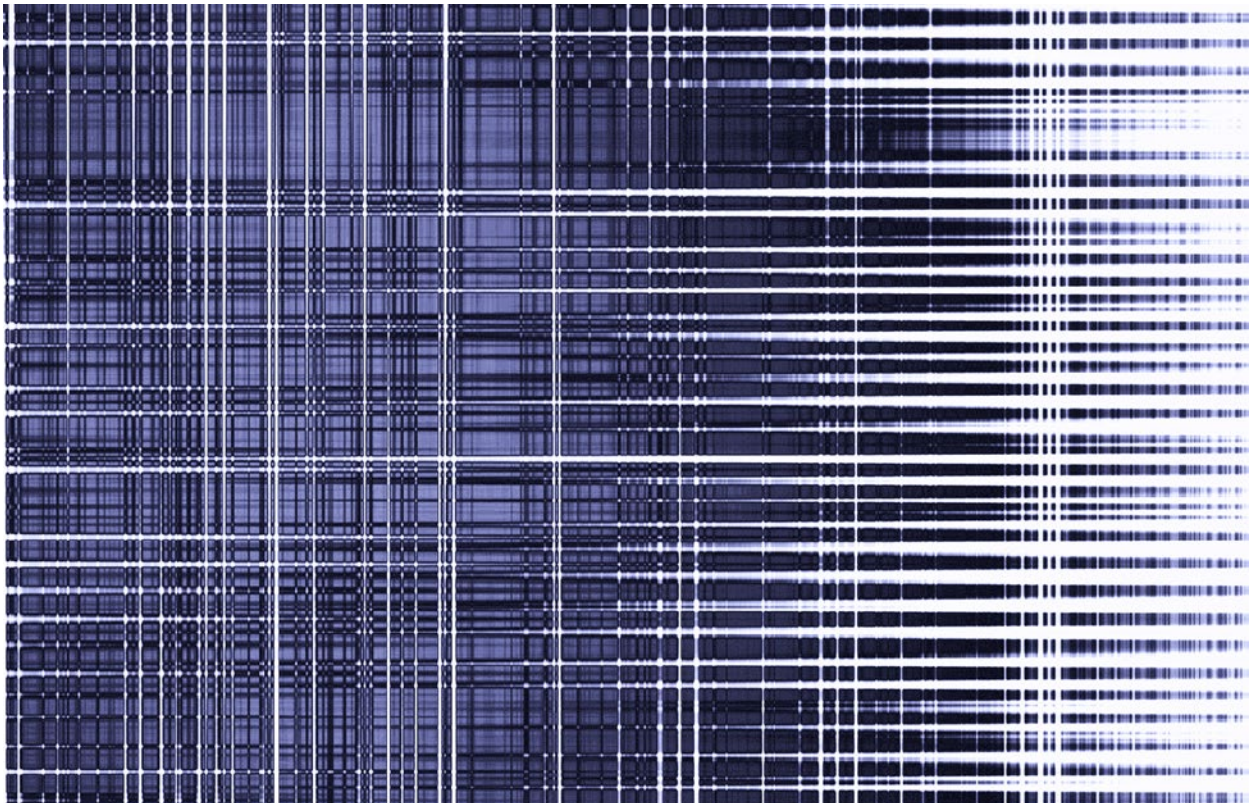
Australian National University

CENTRE FOR SOCIAL RESEARCH & METHODS

Social Research Centre

An ANU Enterprise business



# A universal global measure of univariate and bivariate data utility for anonymised microdata

**S Kocar**

CSRM & SRC METHODS PAPER

NO. 4/2018

# Series note

The ANU Centre for Social Research & Methods (CSRM) was established in 2015 to provide national leadership in the study of Australian society. CSRM has a strategic focus on:

- development of social research methods
- analysis of social issues and policy
- training in social science methods
- providing access to social scientific data.

CSRM publications are produced to enable widespread discussion and comment, and are available for free download from the CSRM website (http://csrm.cass.anu.edu.au/research/publications).

CSRM is located within the Research School of Social Sciences in the College of Arts & Social Sciences at the Australian National University (ANU). The centre is a joint initiative between the Social Research Centre and the ANU. Its expertise includes quantitative, qualitative and experimental research methodologies; public opinion and behaviour measurement; survey design; data collection and analysis; data archiving and management; and professional education in social research methods.

As with all CSRM publications, the views expressed in this Methods Paper are those of the authors and do not reflect any official CSRM position.

**Professor Matthew Gray**
Director, ANU Centre for Social Research & Methods
Research School of Social Sciences
College of Arts & Social Sciences
The Australian National University
August 2018

Methods Paper No. 4/2018

ISSN 2209-184X
ISBN 978-1-925715-09-05

An electronic publication downloaded from http://csrm.cass.anu.edu.au/research/publications.

For a complete list of CSRM working papers, see http://csrm.cass.anu.edu.au/research/publications/working-papers.

**ANU Centre for Social Research & Methods**
Research School of Social Sciences
The Australian National University

# A universal global measure of univariate and bivariate data utility for anonymised microdata


## S Kocar

**Sebastian Kocar** is a PhD candidate at the ANU Centre for Social Research & Methods and a data archivist in the Australian Data Archive at the Australian National University. His research interests include survey methodology, with a particular focus on online data collection, statistical disclosure control and higher education development.


## Abstract


This paper presents a new global data utility measure, based on a benchmarking approach. Data utility measures assess the utility of anonymised microdata by measuring changes in distributions and their impact on bias, variance and other statistics derived from the data. Most existing data utility measures have significant shortcomings – that is, they are limited to continuous variables, to univariate utility assessment, or to local information loss measurements. Several solutions are presented in the proposed global data utility model. It combines univariate and bivariate data utility measures, which calculate information loss using various statistical tests and association measures, such as two-sample Kolmogorov–Smirnov test, chi-squared test (Cramer's *V*), ANOVA *F* test (eta squared), Kruskal-Wallis *H* test (epsilon squared), Spearman coefficient (rho) and Pearson correlation coefficient (*r*). The model is universal, since it also includes new local utility measures for global recoding and variable removal data reduction approaches, and it can be used for data protected with all common masking methods and techniques, from data reduction and data perturbation to generation of synthetic data and sampling. At the bivariate level, the model includes all required data analysis steps: assumptions for statistical tests, statistical significance of the association, direction of the association and strength of the association (size effect).

Since the model should be executed automatically with statistical software code or a package, our aim was to allow all steps to be done with no additional user input. For this reason, we propose approaches to automatically establish the direction of the association between two variables using test-reported standardised residuals and sums of squares between groups.

Although the model is a global data utility model, individual local univariate and bivariate utility can still be assessed for different types of variables, as well as for both normal and non-normal distributions. The next important step in global data utility assessment would be to develop either program code or an R statistical software package for measuring data utility, and to establish the relationship between univariate, bivariate and multivariate data utility of anonymised data.



**Keywords**: statistical disclosure control, data utility, information loss, distribution estimation, bivariate analysis, effect size

# Acronyms

| | |
|---|---|
| ANU | Australian National University |
| CSRM | Centre for Social Research & Methods |
| IL | information loss |
| K–S | Kolmogorov–Smirnov |
| SDC | statistical disclosure control |

# Contents



# Tables and figures

# 1 Introduction

Demand has been increasing for governments to release unit record files in either publicly available anonymised form or restricted-access moderately protected form. Unit record files – which are basically tables that contain unaggregated information about individuals, enterprises, organisations, and so on – are progressively being disseminated by government agencies for both public use (i.e. anybody can access the data) and research use (i.e. only researchers can access the data). These disseminated data files can include detailed information – such as medical, voter registration, census and customer information – which could be used for the allocation of public funds, medical research and trend analysis (Machanavajjhala et al. 2007). Publishing data about individuals allows quality data analysis, which is essential to support an increasing number of real-world applications. However, it may also lead to privacy breaches (Loukides & Gkoulalas-Divanis 2012:9777). Some data are made publicly available, and some data are offered in less safe settings (e.g. to registered researchers on DVDs; see 'five safes framework' in Desai et al. [2016]). For instance, statistical organisations release sample microdata under different modes of access and with different levels of protection. Data may be offered as perturbed public use files, unperturbed statistical data to be accessed in on-site data labs, or even microdata under contract (Shlomo 2010) – also known as scientific use files – with a moderate level of statistical disclosure control (SDC). This means that distributors should focus on data protection. Additionally, unit record files should be more restrictively protected in case they are sensitive. Some variables in datasets can be sensitive as a result of the nature of the information they contain – for example, data on sexual behaviour, health status, income, or if the unit is an enterprise (Dupriez & Boyko 2010).

There is a trade-off between reducing disclosure risk and increasing data utility. We can distinguish between four types of disclosure: identity disclosure, attribute disclosure, inferential disclosure, and disclosure of confidential information about a population or model. SDC is primarily focused on identity disclosure – that is, identification of a respondent (Duncan & Lambert 1989:207–208). On the one hand, releasing no data has zero disclosure risk with no utility; on the other, releasing all collected data including identifying information offers high utility, but with the highest risk (Larsen & Huckett 2010). Most research in the field has focused on the privacy-constrained anonymisation problem – that is, lowering information loss (IL) at certain *k*-anonymity or *l*-diversity values, which is called the direct anonymisation problem. Since this problem can lead to significant IL, data could be useless for specific applications, and the IL could be unacceptable to users (Ghinita et al. 2009). Data producers should also be concerned about data users who would publish an analysis based on statistics in anonymised files that are not similar to corresponding statistics in the original data. In that case, the data producers might have to correct any erroneous conclusions that are reached, and such efforts might require greater resources than those needed to produce data with higher utility (Winkler 1998). The goal is, therefore, to provide data products with reliable utility and controlled disclosure risk (Duncan & Stokes 2004, Templ et al. 2014).

The development of disclosure risk measures has been ahead of the development of data utility measures, mostly as a result of the focus and needs of data disseminators as data protectors, especially when it comes to global measures of those two data protection aspects. Research in the field mostly focuses on developing unbiased approaches to measuring risk of identification, and on investigating which anonymisation methods and techniques are more suitable in different disclosure scenarios (Karr et al. 2006:225). However, the existing data protection

approaches suffer from at least one of the following drawbacks (Ghinita et al. 2009)*:*

- The research has focused on the privacy constraint problem and ignored the accuracy constraint problem.
- The anonymisation is inefficient.
- *L*-diversification causes unnecessary IL.

Very different approaches to SDC modify data in several different ways, leading to challenging disclosure risk and data utility assessments. Synthetic data generation creates a new data matrix with information about a fictitious software-generated population; sampling reduces the size of the sample by multiple times; other data reduction techniques remove variables or recode their values; and data perturbation methods replace original variable values with values generated by applying different perturbation techniques (International Household Survey Network 2017). As a result, protected matrices of different sizes and variables with different ranges of values require the use of adjusted measures of disclosure risk and data utility. Although there are generally accepted and widely used global measures of disclosure risk, mostly calculated as sums of individual per-record risks, data utility assessment is mainly based on estimations of changes in distributions and relations between protected variables (see Templ et al. 2015:28–30), which makes the assessment significantly more local (selected variables) than global (dataset). Methods such as distance measures, distribution estimation, cluster analysis and propensity score can be used to measure reduced utility, but we argue that IL and local (item) utility assessments as local measurements should be combined into data utility assessment as global measurements. The aim of this paper is to propose new solutions for global measurements of data utility for anonymised microdata.

Section 2 of this paper reviews disclosure risk measures, existing IL measures, and how to balance risk and data utility. Section 3 proposes a new user-centred global data utility measure, its structure and the included individual local utility measures. Section 4 establishes which scenarios the proposed model should be applied in, and what future research and development should focus on.

# 2 Balancing disclosure risk and data utility

To release safe data products with high enough data utility for a particular use, a holistic approach to SDC is preferred (Shlomo 2010). Government agencies in Australia are required by law (e.g. the *Privacy Act 1988*) to protect identities of their responding units. Therefore, their primary focus is on reducing disclosure risk. However, data utility assessment should be considered as equally important, because agencies release data to be used by as many (safe) users as possible, and want only unbiased results to be reported. Data utility assessment is especially important because several different data protection solutions are usually possible for the same unit record file – that is, there are different choices of SDC methods and their associated parameters. The trade-off between contradictory goals of decreasing risk and increasing utility involves choosing among different versions of information released (Cox et al. 2011:161). A risk–utility (R-U) confidentiality map (see Duncan & Stokes 2004) can be created to establish which of the solutions represent the best balance between privacy and accuracy. Winkler (1998) argues that users of released files are concerned with their analytic validity: a file is analytically valid if it preserves means and covariances on a small set of subdomains, a few margins and at least one other distributional characteristic. This section reviews both disclosure risk and IL measures as fundamentals for the development of a global data utility measure, which could be used with a global disclosure risk measure in a global R-U confidentiality map.

## 2.1 Disclosure risk measures

The risk of disclosure is a function of the population, as well as the sample. In particular, it is dependent on sample uniques – that is, sample units that are unique within the sample file, where uniqueness is determined by combinations of identifying discrete key variables (Shlomo 2010). Several data characteristics can make potential intruders' jobs easier, including geographical detail, outliers, many attribute variables and census data. Longitudinal or panel data also represent substantial disclosure risk (Duncan & Stokes 2004). The measures for calculating this risk are instruments for estimating the identity disclosure. There are two types of measures: individual per-record and global per-file disclosure risk measures (Shlomo 2010).

Individual measures assess the probability of re-identification of a single responding unit. Methods for assessing individual disclosure risk for sample microdata are generally classified into three types: heuristic, probabilistic record linkage and probabilistic modelling of disclosure risk. Since there is no framework for obtaining consistent global-level disclosure risk measures in the case of heuristics and record linkage approaches, probabilistic modelling seems like the most optimal approach for global disclosure risk assessment (Shlomo 2010). Spruill (1982; cited in Duncan & Lambert 1989) was one of the first to propose a measure of disclosure risk, which was a percentage of 'protected' records closer to their parent record than to any other source record, with the distance computed as a squared distance between the same record in the original and the protected data. Duncan and Lambert (1989) proposed disclosure assessment formulas, which are predictive distributions, for three different scenarios: release to a naive outsider, release to an insider with complete knowledge and release of masked data to an intruder with some knowledge. The last scenario seems to be the only realistic one, since even insiders do not have exact knowledge of all attributes for all units and may not believe respondents entirely. The predictive distributions represent probabilities that any released record is correctly linked, and the formulas take into account the probability that the target record has been released, the number of respondents with the same attributes and the predictive contribution on how the target attributes will appear if released (Duncan &

Lambert 1989). Modern measures for disclosure risk are often based on minimal sample uniques. One of these measures is sample frequencies on subset (SUDA2), a recursive algorithm that generates all possible key variable subsets and scans for uniques patterns (Manning et al. 2008). The calculated risk depends on two aspects and is higher for a larger number of minimal sample uniqueness contained within an observation, or a lower number of variables to determine uniqueness. The risk is calculated as:

$$l_i = \prod_{k=MSUmin_i}^{m-1} (m-k), \qquad i = 1,\ldots,n$$

In this equation, $m$ corresponds to the maximum size of variable subsets, $MSUmin_i$ is the number of minimal sample uniques, and $n$ is number of data file units (Templ et al. 2015).

Global risk measures are aggregated per-record risk measures. They represent the disclosure risk for the entire data file. There are three common disclosure risk measures at the global level: number of sample uniques that are at the same time population uniques, expected number of correct matches/re-identifications (Shlomo 2010:75–76), and number of observations with individual risks higher than a threshold (Templ et al. 2014). Global risk can also be measured using log-linear models (Templ et al. 2015).

## 2.2 Information loss measures

Data utility is based on whether data users can carry out statistical inference and the same analysis on the anonymised data as on the original data. Proxy measures have already been developed to assess data utility, mostly based on measurements of distortions to distributions and the impact on other analysis tools, such as chi-squared statistics (Shlomo 2010). We could classify IL measures in terms of:

- approaches to measuring IL
  - direct distance metrics approach
  - benchmarking approach (Templ et al. 2014)
- type of analysis
  - univariate (e.g. comparing means)
  - bivariate (e.g. comparing measures of association)
  - multivariate (e.g. comparing regression analysis R2) (see Domingo-Ferrer et al. 2001, Shlomo 2010)
- types of variables
  - IL measures for continuous variables
  - IL measures for categorical microdata (Domingo-Ferrer et al. 2001).

Distance metrics are based on measuring distortions to distributions. One of them is average absolute distance per cell ($AAD$), introduced by Gomatam and Karr (2003; cited in Shlomo 2010). It is based on the average absolute difference per table cell in the data and is defined as:

$$AAD\left(D_{\text{orig}}, D_{\text{pert}}\right) = \sum_c \frac{\left|D_{\text{orig}}(c) - D_{\text{pert}}(c)\right|}{n_c}$$

$D_{\text{orig}}$ and $D_{\text{pert}}$ are frequency distributions produced from the microdata, $D(c)$ are frequencies in cell $c$, and $n_c$ is the number of cells. The other common measure, which also takes into account variability of data, is $IL1s$, proposed by Yancey, Winkle and Creecy (2002; cited in Templ et al. 2014). It is defined as:

$$IL1s = \frac{1}{pn} \sum_{j=1}^{p} \sum_{i=1}^{n} \frac{\left|x_{ij} - x'_{ij}\right|}{\sqrt{2} S_j}$$

where $x$ is the original dataset, and $x'$ is the perturbed version of the dataset; there are $n$ records and $p$ variables; and $S_j$ is the standard deviation of the $j$th variable. Some other common distance metrics are a direct comparison of categorical values (for categorical variables), and mean square error, mean absolute error and mean variation (for continuous variables). The listed IL measures for continuous variables are used to compare matrices, covariance matrices (distance metrics), and averages and correlation matrices (benchmarking) (Domingo-Ferrer et al. 2001).

The benchmarking approach is based on comparing statistics computed on the original and perturbed data. At the univariate level, these are statistics such as point estimates, variances or confidence intervals (Templ et al. 2015). At the bivariate level, statistics produced by tests such

as the chi-squared test or ANOVA (goodness-of-fit criterion $R^2$) are compared. The IL measure, based on the chi-squared test and the measure of association between two categorical variables called Cramer's V ($CV$), is often used ($D_{\text{orig}}$ and $D_{\text{pert}}$ are frequency distributions produced from the microdata):

$$RCV\left(D_{\text{pert}}, D_{\text{orig}}\right) = 100\ \frac{CV\left(D_{\text{pert}}\right) - CV(D_{\text{orig}})}{CV(D_{\text{orig}})}$$

For continuous variables, a comparable measure can be used to assess the impact on correlation. Similarly, statistics produced using regression analysis (coefficients, $R^2$) for multiple continuous variables and log-linear modelling for multiple categorical variables could be used.

Global IL metrics are less common than individual item, bivariate or multivariate loss metrics. One of these is global certainty penalty (GCP), introduced by Ghinita et al. (2009), who adopted the normalised certainty penalty (NCP) concept of Xu et al. (2006). NCP measures the amount of distortion introduced by generalisation and is an IL measure for a single equivalence class (variable level, univariate). GCP combines these measures in an IL measure for the entire data file. However, this measure works better with data that include continuous variables only, since distance is often not well defined for categorical variables (Xu et al. 2006). Woo et al. (2009) introduced a set of global utility measures:

- propensity score measure
- cluster analysis measure
- empirical cumulative distribution function measures (using Kolmogorov–Smirnov-type statistics).

Propensity score measures are based on the idea that, if two large groups have the same distributions of propensity scores, they have a similar distribution of covariates. First, original and anonymised data are merged. Propensity scores – that is, the probability of a unit being in the masked dataset – are then calculated. Last, distributions of propensity scores, estimated via logistic regression, are compared.

The similarity of propensity scores, which is, in the end, a data utility measure, can be calculated – for example, by comparing percentiles in each group (masked and original observations):

$$U_p = \frac{1}{N}\sum_{i=1}^{N}[\widehat{p}_i - c]^2$$

where $\widehat{p}_i$ is the estimated propensity score, $N$ is the number of records, and $c$ is the proportion of units with masked data in the merged data file.

Cluster analysis measure is based on the multivariate unsupervised machine-learning method known as cluster analysis, which is carried out with a fixed number of groups on a merged data file including original and protected data. The following equation is used to calculate the utility measure:

$$U_p = \frac{1}{G}\sum_{j=1}^{G} w_j \left[\frac{n_{jo}}{n_j} - c\right]^2$$

where $G$ is the number of groups, $w_j$ is the cluster $j$ weight, $n_{jo}$ is the number of observations from the original data, $n_j$ is the number of units in the $j$th cluster, and $c$ is the proportion of original data in the merged dataset.

# 3 A new user-centred global data utility measure

This section introduces a new global data utility measure, combining univariate distribution-based measures and bivariate measures, calculated from changed coefficients of association between all variables in original and anonymised data files. The model should be used as a relative measure (for comparing global utility of the same unit record file protected in alternative ways) and not as an absolute measure (for comparing the global utility of different data files). This is because global data utility is highly dependent on ranges of mostly demographic variables, as well as other identifying variables, sizes of samples and characteristics of units with an increased identification risk.

To create a global measure, individual (local) measures are also proposed. Based on review of the literature, we concluded that most local and global data utility measures have significant shortcomings. Some can be used with continuous variables only; some are only suitable for univariate utility assessment; some cannot be used with significantly changed sizes of anonymised data matrices or altered measures of variables; and the remainder are more theoretical than user centred.

Generally, there are two main approaches to measuring IL and data utility: direct measures of distances and the benchmarking approach. The latter is generally considered a better solution (Templ et al. 2015). Since there are very different data protection methods and techniques, which produce data matrices of different sizes and include variables with different ranges of values (e.g. bracketing, synthetic data, sampling), unit records may not be the same, and variables might not be of the same type in original and protected data. Consequently, direct measures of distances between original and anonymised data often should not or cannot be calculated. Our proposed user-centred global data utility model is, therefore, based on benchmarking indicators, comparing item distributions and bivariate statistics calculated using original and anonymised data. The idea is that users of protected data should be provided with access to variables with similar (joint) distributions to those in original, non-anonymised versions of the data. In practice, this means that the same or at least similar results would be reported in their research publications as are reported by data producers with access to the original data, even though some characteristics of certain units in the data might differ substantially as a result of SDC procedures applied.

Information loss and overall data accuracy are quite challenging to quantify, partly because of different data masking techniques resulting in different changes to data, but mostly because of users focusing on different research topics and conducting different analyses of the same data. Consequently, we argue that purposely selecting the most important estimates/variables, proposed by other authors (e.g. Templ 2011), could result in accurate measurements of local data utility and IL, but potentially biased estimates of global data utility. Unless the research is focused almost exclusively on measuring a particular concept (e.g. economically active population or unemployment rate in the Labour Force Survey), global data utility measurement should be extended to the majority of, if not all, variables in the data.

## 3.1 Structure of the utility model

The model consists of aggregated local utility measures, which are combined from univariate local item utility measures and bivariate pairs of items utility measures. These measures could potentially be weighted, based on the level of data analysis that would primarily be carried out in secondary use of data – either simple public use (mostly calculating descriptive

statistics) or slightly advanced scientific use of data (calculating both descriptive and bivariate statistics). Creating a global utility measurement follows the principles of creating global risk measurement (see Section 2.1). However, in contrast to the calculation of global risk measurements, which are aggregated record-level risks, global utility measures are aggregated and weighted-average variable-level local utility measures.

This model should be perceived as an expansion of more traditional univariate data utility assessment (e.g. Templ et al. 2015) to the bivariate level, since Lambert (1993) argues that a good anonymisation procedure generally preserves the first two moments of joint distributions, but analysis is not limited to mean and covariance estimations only. On the other hand, the primary focus of our model is not on data utility assessment at the multivariate level, and the model is therefore not particularly suitable for measuring the utility of anonymised data to be used in statistical modelling, although notable changes in univariate and joined distributions generally result in significant changes in multivariate distributions as well. In addition to expanding data utility measurement from the univariate to the bivariate level, we are expanding the benchmarking indicator approach usually applied for continuous variables (Templ et at. 2014) to categorical, including globally recoded variables. Although Xu et al. (2006) listed global recoding as a more severe data reduction technique than local recoding and a source of higher IL, they did not offer any measures for that IL.

The key measures in the global data utility model are the following:

- Average local univariate data utility ($ALDUuni$) is the mean of all local univariate data utility scores ($LDUuni$), calculated for each variable ($k$) separately using approaches and tests described in Sections 3.2 and 3.3.
- Average local bivariate data utility ($ALDUbiv$) is the mean of all ($k$) local bivariate data utility scores, calculated for each variable separately using approaches and tests described in Section 3.4. Each local bivariate data utility score is the mean of all local bivariate data utility measures ($LDUbiv$) – there are $l$ of them per variable, with $l$ being the number of all possible bivariate tests of association (in most cases calculated as $l = k - 1$).
- Global data utility ($GDU$) is a measure combining $ALDUuni$ and $ALDUbiv$ measures (possibly multiplied by coefficients and weights). If univariate utility is equally important to bivariate utility, the following equation should be used (hence the coefficient ½):

$$GDU = \frac{1}{2}ADLUuni + \frac{1}{2}ADLUbiv =$$
$$= \frac{1}{2k}\sum_{i=1}^{k} LDUuni_i + \frac{1}{2k}\sum_{i=1}^{k}\frac{\sum_{j=1}^{l} LDUbiv_j}{l} =$$
$$= \frac{1}{2}\left(\frac{1}{k}\sum_{i=1}^{k} LDUuni_i + \frac{1}{kl}\sum_{i=1}^{k}\sum_{j=1}^{l} LDUbiv_{ij}\right)$$

To calculate local univariate data utility, we have to take into account all the different ways in which data protection affects distributions of protected data. Generally, we propose two different local univariate data measurement approaches: $LDUuni$ for perturbed, suppressed or removed data; and $LDUuni$ for globally recoded data. However, in some cases, a variable can be both globally recoded and suppressed/perturbed.

## 3.2 Local univariate data utility after perturbation, suppression or removal

To create local univariate data utility measures for perturbed, suppressed or removed data (also synthetic data), original and protected data have to be compared at a local level using appropriate statistical tests. Statistical significance ($P$) values returned by these tests indicate the level at which the null hypothesis (i.e. that the two samples were drawn from the same distribution) can be rejected. We argue that local utility, defined as original data local (item) utility minus IL, decreases with a decrease of significance levels of selected distribution comparison tests, such as the two-sample Kolmogorov–Smirnov (K–S) test for univariate continuous variable distributions. Instead of using IL scores with just two possible values – that is, IL = 0 if the two samples were drawn from the same distribution (K–S test significance > 0.05), and IL = 1 if the two samples

were not drawn from the same distribution (K–S significance < 0.05) – we propose to calculate IL from the exponential distribution; an indicator of the protected item not being an accurate measure follows exponential distribution $x \sim \exp(\lambda)$. The probability density function of the exponential distribution is given by:

$$f(x) = \lambda e^{-\lambda x}, \qquad x > 0$$

We have to propose distribution parameters that would return reasonable estimates of IL and data utility at different $P$ values. The scale parameter value should equal 1, which means that, for each variable, IL and local utility values calculated with this equation will always have a value between 0 and 1. For $x$ in the equation, different values have been considered. We suggest $x$ to be 14 times the $P$ value, a statistical significance value of a selected test. With $x = 14P$ and with $P = 0.05$, which is the level traditionally used in hypothesis testing, local data utility value equals IL (0.50). However, at $P = 0.01$, the probability of samples being drawn from the same distribution is much lower; hence local data utility is significantly lower, at 0.13 (see Figure 1). These seem to be quite reasonable estimates of univariate data utility.

The following equation should be used to calculate local univariate data utility after perturbation, suppression or removal ($LDUunipsr$):

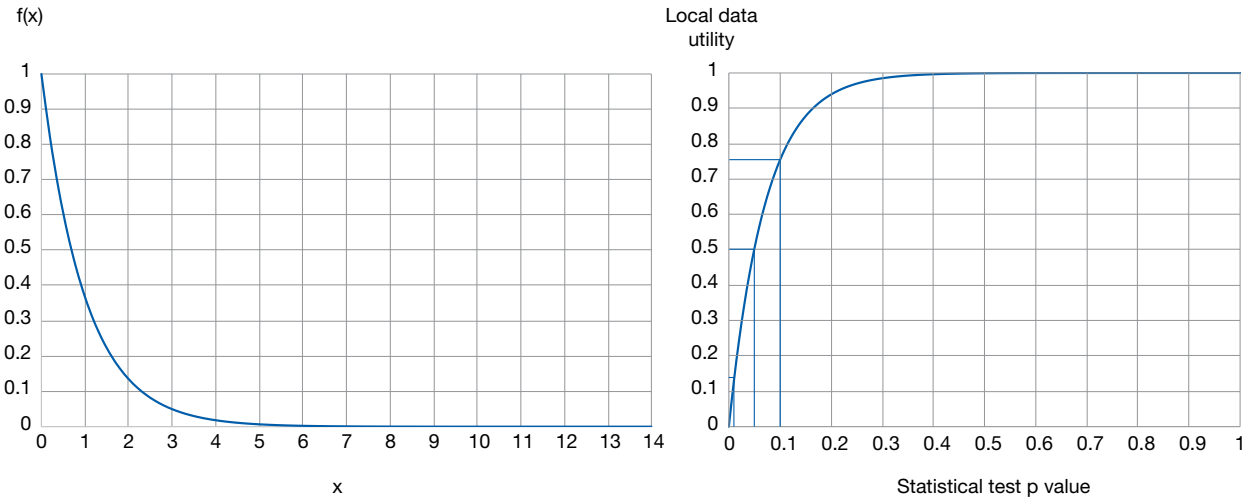
$$LDUunipsr = 1 - e^{-14p}$$

We decided to use this distribution and its parameters to calculate IL, since very little increase in $P$ value is required to retain the null hypothesis (no distribution differences between groups), which means that IL levels should decrease very quickly. The exponential distribution is a continuous probability distribution, applied to a continuous random variable with sets of possible values that are infinite and uncountable. However, in our case, the density value for $x$ values between 0.4 and 1 is <0.01, and IL is negligible. Therefore, we can use it in practice despite having a defined range of variable $x$ values (see Figure 1).

To compare data and calculate $P$ values, different univariate statistical tests should be used. Since there are different types of variables with different distributions, various tests have been considered. The following tests are used in our model at the univariate level:

- For univariate-level interval and ratio variables, nonparametric two-sample K–S test (Massey 1951) for equality of distributions for continuous variables should be used to determine significance levels. The distributions of original and protected variables are compared, given that the protected continuous variables have not been globally recoded into categorical variables or bottom/top coded. Two-sample *t* test as a parametric version of the K–S test could potentially be used. However, it is not sensitive to variability and

**Figure 1** **Exponential distribution of information loss, the function of local data utility**



would not be a suitable solution for comparing non-normal distributions; we would prefer to use a universal distribution-free test. We could theoretically use the paired-samples version of the test. However, it focuses on distances between variable values of the same units, and therefore cannot be used for all data protection methods and techniques (e.g. sampling – data reduction).

- For univariate-level categorical variables (nominal and ordinal), chi-squared test seems to be the most popular and reasonable solution. The test would compare discrete distributions of original and protected variables as 'groups' and report $P$ values.

## 3.3 Local univariate data utility after global recoding

There are several existing distance measures for perturbed or suppressed categorical variables (see Domingo-Ferrer et al. 2001:107–109). However, developing an unbiased local utility measure for globally recoded variables is more challenging because collapsing of variable values changes their ranges or even their types. Therefore, we have to consider how these kinds of data reduction techniques influence the analysis of data. After the initial collapsing of values, we also have to pay attention to distributions of already globally recoded variables, since aggregation is often combined with other data reduction techniques, such as local suppression, or data perturbation techniques, such as data swapping.

In this paper, we consider three different sources of local IL after global recoding:

- limited comparison of groups' statistics (categorical → categorical variable)
- potentially biased allocation of values into value groups (continuous → categorical variable)
- increased heterogeneity of groups (continuous → categorical variable).

Consequently, we propose three different measures of data utility after global recoding; some are more rigorous and some slightly less, and users need to select the most suitable measure for their case.

First, we have to consider how much we limit comparisons between groups of categorical variables due to the reduction in the number of categories. For example, we might want to mask an age-group variable with ten 10-year age groups (0–9, 10–19, …, 90–99 years), recoding it into a variable with five 20-year age groups (0–19, …, 80–99 years). In practice, this means that comparisons of proportions of people aged 20–29 with those aged 30–39, as well as calculation of population estimates for these two groups (individually), would no longer be possible. In this specific case, the total number of comparisons between pairs of groups $(n(n-1)/2)$ decreases from 45 to 10, by almost 78%. We argue that local univariate data utility after global recording ($LDUnigr$) should decrease in the same way. The following formula for $LDUnigr$ should be used in the model:

$$LDUunigr = \frac{(n-m)(n-m-1)}{n(n-1)}$$

In the equation, $n$ represents the original number of variable values, and $m$ represents the reduced number of variable values in the protected data.

Second, when values of continuous variables are collapsed into groups, they might end up in groups with less similar values (e.g. as a result of allocation). One way to measure this source of IL is to calculate the percentage of units that are closer to any unit in the neighbour groups than to any unit in their own group (e.g. age value 39 in the 30–39 age group is closer to value 40 in the 40–49 age group than to value 33 in the 30–39 age group). The following formula for $LDUnigr$ is proposed:

$$LDUunigr = 1 - \frac{n_{ba}}{n}$$

In the equation, $n$ represents the number of all units, and $n_{ba}$ represents the number of all biased allocations at collapsing of values in anonymised data. Generally, the fewer groups are created, the higher the chance of biased allocations due to an increased heterogeneity of groups.

Third, we introduce another group heterogeneity-based measure of local data utility after global

recoding. When we collapse values into a group, data analysis treats all these values as the same; ideally, this would not be the case. If the number of groups is lower, the groups might be very heterogeneous. To measure heterogeneity, we could calculate average distances – that is, mean absolute deviations (MAD) – from the mean of newly created groups. We would calculate MAD between unit values (before recoding) and corresponding group averages (of groups after recoding), and compare them with the maximum theoretical average distances if all units were in one group only. The following formula for $LDUunigr$ is proposed:

$$LDUunigr = 1 - \frac{n\sum_{j=1}^{p}\sum_{i=1}^{r}(x_{ij} - \bar{x_j})}{s\sum_{i=1}^{n}(x_i - \bar{\bar{x}})}$$

There are $r$ records in the original dataset and $s$ records in the anonymised dataset, with $p$ groups/categories and $r$ units in a group. $\bar{\bar{x}}$ is the grand average (as if there were just one group), while $x_j$ are means of newly created groups in anonymised data, calculated from values in the original data. If no units had been removed and no sampling carried out, then $n = s$.

We also have to take into consideration other possible changes to already recoded data, such as changes after carrying out local suppression or micro-aggregation. Therefore, we have to compare further distributions of recoded variables (original and protected data) using approaches described for nonaggregated univariate-level interval or ratio variables, or categorical variables – that is, two-sample K–S or chi-squared tests ($LDUunipsr$). Local univariate data utility ($LDUuni$) for globally recoded and further protected variables is a product of local item data utility after global recoding ($LDUnigr$) and local item data utility after perturbation ($LDUunipsr$):

$$LDUuni = LDUunigr\ LDUunipsr$$

## 3.4 Local bivariate data utility

To calculate local bivariate data utility ($LDUbiv$) scores, original and protected data can be compared at a local level using benchmarking statistics generated by appropriate bivariate statistical tests. Our local bivariate measures will be based on the Cramer's *V* IL measure (see the equation in Section 2.2), proposed by Shlomo (2010). Cramer's *V* is a chi-squared-based measure of nominal association for tables sized 2 × 2 and larger. For measures of ordinal, interval and ratio associations/effect sizes, other statistical tests reporting coefficients measuring the strength of the relationship between two variables should be applied. Table 1 lists statistical tests and association measures for pairs of variables based on their type and normality of the distributions (McDonald 2009:308–313; Fritz et al. 2012).

**Table 1** **Bivariate tests and statistics used for local bivariate data utility calculation**

| Type of variables | Nominal | Ordinal | Non-normally distributed interval/ratio | Normally distributed interval/ratio |
|---|---|---|---|---|
| Nominal | Chi-squared test (Cramer's *V*) | – | – | – |
| Ordinal | Kruskal–Wallis *H* test (epsilon squared) | Spearman correlation (Spearman's rho) | – | – |
| Non-normally distributed interval/ratio | Kruskal–Wallis *H* test (epsilon squared) | Spearman correlation (Spearman's rho) | Spearman correlation (Spearman's rho) | – |
| Normally distributed interval/ratio | ANOVA *F* test (eta squared) | Spearman correlation (Spearman's rho) | Spearman correlation (Spearman's rho) | Pearson correlation (Pearson's *r*) |

In total, five different bivariate tests and corresponding association measures should be included in our data utility model, in addition to an optional one-sample K–S test for establishing whether a continuous variable is normally distributed. ANOVA *F* test and Kruskal–Wallis *H* test could be used for measuring the association between one interval/ordinal and one nominal variable with only two categories, although *t* test and Mann–Whitney test would generally be considered as the primary test options. When performing bivariate analysis tests, the following should generally be explored: assumptions for statistical tests, statistical significance of the association, direction of the association and strength of the association (effect size) (Page 2014). Therefore, in all cases, after the test has been selected and the results obtained, the following three steps should be followed to calculate local bivariate data utility:

Step 1: Establish whether significance levels for bivariate statistical tests are above a threshold ($P_t$; e.g. $P \geq 0.05$ or $P \geq 0.01$) for the same pairs of variables in original and protected data:

- If both $P$ values are above the threshold, there is no association in either case, and $LDUbiv = 1$.
- If one $P$ value is above and the other below the threshold, there is a significant difference in associations, and $LDUbiv = 0$.
- If both $P$ values are below the threshold, there are statistically significant associations, and further review of the results is required (continue with the next step).

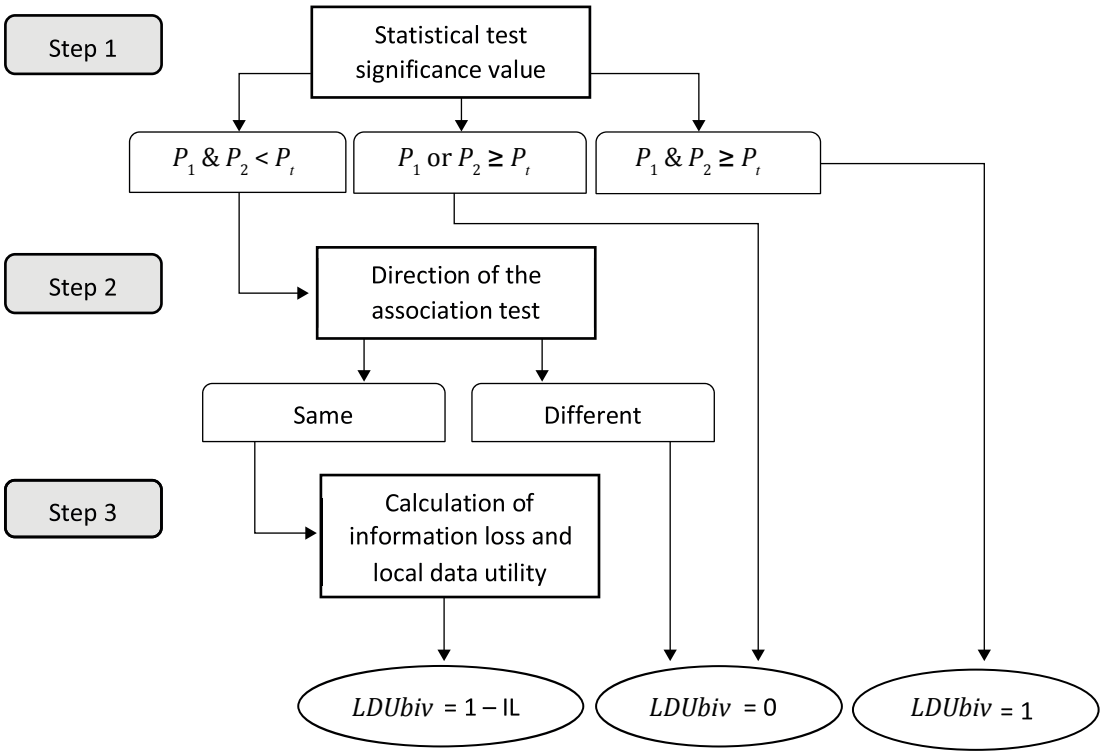
Step 2: Establish whether the direction of the association is the same for pairs of variables in original and protected data (comparing signs of correlation coefficient numbers, standardised residuals of chi-squared tests, the total sum of squares between groups for ANOVA and *H* test – see details below):

- If there are opposite directions, the relationship between variables is significantly different as a result of data protection, and $LDUbiv = 0$.
- If the directions are the same, further review of results is needed (continue with the next step).

Step 3: Calculate IL at the bivariate level using measures of association between two variables, and calculate local bivariate data utility ($LDUbiv$) using IL scores.

We can present the procedure in an algorithmic way (Figure 2).

**Figure 2** **Local data utility calculation procedure**



Establishing the direction of association, and calculating IL and local bivariate data utility using statistics of different bivariate tests are described below.

### 3.4.1 Tests for pairs of variables including at least one nominal variable – chi-squared test, Kruskal–Wallis *H* test, ANOVA *F* test

To establish whether the direction of the association is the same for pairs of variables in original and protected data (step 2), we should review and compare measures of the strength of the difference between either:

- observed and expected values, or
- group means and the grand mean, or
- group average ranks and the grand average rank.

The measures used in our model are chi-squard test standardised residuals, ANOVA *F* test sums of squares between groups, and Kruskal–Wallis sums of squares between groups (from average ranks). We prefer to introduce solutions that reduce the amount of manual input, reduce the need for manual review of bivariate results, and calculate IL and data utility automatically (e.g. using R code or an R package). One of the solutions is to automatically establish the change in those measures of the strength of the difference. In general, if the direction of the association between two categorical variables changes in the protected data, standardised residuals ($SR$) change sign. Although it is straightforward to notice a difference in 2 × 2 tables, larger contingency tables are more difficult to review manually or automatically. At least four approaches can be used to investigate a statistically significant chi-squared test result: calculating residuals, comparing cells, ransacking and partitioning (Sharpe 2015). Calculating standardised residuals seems to be the best solution for automatic estimation (or detection of a change) of the direction of the association. The residuals are calculated for a single cell in a contingency table as:

$$SR_i = \frac{O_i - E_i}{\sqrt{E_i}}$$

where $O_i$ is the observed value, and $E_i$ is the expected value in the cell. On the other hand, ANOVA *F* test and Kruskal–Wallis *H* test sums of squares between groups ($SSB_i$) for individual groups, not a total sum of squares between for all groups ($SSB$), should be compared. ANOVA *F* test sums of squares between groups (contributions of individual groups) are calculated with an adjusted Fisher's (1925) equation for total $SSB$ (removing the summation notation):

$$SSB_i = n_i(\bar{x}_i - \bar{\bar{x}})^2$$

where $\bar{x}_i$ is the mean of the group, $n$ is the number of units in that group, and $\bar{\bar{x}}$ is the grand mean. For the Kruskal–Wallis test, the approach is based on the same principles of $SSB_i$ calculation for normally distributed continuous variables; ranked variables with group means ($\bar{x}_i$) are replaced with average group ranks ($\bar{r}_i$), and grand means ($\bar{\bar{x}}$) are replaced with grand average rank in the equation ($\bar{\bar{r}}$):

$$SSB_i = n_i(\bar{r}_i - \bar{\bar{r}})^2 = n_i\left(\bar{r}_i - \frac{N+1}{2}\right)^2$$

where $n_i$ is the number of units in that group, and $N$ is the total number of units in all groups.

For the calculation of the direction, it is important to note that:

$$\text{if } \bar{x}_i \leq \bar{\bar{x}} \text{ then } SSB_i \leq 0$$

$$\text{if } \bar{r}_i \leq \bar{\bar{r}} \text{ then } SSB_i \leq 0$$

This means that we have to change the sign of the $SSB$ value in the equation below if the mean rank of the group is lower than the grand mean rank. For standardised residuals, there is no need for the change of sign, since they are already negative for cells with observed values lower than expected values. Taking all those conditions into account, we propose to calculate the possible change of the sign in the following way:

$$\sum_{i=1}^{n} \left| MSD_i(D_{\text{orig}}) - MSD_i(D_{\text{prot}}) \right| - \sum_{i=1}^{n} \left| MSD_i(D_{\text{orig}}) \right|$$

Measures of the strength of the difference ($MSD$) used in the formula are, as explained above, either standardised residuals ($SR_i$) or sums of squares between groups ($SSB_i$ for either ANOVA *F* test or Kruskal–Wallis *H* test). In the equation, $n$

stands for either the number of contingency table cells (chi-squared test) or the number of all one-way compared groups (ANOVA, Kruskal–Wallis). The calculation is based on the following:

- If all measures of the strength of the difference values in protected data $D_{\text{prot}}$ equalled 0 (no association between the two variables), the expression above would also equal 0.
- If measures of the strength of the difference values in the protected data were closer to the original $D_{\text{orig}}$ values than 0, the expression above would be negative (the same direction for the association).
- If measures of the strength of the difference values in the protected data changed the sign for most cells/groups and/or higher $SR/SSB$ values, the expression above would be positive (changed direction of the association).

To calculate IL in step 3, Cramer's *V* IL measure proposed by Shlomo (2010) should be adjusted. Since IL and local data utility in our model can take values between 0 and 1, the multiplier 100 should be removed. Also, the Cramer's *V* for the protected data, as well as epsilon squared and eta squared, could theoretically be higher than the original coefficients of association. Therefore, the difference between the coefficients in the denominator should be absolute. Although it is not very likely, it should still be noted that the theoretical value of the IL score could be higher than 1 if the original data ($D_{\text{orig}}$) coefficient of association were much lower than the protected data ($D_{\text{prot}}$) coefficients (e.g. 0.5 and 0.2), using the proposed equation. Therefore, the higher of the two coefficients should be in the denominator. Last but not least, we decided to use squared coefficients of association in the equation for consistency, since some size-effect measures are already squared (eta and epsilon). This means that IL increases more quickly with an increase in the difference between the coefficient values; in our opinion, this better reflects the decreased local bivariate data utility. The adjusted equation for IL and local bivariate data utility ($LDUbiv$) is:

$$LDUbiv = 1 - IL = 1 - \frac{\left|coeff^2(D_{\text{prot}}) - coeff^2(D_{\text{orig}})\right|}{\max coeff^2(D_{\text{prot}}), coeff^2(D_{\text{orig}})}$$

We can use this formula to measure IL with three different squared coefficients of association/size effect measures ($coeff^2$) (Fritz et al. 2012):

- Cramer's *V* (squared), calculated as

$$V^2 = \frac{\chi^2/n}{\min(k-1),(r-1)}$$

- epsilon squared, calculated as

$$\varepsilon^2 = H\frac{(n^2+1)}{(n^2-1)}$$

- eta squared, calculated as a ratio between sum of squares between groups ($SSB$) and total sum of squares ($TSS$)

$$\eta_p^2 = \frac{SSB}{TSS}$$

### 3.4.2 Tests for pairs of variables that are at least ordinal – Spearman correlation and Pearson correlation coefficient

To establish whether the direction of the association is the same for pairs of variables in original and protected data, this time no additional tests are needed, and the signs of the correlation coefficient just need to be checked.

To calculate IL and local bivariate data utility for continuous or ranked variables (i.e. ordinal or non-normally distributed continuous variables), the same equation as for IL for other tests in this model should be used, just with squared Spearman $\rho$ or Pearson $r$ correlation coefficients ($corr\_coeff^2$). The following equation for IL and utility is proposed:

$$LDUbiv = 1 - IL = 1 - \frac{\left|corr\_coeff^2(D_{\text{prot}}) - corr\_coeff^2(D_{\text{orig}})\right|}{\max corr\_coeff^2(D_{\text{prot}}), corr\_coeff^2(D_{\text{orig}})}$$

### 3.4.3 Tests for variables with changed type due to data protection

If SDC leads to variables changing their type, either from continuous to categorical or from normally to non-normally distributed, it is more challenging to compare measures of association – that is, effect sizes measured with different statistical tests. An eta squared value represents

a significantly different effect size from the same Pearson $r^2$ value (Vacha-Haase & Thompson 2004:2–4). Moreover, Cohen (1988:224–227) provided a guide to the magnitude of the effect size for the chi-squared test (Cramer's *V*: 0.1 = small effect, 0.3 = medium effect, 0.5 = large effect for 2 × 2 contingency tables). However, he noted that the degree of association does not depend on Cramer's *V* coefficient only, but also on the number of cells in contingency tables (degrees of freedom). Therefore, just like at the univariate level, local bivariate data utility for these variables might have to be calculated in an alternative way.

It is easier to compare effect sizes if one or both variables, measured at the continuous level, are recoded to the ordinal level and later perturbed, suppressed or not additionally protected. In that case, local bivariate data utility can be calculated by comparing effect sizes $r^2$ or $\rho^2$ (original data) with $\rho^2$ (protected data) using the $LDUbiv$ equation for pairs of variables that are at least ordinal. Generally, effect sizes decrease with global recoding, and top and bottom coding, and there is very little difference in effect sizes measured with Pearson and Spearman correlation coefficient tests for both normal and skewed distributions.

On the other hand, if one of the variables in the pair is measured at the nominal level, while the other one changes distribution or type (e.g. from interval to ordinal or binary), then eta squared, epsilon squared and Cramer's *V* coefficient values cannot be directly compared using the equations proposed in this paper. In that case, we would have to review how effect sizes of different statistical tests should be compared and interpreted.

## 3.5 Data utility of data with removed variables

Removing variables is considered one of the most extreme measures of data protection, unless only direct identifiers, identification numbers or variables irrelevant for analytical purposes are deleted from disseminated data. Although removing individual cases and its effect on local data utility is already taken into account with local utility calculations for perturbed, suppressed or removed data, removing variables is a special case. The utility of data, calculated with measures presented in Sections 3.1–3.4, decreases linearly with an increase in variables removed, since local univariate data utility and average local bivariate data utility for a removed variable equal 0. However, removing variables means removing any valuable information about a measured concept, results in additional effects on global data utility and should be additionally penalised. As in the case of global recoding (decreasing values/categories), decreasing the number of variables limits bivariate analysis because it reduces the total number of pairs of variables that could be investigated at the bivariate level. A similar equation to the $LDUnigr$ formula for categorical variables after global recoding can be used for this global reduction coefficient ($GRC$):

$$GRC = \frac{n(D_{\text{prot}})\big(n(D_{\text{prot}}) - 1\big)}{n(D_{\text{orig}})\big(n(D_{\text{orig}}) - 1\big)}$$

In the equation, $n(D_{\text{orig}})$ represents the number of variables in the original data, and $n(D_{\text{prot}})$ represents the number of variables in the protected data. If, for example, we removed 5 variables relevant for analytical purposes from a dataset with 100 variables, the $GRC$ would equal 0.902. We would then multiply global data utility with this $GRC$, if required.

## 3.6 Multivariate data utility

Although data utility assessment at the multivariate level is not the primary focus of the proposed global data utility measure, it could nevertheless be integrated into the model under certain conditions. The inclusion of multivariate assessment would contribute to the overall estimation of data quality, since microdata are more often than not used for statistical modelling. However, the most notable issue of an unbiased multivariate global utility assessment is a subjective selection of variables for multivariate analysis. If not all variables of analytical interest are included in the model(s), the assessment is more local than global. Consequently, multivariate data utility could feasibly be extended to the special case of research data that focus almost

exclusively on measuring a particular concept (e.g. Labour Force Survey).

Our adjusted Cramer's *V* IL measure, originally proposed by Shlomo (2010:84), could be used with different multivariate coefficients, such as:

- linear regression adjusted $R^2$ statistic (including all standardised beta coefficients in the model) for normally distributed continuous variables
- logistic regression Cox & Snell pseudo $R^2$ statistic (including all odds ratios) for binary dependent variables, and continuous and binary independent variables
- Cronbach alpha coefficient (including all if-item-deleted coefficients) for correlated continuous variables deriving an index.

Multivariate data utility – both local (see Shlomo 2010) and global (see Woo et al. 2009) – has already been studied in the literature. However, how these approaches compare and how they could be integrated into benchmarking-based global utility models are yet to be investigated. This type of global data utility assessment should therefore be an area of ongoing research and development.

# 4 Discussion

Data protection literature still offers little guidance on how to efficiently measure global data utility. Although much research has been done on how to calculate identification risk and how to protect data to minimise it, less research has looked at how to measure IL at the unit record file level, and how to find the proper balance between global risk and global data utility. In this paper, we propose a solution to an effective global data utility assessment. We prepared a user-centred global data utility measurement model, combining local univariate data utility and local bivariate data utility measures. We argue that utility assessment should be expanded at least to the bivariate level and that all variables, except for weights, other derived variables, and ID numbers, should be included in the estimations.

Since we have proposed several individual steps and individual local measures, this model can be used for either univariate or bivariate utility assessment only, or for local utility measurement focusing on a limited number of distributions and associations. Five different bivariate measures are proposed to calculate local bivariate utility for pairs of variables of different types and distributions. It should be noted that the proposed procedure could theoretically be fully automated using a software environment for statistical computing such as R. For this reason, approaches to automatically establish the direction of association, using standardised residuals and sums of squares between groups, are proposed.

Because the model is based on a benchmarking approach, not on distance measures, it can be used with data protected by sampling or perturbed, and synthetic data techniques. One possible application of the model is first to investigate general global dataset data utility and then to use the results to focus on local IL, using a top-down approach. One important aspect of this model is that certain data utility measurement results (e.g. combinations of variables with less accurate estimations of associations) can be shared with data users, whereas data protection details, based on disclosure risk measurements, often cannot be provided because they have the potential to reveal information that would help intruders to re-create original data and identify individuals. Since this is a global measure, based on univariate and bivariate test-reported statistics (nonindividual records approach), it could be used to compare data for other purposes, such as evaluating different survey data weighting schemes, and comparing data collected with different survey modes or interviewers (investigating measurement errors).

To fully exploit this model, either program code or an R package would have to be developed. This would enable other users to assess data utility with as little effort and manual input as possible. They would only need to provide information about variable types (nominal, ordinal, continuous) in original and protected data, and about performed global recoding and bottom and top coding, and provide both original and protected data. One of the future challenges is to develop unbiased global measures of data utility at the multivariate level for different types of variables, knowing that even less restrictive data protection approaches may result in considerable information loss in complex statistical models. To start with, it would be particularly useful to establish the relationship between local univariate and bivariate data utility scores and local multivariate utility scores for the same range of variables, calculated using the data utility measures proposed in this paper.

# References


Cohen J (1988). *Statistical power analysis for the behavioral sciences*, Lawrence Erlbaum Associates, New Jersey, 2.

Cox LH, Karr AF & Kinney SK (2011). Risk-utility paradigms for statistical disclosure limitation: how to think, but not how to act. *International Statistical Review* 79(2):160–183.

Desai T, Ritchie F & Welpton R (2016). *Five safes: designing data access for research*, working paper, University of the West of England.

Domingo-Ferrer J, Mateo-Sanz JM & Torra V (2001). Comparing SDC methods for microdata on the basis of information loss and disclosure risk. In: *Pre-proceedings of ETK-NTTS*, vol. 2, 807–826.

Duncan G & Lambert D (1989). The risk of disclosure for microdata. *Journal of Business & Economic Statistics* 7(2):207–217.

——— & Stokes SL (2004). Disclosure risk vs. data utility: the R-U confidentiality map as applied to topcoding. *Chance* 17(3):16–20.

Dupriez O & Boyko E (2010). *Dissemination of microdata files: principles, procedures and practices*, International Household Survey Network.

Fisher RA (1925). *Statistical methods for research workers*, Genesis Publishing.

Fritz CO, Morris PE & Richler JJ (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of Experimental Psychology: General* 141(1):2.

Ghinita G, Karras P, Kalnis P & Mamoulis N (2009). A framework for efficient data anonymization under privacy and accuracy constraints. *ACM Transactions on Database Systems* 34(2):9.

International Household Survey Network (2017). Reducing the disclosure risk, www.ihsn.org/anonymization-risk-reduction.

Karr AF, Kohnen CN, Oganian A, Reiter JP & Sanil AP (2006). A framework for evaluating the utility of data altered to protect confidentiality. *American Statistician* 60(3):224–232.

Lambert D (1993). Measures of disclosure risk and harm. *Journal of Official Statistics* 9(2):313.

Larsen MD & Huckett JC (2010). Measuring disclosure risk for multimethod synthetic data generation. In: *Proceedings: SocialCom 2010 – the Second IEEE International Conference on Social Computing*, Minneapolis, 20–22 August, Institute of Electrical and Electronics Engineers, 808–815.

Loukides G & Gkoulalas-Divanis A (2012). Utility-preserving transaction data anonymization with low information loss. *Expert Systems with Applications* 39(10):9764–9777.

Machanavajjhala A, Kifer D, Gehrke J & Venkitasubramaniam M (2007). $L$-diversity: privacy beyond $k$-anonymity. *ACM Transactions on Knowledge Discovery from Data* 1(1):3.

Manning AM, Haglin DJ & Keane JA (2008). A recursive search algorithm for statistical disclosure assessment. *Data Mining and Knowledge Discovery* 16(2):165–196.

Massey FJ Jr (1951). The Kolmogorov–Smirnov test for goodness of fit. *Journal of the American Statistical Association* 46(253):68–78.

McDonald JH (2009). *Handbook of biological statistics*, 2nd edn, Sparky House Publishing, Baltimore.

Page P (2014). Beyond statistical significance: clinical interpretation of rehabilitation research literature. *International Journal of Sports Physical Therapy* 9(5):726.

Sharpe D (2015). Your chi-square test is statistically significant: now what? *Practical Assessment, Research & Evaluation* 20(8):1–10.

Shlomo N (2010). Releasing microdata: disclosure risk estimation, data masking and assessing utility. *Journal of Privacy and Confidentiality* 2(1):73–91.

Templ M (2011). *Estimators and model predictions from the structural earnings survey for benchmarking statistical disclosure methods*, Research Report CS-2011-4, Department of Statistics and Probability Theory, Vienna University of Technology.

——— , Meindl B, Kowarik A & Chen S (2014). *Introduction to statistical disclosure control (SDC)*, IHSN Working Paper 005, International Household Survey Network, http://www.ihsn.org/home/sites/default/files/resources/ihsn-working-paper-007-Oct27.pdf.

——— , Kowarik A & Meindl B (2015). Statistical disclosure control for micro-data using the R package sdcMicro. *Journal of Statistical Software* 67(1):1–36.

Vacha-Haase T & Thompson B (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology* 51(4):473.

Winkler WE (1998). *Producing public-use microdata that are analytically valid and confidential*, Statistical Research Report Series RR98/02, US Census Bureau, Statistical Research Division, Washington, DC.

Woo MJ, Reiter JP, Oganian A & Karr AF (2009). Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality* 1(1):7.

Xu J, Wang W, Pei J, Wang X, Shi B & Fu AWC (2006). Utility-based anonymization using local recoding. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, 20–23 August, Association for Computing Machinery, New York, 785–790.

CENTRE FOR SOCIAL
RESEARCH & METHODS

+61 2 6125 1279
csrm.comms@anu.edu.au

The Australian National University
Canberra ACT 2601 Australia

www.anu.edu.au