. Differences between original and perturbed data for $\alpha = 0.7$

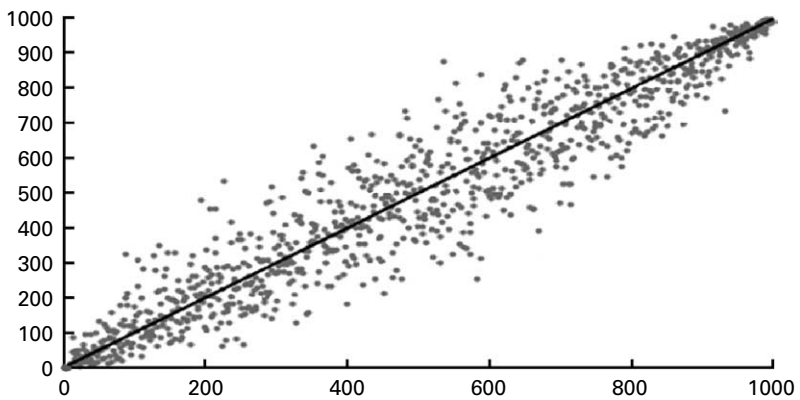

Fig. 7. Initial vs. perturbed ranks for $\alpha = 0.95$

Rank changes reinforce the fact that the greater the dissimilarity the lower the disclosure risk for the disseminated microdata perturbed by this method. Data points that are further away from the sample mean can be more easily identified due to two distinct problems: the classical issue of protection of the value recorded, plus a distance effect, i.e., while perturbed, an observation away from the mean could again face high disclosure risk by still remaining far from it. Changes in ranks circumvent this additional problem. This mechanism is a (welcome) by-product of the present methodology.

Changes in the ranks, however, can also be a drawback, as they will perturb the covariance with other variables. In fact, the lower α is, the lower the correlation between the original and the perturbed variable will be (Table 1); this will also imply higher perturbation of covariance with other variables.

Through its similarity parameter, the univariate multiplicative method presented here allows preserving the covariance with any other variables, but with a trade-off regarding the degree of securization that one wants to achieve in the disseminated data. This trade-off represents an inherent limit to the multiplicative masking structure. For example, one

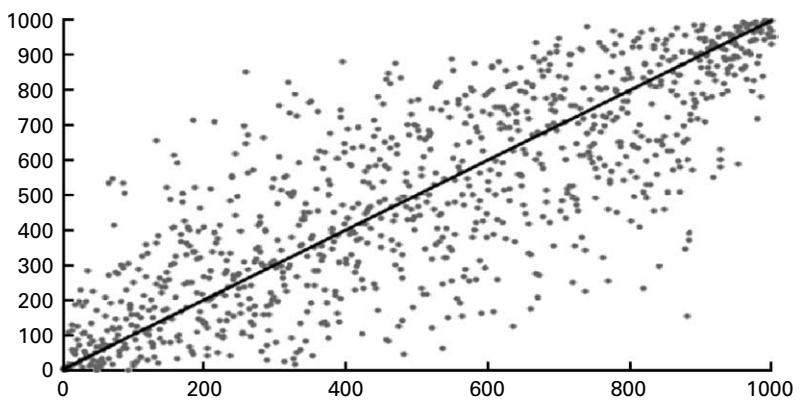

Fig. 8. Initial vs. perturbed ranks for $\alpha = 0.7$

Table 1. Correlation between original and perturbed variable coefficients for different similarity degrees

α	0.999	0.95	0.9	0.8	0.7
Correlation coefficient	0.99	0.60	0.41	0.24	0.18

cannot adapt the perturbation process by introducing a nonconfidential variable in order to preserve exactly some set of covariances: a necessary condition to do that would be that the nonconfidential variable also itself follows a lognormal distribution. But a heavy-tailed nonconfidential variable is a very unlikely configuration. In other cases, the use of the perturbation method with any nonlognormal distribution would induce a distribution of the perturbed variable having a different functional existing in neither exact nor closed form, or being too cumbersome an approximation to be tractable in a simple disclosure control environment (Laeven et al. 2005). This is true at finite distance but one has to note that asymptotic exactness is possible using copula (Sarathy et al. 2002).

4. Further Remarks and Conclusion

When using statistical disclosure control techniques to generate perturbed data, analysis performed on the altered datasets should yield results that are identical or at least very close to those that would have been obtained using the original data. The assumption of normality in the distribution of the original variable and in the error term is a convenient way to achieve this objective. Unfortunately, many economic variables are distributed according to a heavy-tailed, asymmetric form that makes the Gaussian framework limited. Moreover, as underlined by many recent studies (Piketty and Saez 2003), fat tails are important for economic analysis as their impact could be substantial. Nonetheless, one has to note that data points generating a heavy-tailed distribution are often scarcely present in microdata sets, especially those coming from survey-based data (except if specific oversampling procedures are used).

Two reasons account for this under-representation of high values. The first is simply the sampling scheme, as observations away from the mean are less likely to be observed in surveys. The second is that, as observations away from the mean face a higher disclosure risk than data points closer to it, control of these risks forces data producers to rely on top coding, i.e., values above a certain amount are automatically censored to that amount. As a result, sample data skewness is only a partial measure of the true population skewness. In this case, one can still reasonably assume that normality is a sufficient assumption for sample data perturbation, but further research will have to be conducted to determine the relative performances of these additive masking methods when the original data differ from a normal distribution.

The case of register-based microdata is quite different from that of surveys, as entire the population is generally included. In this case, skewness is likely to occur very often, and our methodology will perform better than methods such as the Muralidhar-Sarathy hybrid generator. Moreover, as only heuristic rules are possible in practice for preserving covariances (one being, for example, choosing a degree of similarity between 0.99 and 0.95 that will protect observations away from the mean while preserving sufficiently the covariance), register-based data are favoured; due to their nature and the fact that they are

not originally collected for analysis purpose, fewer variables are available than in a survey for covariance computations.

In conclusion, this article has presented a simple technique that allows data producers to generate perturbed datasets according to a selectable degree of similarity when the underlying distribution is positively skewed, using the properties of a lognormal distribution. The range of applications for this technique is potentially large, particularly when one is interested in the descriptive features of a distribution. For example, this method avoids the use of interpolation for the computation of inequality as practised in Atkinson et al. (2010). For a low value of the dissimilarity parameter, administrative records could easily be made available as public use files. As argued by Sen and Foster (1997) in the case of income distribution: "The log-normal form gives good fits for many countries, though for high levels of income as such the best fits often seem to take the Pareto-form." This means that lognormal distributions, while useful and reasonable approximations, do not always conform to the heavy tails observed for some economic variables. However, transformations of distributions exhibiting more skewness than the lognormal form such as Pareto can only be achieved through approximation, rendering a disclosure control framework based on them intractable in practice. What is sure is that more research is needed in this field. Because of the growing demand for microdata access, the simple methodology presented here could provide a useful starting point upon which more refined masking techniques preserving skewness could be built.

5. References

Atkinson, A.B., Piketty, T., and Saez, E. (2010). Top Incomes in the Long Run of History. Journal of Economic Literature. Forthcoming.

Brand, R. (2002). Microdata Protection through Noise Addition. Inference Control In Statistical Databases, LNCS 2316, 97-116. Berlin Heidelberg: Springer-Verlag.

Burridge, J. (2003). Information Preserving Statistical Obfuscation. Statistics and Computing, 13, 321–327.

Domingo-Ferrer, J. and Gonzalez-Nicolas, U. (2010). Hybrid Microdata Using Microaggregation. Information Sciences, 180, 2837–2844.

Fuller, W.A. (1993). Masking Procedures for Micro-data Disclosure Limitation. Journal of Official Statistics, 9, 383–406.

Gao, X., Xu, H., and Ye, D. (2009). Asymptotic Behaviour of Tail Density for Sum of Correlated Lognormal Variables. International Journal of Mathematics and Mathematical Sciences, 2009, 1–28.

Höhne, J. (2004). Varianten von Zufallsüberlagerung. Working Paper of the Project Group "De Facto Anonymisation of Business Microdata". Wiesbaden. [In German].

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J., Schulte Nordholt, E., Seri, G., and De Wolf, P.-P. (2010). Handbook on Statistical Disclosure Control. ESSNet SDC.

Kim, J.J. and Winkler, W.E. (2001). Multiplicative Noise for Masking Continuous Data. Proceedings of the American Statistical Association, Section on Survey Research Methods, CD-ROM.

Kleber, C. and Kotz, S. (2003). Statistical Size Distributions in Economics and Actuarial Sciences. Wiley Series in Probability and Statistics. Wiley-Interscience.

Krishnamoorthy, K. (2006). Handbook of Statistical Distributions with Applications. Statistics: Textbooks and Monographs. Chapman & Hall/CRC.

Laeven, R.J.A., Goovaerts, M.J., and Hoedemakers, T. (2005). Some Asymptotic Results for Sums of Dependent Variables, with Actuarial Applications. Insurance: Mathematics and Economics, 37, 154–172.

Lydall, H.F. (1966). The Structure of Earnings. Oxford: Clarendon Press.

Mostafa, M.D. and Mahmoud, M.W. (1964). On the Problem of Estimation for the Bivariate Lognormal Distribution. Biometrika, 51, 522–527.

Muralidhar, K., Batra, D., and Kirs, P.J. (1995). Accessibility, Security, and Accuracy in Statistical Databases: The Case for the Multiplicative Fixed Perturbation Approach. Management Science, 41, 1549–1564.

Muralidhar, K. and Sarathy, R. (2005). An Enhanced Data Perturbation Approach for Small Data Sets. Decision Sciences, 36, 513–529.

Muralidhar, K. and Sarathy, R. (2008). Generating Sufficiency-based Non-synthetic Perturbed Data. Transactions on Data Privacy, 1, 17–33.

OECD (2010). Report to the MacArthur Foundation: Feasibility Study of a Harmonised Access to Labour Force and Migration Statistics (Phase I). OECD.

Oganian, A. and Karr, A.F. (2011). Masking Methods that Preserve Positivity Constraints in Microdata. Journal of Statistical Planning and Inference, 141, 31–41.

Piketty, T. and Saez, E. (2003). Income Inequality in the United States 1913–1998. The Quarterly Journal of Economics, CXVIII, 1–39.

Sarathy, R., Muralidhar, K., and Parsa, R. (2002). Perturbing Non-Normal Confidential Attributes: the Copula Approach. Management Sciences, 48, 1613–1627.

Sen, A. and Foster, J.E. (1997). On Economic Inequality. Oxford: Oxford University Press.

Received February 2011
Revised September 2011