A Probabilistic Procedure for Anonymisation and Analysis of Perturbed Datasets

Harvey Goldstein
University of Bristol and University College London
and
Natalie Shlomo
University of Manchester

Abstract

The requirement to anonymise datasets that are to be released for secondary analysis needs to be balanced by the need to allow their analysis to provide efficient and consistent parameter estimates. The proposal in the present paper is to use the addition of random noise to some or all variables in a released (already pseudonymised) data set where the values of some identifying variables for individuals of interest are also available to an external 'attacker' who wishes to identify those individuals so that they can interrogate their records in the dataset. To avoid such identification enough noise needs to be generated and added to these identifying variables. The noise so generated then needs to be accounted for at the analysis stage to provide required parameter estimates. Where the characteristics of the noise are made available to the analyst by the agency providing the data, we propose a method that allows a valid analysis. This is formally a measurement error model and there exist procedures for model fitting that recovers consistent estimates of the true model parameters. The paper shows how an appropriate noise distribution can be determined and at the analysis stage describes a Bayesian MCMC algorithm that allows for noise removal.

Keywords

Additive noise, Anonymisation, Measurement error, Record linkage.

Address for correspondence

Professor H. Goldstein
Graduate School of Education
University of Bristol
Bristol, BS8 1JA
UK
h.goldstein@bristol.ac.uk

Number of words excluding abstract, tables, references, acknowledgement and appendix: 9659

1. Introduction

Providers of datasets for research purposes are typically confronted by a tension between making available useful fit-for-purpose data retaining the fine grain with which the data were obtained, and 'perturbing' the data sufficiently so that, even without obvious identification information such as name, birth data, address location, an 'intruder' or 'attacker' cannot easily obtain the identity of any given person within the dataset. A review of procedures for anonymization of such public use datasets is given by Willenborg and De Waal (2001) and Hundepool, et al. (2012) and references therein.

In the present paper we develop an approach that seeks to perturb data in such a way that the disclosure risk can be quantified while at the same time allowing a data analyst to fit models that respect the fine grain of the original data. The general idea is to use the addition of random noise to some or all variables in a data set where the values of those variables for individuals of interest are also available to an external attacker who wishes to identify those individuals so that they can interrogate their records in the dataset. The idea is that this avoids identification by an attacker via the linking of patterns based on the values of such variables. The noise so generated can then be removed at the analysis stage if its characteristics are known, requiring disclosure of the distribution parameters generating the noise by the statistical agency. This leads to consistent model parameter estimates, although a loss of efficiency will occur. In contrast, the usual method of anonymisation by coarsening the data such as grouping, would not allow the retrieval of the model parameter estimates.

The basic concept is discussed in some detail by Fuller (1993) and also by Winkler (1998). Fuller (1993) points out that the optimum approach is where the random noise is added independently for each variable in the dataset and we shall also make this assumption. He treats the case of normal measurement errors and true data that have a multivariate normal distribution (also used as an approximation for discrete data). He derives a probability that any given record in the released data can be identified as the 'correct' one based upon a subset of the variable values that are known to the attacker. His method of constructing the perturbed data is designed to provide almost unbiased inferences for linear models and he discusses some of the difficulties for non- linear and non- additive models.

Another approach for additive random noise is to add correlated random noise (see: Kim, 1986, Little, 1993, Ting, Fienberg and Trottini, 2008 and Shlomo, 2010). Here the noise that is added is a linear function of the variables to be perturbed. This preserves sufficient statistics in the form of means and covariance matrices, without requiring knowledge of the precise parameter values used to generate the noise, which Fuller's procedure requires. The main drawback, is that it is restricted to models that can be fitted using sufficient statistics such as linear regression, and thus excludes, for example, generalised linear models and multilevel models. It also does not allow diagnostics based on residuals since the production of residual estimates requires knowledge of the noise parameters. Our approach allows for generalised linear and nonlinear models to be fitted to noise-added data within a more general measurement error modelling framework than that of Fuller (2006), and requires knowledge of the parameters used to generate the noise.

Releasing parameters used to generate the noise is common practice in the cryptography literature in Computer Science in order to be able to decode encryptions although it is rarely done at statistical agencies. Cox et al. (2011) discuss the need for transparency in which a statistical agency releases information about the disclosure control processes used to transform the original data to the masked released data. They distinguish between legitimate users and intruders and advocate controlled release of the parameters used to generate the noise so that legitimate users can carry out statistical inferences. Hence we will assume here that the parameters for generating noise are known, either released to trusted users or are in the public domain as is the case for the computer science additive noise approach of Differential Privacy (Dwork, 2006).

It has long been recognized by statistical agencies and data custodians that there is always a trade-off between reducing disclosure risk through statistical disclosure control methods and preserving the analytical properties of the data (Winkler, 1998). However, if stochastic perturbation methods are used to anonymise the data, then, as we demonstrate, the statistical analysis is able to   account for both the   measurement errors and the substantive model of interest. In particular, the greater degree of noise that is added, the lower the statistical efficiency in terms of larger interval estimates for parameters.

Section 2 discusses the technique of anonymisation by adding random noise thus inducing measurement errors into statistical models and how the disclosure risk can be quantified. Section 3 contrasts this technique with other common approaches of anonymisation. Section 4 presents a simulation study on disclosure risk assessment under the proposed approach. Section 5 describes how the anonymisation by adding random noise can be applied to categorical variables and Section 6 on other extensions using varying noise parameters. Section 7 describes how we can compensate for the induced measurement error in statistical analysis with a more detailed description in the appendix where the proposed approach is shown to apply generally to complex models including nonlinear and multilevel models. Section 8 presents a simulation study fitting statistical models under induced measurement error and Section 9 presents both the disclosure risk assessment and statistical modelling on real datasets. We conclude with a discussion in Section 10.

2. Measurement errors

Consider a subset of q variables, y, that are to undergo a statistical disclosure control method. We may also have other variables, say x, that are available to the data analyst but which singly or jointly have little relationship with this subset. In an extreme case such variables may not exist so that effectively the subset q is the complete set of available variables. We assume that the attacker   has knowledge of the anonymisation techniques being used. We deal first with the case of a set of continuously distributed variables assumed to be multivariate normal (MVN).

We suppose that the attacker has a set of values on the q variables, say y^*, that she intends to match against records in the dataset. We introduce random noise in the form of q variates m and for simplicity we shall consider the special case where these variates are independent. We have

$$z = y + m,\ y \sim MVN(\mu, \Omega), \quad m \sim MVN(0, \Omega_m), \quad z \sim MVN(\mu, \Omega + \sigma_m^2 I), \quad \Omega_m\ diagonal$$

The value of Ω_m will determine the strength of the resistance to attack. In the appendix we will introduce a more general notation.

We now form a measure of the distance between y^* and all possible values of z and rank these distances. We shall assume that the values $y,$ are without error, although our procedure can be extended to that case straightforwardly. We also assume that for each record belonging to the attacker there does correspond a record in the dataset and that some or all the variables may have their true values. From the attacker's perspective the best case scenario is where all her variables have the true values, and we shall assume that this is the case in our simulations and substantive example.

We first consider measuring the distance between the attacker's data and each record in the dataset. We shall then briefly consider how an attacker might be able to improve their chance of detecting the desired record.

A general distance measure can be written in the form

$D^* \propto (z - y^*)^T W (z - y^*)$, and where in the case of independence we have the Euclidean distance for each comparison record i and we have

$$D_i^* = \sum_{j=1}^{q}(z_{ij} - y_j^*)^2, \quad D_i = \sum_{j=1}^{q}(y_{ij} - y_j^*)^2, \quad i = 1, \dots., n \qquad (1)$$

For a given attacker record we form $r_i^* = rank(D_i^*),$ and $r_i = rank(D_i),$
and let i^* be the value of i for $r_i^* = 1$, that is the closest record for the attacker in terms of the distance measure. We define $h = r_{i^*} - 1$ which is the difference in ranks between the record identified by the attacker and the rank of the actual closest record and we refer to it as the h-rank disclosure index for y^*, or simply as the h-index. Thus if h=0 we have the correct match. We also note that the record held by the attacker may not correspond to any individual in the dataset so that a correct match will never occur.

We therefore need to determine σ_m^2 such that $E(h)$ is large enough (for example taking the value of 3) to create sufficient unreliability of determining the correct record for the attack not to be worthwhile. Alternatively, we could require

$$\Pr(h < p) < \epsilon \qquad (2)$$

where a suitable choice might be, say, $p = 3, \epsilon = 0.1$.

The distribution of h will in general be a function of y^*, for example if y^* is a multivariate 'outlier' the attacker is more likely to find the correct match. In the simulation discussed in Section 4 we will examine the performance of the procedure with respect to values of percentiles of D. We shall not consider the case where we can randomly create missing data except to note that this will contribute to the unreliability of the matching process. In fact data values may be missing in any given dataset so that our computations in that respect represent a best case scenario.

In principle, an attacker who has access to the noise parameters may be able to utilise this to improve their attack strategy by making use of this information.

Thus, with knowledge of Ω_m rather than utilising $z,$ the attacker could obtain more precision since they would be able to estimate

$$z^* = E(y|z) = \left(cov(z)\right)^{-1} cov(y) \times z \qquad (3)$$

In the case of independent random noise for normal variables a simple procedure would be to use $z_j R_j$ for variable j, where $R_j = \sigma_y^2/(\sigma_m^2 + \sigma_y^2)$ is the 'reliability' of the observed variable. We will investigate any advantage to the attacker that the use of (3) might bring in the simulation discussed in Section 4. We also point out in the discussion that information about Ω_m can be limited to accredited data analysts through secure settings.

3. Similarity to other methods

Standard procedures for anonymisation are typically based upon a distinction that is made between primary individual identifiers such as name and birth date, quasi identifiers such as ethnic group and sensitive identifiers such as disease diagnostic categories that may have disclosive values. Thus *k*-anonymity models and extensions such as *l*-diversity and *t*-closeness consider two stages. In the first the values for the primary identifiers are coarsened to the extent that in the final data set there are at least *k* records that have identical identifier values, for any given set of identifier values (a *k*-anonymity data cell). In the second stage the simplest form of *l*-diversity requires that for any quasi-identifier there are at least *l* distinct values for each *k*-anonymity data cell. For *t*-closeness the difference (using a suitable measure) between the distribution of the *l* values in the data cell and the distribution in the whole sample should not exceed a threshold *t*. These methods may also lead to the suppression of certain variable values to ensure the required anonymity level. (see for example, the Statistical Disclosure control package http://cran.r-project.org/web/packages/sdcMicro/sdcMicro.pdf). One of the drawbacks of such methods is that they may remove or degrade too much identifying data that is of importance to the data analyst.

Other approaches that focus on estimating the risk of re-identification of a data subject with noise-added data are found in Winkler (1998) based on probabilistic record linkage and in Reiter and Mitra (2009) and Shlomo and Skinner (2010) on using probabilistic modelling to estimate a probability of re-identification whilst accounting for the perturbation. Reiter and Mitra (2009) also accounts for intruder knowledge. Winkler (1998) in particular concluded that even moderate amounts of additive noise where some of the analytical properties of the data are preserved may still have considerable disclosure risks.

Polettini and Arima (2015) propose a method for inference under perturbed data that has similarities to our own. They develop it in the context of small area estimation where the predictor variables have been masked using the post-randomization method (PRAM) (Gouweleeuw, et al., 1998). They apply a Bayesian algorithm to the data measured at the aggregate small area level, where categorical data are approximated by a multivariate normal distribution and where the adequacy of the approximation is a function of the number of individual records within an area. Our own procedure primarily operates at the level of individual records and does not involve such approximations. It can also be used to deal with data aggregated to higher levels as explained below. Woo and Slavkovic (2014) also discuss logistic regression with variables subjected to PRAM.

We make no distinction between primary, quasi and sensitive identifiers, although all could be incorporated in the distance computation if needed. Our proposed method is essentially probabilistic rather than one that guarantees a given level of anonymity, as in the standard k-anonymisation methods. The key element, however, is that from the data analyst's viewpoint there is no data coarsening, for example by grouping or truncating extreme values, with potentially enhanced quality for the inferences. It is applicable to all statistical models, in particular the family of generalised linear models, for which measurement error methods can be applied. Where there are established procedures for diagnostics these can also be utilised since the estimation as described in the appendix provides the necessary parameters for the model of interest.

Adding noise is similar in some ways to the so-called fully synthetic data approach where data is generated from a series of predictive distributions (Rubin, 1993 and Reiter, 2005). Our proposed approach for measuring disclosure risk and analysis is applicable in this setting as well.

4. A simulation for disclosiveness

We generate a series of simulated datasets with 1000 records with a mean vector of zero, q=5 and $\sigma_m^2 = 0.1$ and Ω has all variances =1 and covariances = 0.25. This value of q is chosen since it will typically represent the number of identifiers available, but we have varied this as well below.

For each true value record, treating it as a potential attacker record we generate D_i as in (1). We choose 9 values representing approximately deciles of the distribution of D_i, to define suitable attacker records y^*. The distribution of h varies by decile with greater precision of attack at extreme values and we show results for different deciles. The choice of standardised variates simplifies the computations somewhat.

Routines are written in MATLAB. Based on 1000 simulations we obtain the following results in Table 1 for the cumulative distributions.

Table 1. Cumulative percentage distribution for *h* for deciles of the distribution of D_i

h	Decile								
	10	20	30	40	50	60	70	80	90
0	52.2	49.4	43.9	41.3	41.7	43.5	40.5	47. 0	56.2
1	62.9	60.7	56.1	53.1	53.1	54.3	53.0	58.3	67.3
2	70.0	65.3	62.0	61.2	60.8	61.8	60.9	64.7	73.8
3	74.7	70.2	68.6	66.0	65.8	66.6	65.8	70.3	77.9
4	78.5	74.4	72.8	70.1	68.7	70.0	69.6	73.8	81.8
5	80.8	77.5	76.5	72.7	71.5	72.1	72.8	76.8	84.3
6	83.1	79.9	78.1	74.3	74.1	74.8	75.5	78.7	86.2
7	84.4	82.4	80.6	76.5	76.1	76.2	78.0	81.2	87.3
8	85.9	83.9	81.7	78.7	78.2	77.8	80.0	83.9	88.9
9	87.7	84.9	83.5	79.5	79.4	80.2	81.8	84.9	90.7
10	89.1	86.4	85.0	81.1	81.2	81.8	83.2	86.6	91.9
11	90.1	87.1	85.4	82.4	83.5	83.2	84.1	87.4	92.9
12	91.1	88.3	86.1	83.8	84.6	84.7	85.3	88.2	93.9
13	91.9	88.8	87.1	85.3	85.6	85.8	86.6	89.1	94.3
14	93.0	89.3	87.8	86.5	86.1	86.6	87.5	90.0	95.0
15	94.2	90.1	88.1	87.4	86.6	87.4	88.0	90.2	95.1

16	94.9	91.1	88.8	88.0	87.5	87.9	88.5	90.9	95.1
17	95.2	91.7	89.3	88.8	88.5	88.5	89.5	91.4	95.6
18	95.7	91.9	89.8	90.0	89.6	89.1	90.1	91.8	95.9
19	96.5	92.6	90.3	90.9	90.0	89.9	90.5	92.4	96.6
20+	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Thus, from the viewpoint of the attacker, in more than 40% of the cases the nearest record is not the true one. For the 10th decile $pr(h > 3) = 0.25$ and $pr(h > 5) = 0.19$. For the median an attacker has a harder time with $pr(h > 5) = 0.29$. Depending of course on the degree of disclosure risk that can be tolerated, it could be argued that this is adequate to make an attack too unreliable to be worthwhile.

For the weighted distance case in (1) where $W = \Omega$ we obtain essentially similar results.

We have also run the simulation with larger sample sizes and we find that at the lowest decile for a greater sample size of 10,000 $pr(h > 3) = 0.60$ and $pr(h > 5) = 0.54$.

We now look at a range of values for $\Omega \; and \; \sigma_m^2$ and different sample sizes. We study just the case for the lowest decile: an individual with more extreme values presents a target that is more favourable to an attacker. We present results for $pr(h > 5)$ and for two different sample sizes, in Table 2.

Table 2. Lowest decile estimates for *h*

pr(h>5) for combinations of $\Omega \; and \; \sigma_m^2$ where Ω always has unit diagonal elements and equal off-diagonal elements (given by columns) are shown.					
Sample size =1000					
$\sigma_m^2.$	0.1	0.2	0.3	0.4	0.5
0.1	0.15	0.16	0.19	0.23	0.24
0.2	0.45	0.43	0.46	0.50	0.54
0.3	0.58	0.63	0.63	0.65	0.70
0.4	0.73	0.74	0.74	0.76	0.77
Sample size =5000					
$\sigma_m^2.$	0.1	0.2	0.3	0.4	0.5
0.1	0.41	0.48	0.48	0.50	0.55
0.2	0.72	0.71	0.75	0.77	0.80
0.3	0.84	0.84	0.87	0.88	0.89
0.4	0.90	0.90	0.92	0.90	0.93
pr(h=0). For combinations of $\Omega \; and \; \sigma_m^2$ where Ω always has unit diagonal elements and equal off-diagonal elements (given by columns) are shown.					
Sample size =1000					
$\sigma_m^2.$	0.1	0.2	0.3	0.4	0.5
0.1	0.56	0.54	0.53	0.49	0.45
0.2	0.27	0.27	0.24	0.23	0.18
0.3	0.16	0.13	0.15	0.12	0.09
0.4	0.10	0.09	0.010	0.08	0.07

We see that even with the smaller sample size of 1000, and moderate proportions of noise (10% of the variance of the true values), we have reasonably high probabilities of h exceeding a value of 5 and small probabilities that the nearest record is the correct one.

We now study what effect the use of (3) has, that is when the attacker makes use of information about the parameters of the noise distribution.

Table 3. Cumulative percentage distribution of *h* for deciles of the distribution of D_i where noise parameters are known and expected values of identifiers with noise are used

h	Decile				
	10	20	30	40	50
0	54.7	48.7	44.3	41.3	45.4
1	67.1	61.8	54.5	55.6	57.0
2	73.1	68.2	62.8	64.1	63.7
3	78.6	72.4	66.7	69.3	68.4
4	81.9	75.3	70.6	73.1	72.4
5	84.5	79.3	73.7	75.4	74.9

Table 3 shows, for $h = 1, \dots 5$, and for the 10th-50th percentiles, the percentage distributions of h when knowledge of the noise parameters is used as described in (3) under the same simulation conditions as Table 1. Comparing with Table 1 we see generally small increases in the cumulative probability that an attacker selects a record close to the correct one.

We would expect that the probability of the attacker selecting a record close to the correct one will increase with the number of distinct identifiers used. Thus, for example, if there are 10 identifiers and we simulate with the same value for Ω as before, we now need a noise parameter $\Omega_m = 0.34$ rather than $\Omega_m = 0.1$ to obtain approximately the same values for the distribution of h. This suggests that careful consideration needs to be given to the likely number of identifiers available to the attacker. We have also varied the size of the covariances from 0.1 to 0.5, but this has only a small effect on the distribution of h, with a decrease in the covariance associated with a slightly higher risk of disclosure, for example for the 5th percentile the $pr(h = 0)$ is 0.47 for a covariance of 0.5 as opposed to 0.53 with a covariance of 0.1 as in Table 1.

5. Misclassifications for categorical variables

Consider, for simplicity, a series of q independent binary identification variables. We assume that one of the categories is small e.g. $\pi = pr(y = 1) = 0.1$. Thus, if q=3, with $\pi_i = 0.1, i = 1, \dots, 3$ then the most favourable vector y^* is $y_i = 1 \; \forall \; i$.

Suppose now we introduce a simple misclassification where for each $i | y_i = 0$ independently, we randomly assign $y_i = 1$ with probability 0.1. Thus for each binary variable we now have $pr(z_i = 1) = 0.19$. The probability that all three variables have value 1 i.e. $pr(z_i = 1), \forall \; i = \; 0.19^3 = 0.007$, whereas $pr(y_i = 1), \forall \; i = \; 0.1^3 = 0.001$. Thus of those identified only 15% are correctly identified. Such procedures for categorical variables have been implemented in the PRAM method (Gouweleeuw, et al., 1998).

For multicategory, including ordered and unordered variables, we can alternatively consider the following, simpler, procedure. Numbering the categories $j = \{1, \ldots., p\}$ we independently add to the true category code noise

$$m_j \sim N(0, \sigma_m^2\ ), 1 \leq m_j \leq p \qquad (4)$$

This results in a truncated normal distribution and the variable enters (1) along with the continuous variables. The purpose of the truncation is to avoid easy detection for the extreme category codes. We could also generalise (4) to allow different variances for each category. We note that the noise is simply added to the category codes $\{1, \ldots., p\}$, irrespective of whether this is an ordered or unordered variable. The MCMC steps described in the appendix show how we may then draw from the posterior distribution of the true (unknown) category codes. To avoid the potential objection that it may be confusing to release categorical variables (with added noise) as continuous, we also describe in the appendix, how a rounding of the noise-added values to integer values can also be used, although this will result in less efficient estimates.

6. Fitting models with known noise or measurement error

Bayesian procedures for fitting models with measurement errors have been proposed (Richardson and Gilks, 1993) and these have been further developed to fit multilevel data structures and allow models that include interaction and power terms. The noise that is added in our procedure has the characteristics of measurement error and can be treated as such. An outline of a general algorithm with details specific to data anonymisation is given in the appendix, and further details of the estimation algorithm can be found in Goldstein and Browne (2017). Other methods for estimation of models with measurement errors are also available, such as the simulation-extrapolation method, SIMEX (Delaigle and Hall, 2008) and moment based estimators described by Fuller (2006), and these can also be used. . The Bayesian model procedure that we describe has the advantage that it is a fully specified probabilistic model that is readily generalised to handle complex data structures including multilevel and generalised linear models without approximations, with straightforward computation of interval estimates.

Full estimation details are given in the appendix, and can be summarised as follows. For ease of exposition we assume a single level linear model with just a single predictor variable that has added noise. The case of several predictors with independent added noise and the case where we have a generalised linear model and the multilevel case follow straightforwardly. We assume here multivariate normality for the noise and discuss the modifications needed to fit generalised linear models and multilevel models in the appendix.

Define the true values of the variable with added noise as X_1 and those variables without added noise as X_2, and $X = [X_1\ X_2]$ where X_2 is known.

Define the joint model – the noise or measurement error model (MEM) in two parts, (5a) and (5b) and the model of interest (MOI) (5c).

$$x_1 = X_1 + \gamma_1 \qquad (5a)$$

$$X_1 = X_2^T \alpha + \gamma_2 \qquad (5b)$$

$$Y = X\beta + e \qquad (5c)$$

where $\gamma_1 \sim N\left(0, \sigma^2_{\_\gamma_1}\right)$, $\gamma_2 \sim N\left(0, \sigma^2_{\_\gamma_2}\right)$, $e \sim N(0, \sigma^2_e)$. We note that the MOI (5c) may contain functions of the X_1, such as interaction or power terms. Lower case variables define observed and upper case true values, and we assume the residual terms in (5a)-(5c) are independent.

The appendix details the MCMC steps required to fit this model. In brief this involves the following steps:

1). Update the true values using a metropolis step, conditionally on the current values of the other parameters

2). Update the α parameters using a Gibbs step, conditionally on current values of the other parameters

3). Update the β parameters using a Gibbs step, conditionally on current values of the other parameters

4). Update the variance and covariance parameters, conditionally on current values of the other parameters

In the appendix we also discuss the following extensions and the difficulties associated with them. Where we have added noise in the response variable, the general effect of correcting these is to shrink the estimate of the residual variance and hence inflate the standard errors for the parameters. Where the response is categorical, notably binary, a further step in the algorithm is involved. We also discuss in the appendix how to deal with variables that have been truncated, for example to handle large outliers, and how to handle categorical variables that are presented rounded to the nearest integer value.

After fitting a suitable model such as the above, the original variable scales and relationships are fully recovered, albeit with a loss of efficiency – which with very large datasets may not be an important issue. The loss of efficiency, in terms of interval estimates or standard errors associated with parameter estimates can be estimated for any proposed model to be fitted to the perturbed data, given the noise parameters. Thus, for example, in the simple regression case where independent normal noise with a common variance σ^2_m has been added to the set of predictors X, we can use as a simple overall measure for the relative efficiency the determinental ratio $(|X^T X| / |X^T X + N\sigma^2_m|)$ where X includes those with and those without measurement errors and N is the sample size. The data provider would be able to supply such estimates, but perhaps of more use will be estimates of the inflation of standard errors, for some typical models, associated with individual parameters. These could be provided alongside the released data or possibly requested from the data provider by a data analyst with respect to any given fitted model. We illustrate the effects on standard errors in our example analysis in Section 9.

Where the original variables are also subject to measurement errors with known distributions the fitted model will be based upon the total measurement error.

In the case where the data provider wishes to impose additional, known, data constraints (see discussion in Section 9) some modification to the estimation algorithm will be needed to incorporate these. The appendix also indicates how these can be incorporated.

7. A simulation of a measurement error model

We carry out a simple simulation from the following model, in order to illustrate that we can readily recover the signal from noisy data for both a binary and continuous predictor. The model we simulate from is given by

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \qquad e_i \sim N(0,1),\ \ \beta_0 = \beta_1 = \beta_2 = 1 \qquad (6)$$

$$\begin{pmatrix} x_1 \\ x_2^* \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right),\ \ x_2 = 0\ if\ x_2^* \leq 0;\ \ \ x_2 = 1\ if\ x_2^* > 0;$$

Independent noise with variance $\sigma_m^2 = 0.2$ is added to x_1, x_2. We carry out two sets of 100 simulations with 1000 records. The results are given in Table 4.

Table 4. Estimates from addition of noise (standard errors in brackets*). MCMC burn in =500 iterations=500. Simulations from model (5) Sample size = 1000, number of simulations = 100

Parameter	Simulation parameters	Noisy data no adjustment	Noisy data adjusted	Noisy data adjusted % bias
β_o	1.0	0.974 (0.004)	0.997 (0.003)	-0.3
β_1	1.0	0.887 (0.002)	1.004 (0.003)	0.4
β_2	1.0	1.051 (0.005)	1.003 (0.005)	0.3
σ_e^2	1.0	1.0	1.0	0

*The standard error estimates are the standard deviations computed from the MCMC chains.

We see negligible bias, no more than 0.5% for the adjusted estimates and in all cases the 95% confidence interval overlaps the true value.

8. Example analyses

Our first example illustrates just the use of a measurement error model for two level data where we have added noise and used the procedures in the appendix to estimate the true model parameters. The data will be referred to as the Tutorial dataset (Goldstein et al., 1993). The response is a normalised examination score taken at age 16 by 4059 students in 65 schools in Inner London. The predictor variables are a standardised reading test score taken at age 11 (x_1) before pupils attended their secondary school and the binary variable gender (x_2). A 2-level variance components model is fitted:

$$y_{ij} = \beta_0 + \beta_1 x_{iij} + \beta_{2ij} x_{2ij} + u_j + e_{ij}, \quad u_j \sim N(0, \sigma_u^2), \quad e_{ij} \sim N(0, \sigma_e^2) \qquad (7)$$

To illustrate our procedure we add normally distributed noise with mean 0 and variance $\sigma_m^2 = 0.2$ independently to the reading score and gender. For gender, values less than 0 are set to 0 and values greater than 1 are set to 1. The principal purpose of the first example is to show how adjusting for the added noise produces consistent estimates rather than to explore a range of values against disclosure risks.

Table 5 shows the results from fitting the model before adding noise, fitting the model with added noise but without adjusting for measurement error and fitting adjusting for measurement error. Only one application of measurement error addition has been used to produce these results and the response variable is not perturbed.

Table 5. Tutorial dataset. Estimates from addition of noise (standard errors in brackets). MCMC burn in =500 iterations=500. Standard errors in brackets using MCMC chain standard deviation estimates. Normal noise variance 0.2.

Parameter	Original model	Noisy data no adjustment	Noisy data adjusted
β_o	-0.097 (0.050)	-0.082 (0.044)	-0.093 (0.042)
β_1	0.559 (0.013)	0.460 (0.012)	0.549 (0.014)
β_2	0.174 (0.033)	0.109 (0.030)	0.155 (0.035)
σ_u^2	0.097 (0.021)	0.102 (0.022)	0.100 (0.020)
σ_e^2	0.563 (0.013)	0.614 (0.013)	0.562 (0.014)

We note that the adjusted estimates are close to those using the original data whereas ignoring the measurement error produces estimates with considerable biases.

Our second example uses a dataset from a 1982 survey of the sugar cane farm industry in Queensland, Australia that was used in Chambers and Dunstan (1986).

It illustrates both the measurement error model and the computation of the h-index for disclosiveness. In order to compare our approach for accounting for the measurement error in our analyses with the case of adding correlated noise and preserving sufficient statistics, we continue using a linear regression model. We note that under more complex models such as generalized linear models, the correlated additive noise approach would not provide valid results.

The model of interest has the sugar cane yield receipt as response (y) and predictors are region (x_1, Northern=1, Southern=0), sugar cane harvest (x_2, continuous in tonnes) and cost (x_3, in Australian dollars). There are 333 farms in the dataset with no missing data.

The model to be fitted is the linear regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i \qquad e_i \sim N(0, \sigma_e^2) \qquad (8)$$

We have added noise to the data in two different ways. In the first case we use correlated Gaussian noise as described in Shlomo (2010) and in the second case we add independent Gaussian noise to each variable as described in Section 2. In both cases, we add two levels of noise: the variance of noise is 0.05 and 0.17 times the variance of the corresponding variance of the true values for each variable.

We first present the results of fitting (8) to each of the noisy datasets in Tables 6a and 6b and then present the results of computations on disclosiveness in Tables 7a and 7b. We note that for correlated noise the model is simply fitted to the observed data after a single draw and the reported standard errors are analytical resulting from the model fit. For independent additive noise we use the procedures described in the appendix where reported standard errors are empirical resulting from the MCMC chains. Table 6a is based on the case where noise is added to the predictors only and Table 6b are based on the case where noise is added to both the predictors and the response variable.

Table 6a. Sugar cane farm data. Results of fitting (7) to 5% and 17% relative variance noisy data. Standard errors in brackets. Noise added to predictors only

Predictors Only					
Parameter	True data with no added noise	Correlated Noise with 5% of variance added	Independent Noise with 5% of variance added	Correlated Noise with 17% of variance added	Independent Noise with 17% of variance added
β_o	-7.89 (1.70)	-6.01 (2.25)	-5.71 (1.70)	-3.56 (3.35)	-9.16 (2.27)
β_1	11.35 (1.37)	10.13 (1.83)	10.03 (1.38)	14.20 (2.68)	13.05 (1.83)
β_2	0.022 (0.0007)	0.022 (0.0009)	0.022 (0.0007)	0.021 (0.0014)	0.021 (0.0011)
β_3	0.00013 (0.00004)	0.00008 (0.0006)	0.00010 (0.00004)	0.00015 (0.00008)	0.00022 (0.000070)
σ^2	146.2 (11.23)	254.02	150.1 (11.63)	530.2	170.8 (17.86)

Table 6b. Sugar cane farm data. Results of fitting (7) to 5% and 17% relative variance noisy data. Standard errors in brackets. Noise added to predictors and response

Predictors and Response Variable					
Parameter	True data with no added noise	Correlated Noise with 5% of variance added	Independent Noise with 5% of variance added	Correlated Noise with 17% of variance added	Independent Noise with 17% of variance added
β_o	-7.89 (1.70)	-7.33 (1.71)	-8.95 (2.88)	-7.56 (1.81)	-7.04 (4.39)
β_1	11.35 (1.37)	11.82 (1.41)	11.91 (2.26)	11.36 (1.43)	11.38 (3.64)
β_2	0.022 (0.0007)	0.022 (0.0007)	0.021 (0.0014)	0.022 (0.0007)	0.022 (0.0031)
β_3	0.00013 (0.00004)	0.00011 (0.0004)	0.00023 (0.000086)	0.00014 (0.0004)	0.00016 (0.00019)
σ^2	146.2 (11.23)	148.01	180.7 (28.32)	150.46	179.8 (73.3)

We see in Table 6a the negative effect of releasing one draw of correlated noise to users without providing the parameters of the noise distribution when only some of the variables intended for the statistical modelling have added correlated random noise. Users will obtain biased estimates with very large standard errors. This is not the case in Table 6b where all variables intended for the statistical modelling have correlated noise added to them. In that case, we obtain similar parameter estimates and standard errors to the model run on the original true data. Given that it is generally unknown what types of analysis will be carried out on the released perturbed data, it is essential that users obtain the noise parameters and be able to analyse the data under measurement error using the procedures described in the appendix. We see in Tables 6a and 6b under these procedures that we obtain unbiased parameter estimates taking into account the measurement error regardless of which variables have been perturbed. We also see some considerably increased standard errors as greater amounts of noise are added. For example, under the 17% of the variance of the true response variance for receipts, the variance of the response noise was set at 411 and this is nearly three times the true residual variance as estimated in Tables 6a and 6b. This governs the size of the standard errors. Even when the variances are chosen to be just 5% of the variances of the true values, we still see an increase in the standard errors.

Table 7a presents the values of the *h*-index for individual records chosen to represent both the centre and extremes of the data distribution, for different amounts of random noise on the correlated noise addition on all four variables (predictors and response variable): receipts, region, harvest and costs and similarly Table 7b presents the independently added random noise. To avoid skewness in these variables, the distances were calculated on standardized variables although given the nature of this data, some skewness remains.

Table 7a. Sugar cane data. Values of *h*-index for records at different multivariate distance quantiles. Random correlated noise addition all variables as percentage of true variances

Distance quantile	h-index. 5% of true variance	*h*-index. 17% of true variance	*pr*(h=0) 5% of true variance	*pr*(h=0) 17% of true variance
5	11.2	23.2	0.11	0.04
10	17.2	29.8	0.05	0.05
50	17.5	37.9	0.08	0.04
90	2.7	10.7	0.43	0.17
95	0.0	0.5	0.98	0.78

Table 7b. Sugar cane data. Values of *h*-index for records at different multivariate distance quantiles. Random independent noise addition all variables as percentage of true variances

Distance quantile	h-index. 5% of true variance	*h*-index. 17% of true variance	*pr*(h=0) 5% of true variance	*pr*(h=0) 17% of true variance
5	7.0	16.0	0.12	0.06
10	12.6	23.0	0.07	0.04

50	12.7	27.8	0.08	0.03
90	1.7	6.8	0.38	0.15
95	0.0	0.2	1.00	0.89

From both Tables 7a and 7b, it is clear that in terms of their multivariate distance from the data centroid, the h-index values show a high level of disclosure protection for both the 5% and 17% relative variance added noise values except for the far right tail due to the skewness of the data in the sugar farms dataset. In the latter case the data provider may well decide to group (truncate) the very large values, as we describe in the discussion below. Comparing correlated random noise addition in Table 7a with the independent random noise addition in Table 7b, we see quite similar results for $pr(h = 0)$, although the value of h overall does tend to be greater for the correlated noise. The 17% relative variance added noise values offers greater disclosure protection than the 5% relative variance, but this may still not be satisfactory and the data provider would then have a choice of increasing the amount of added noise, adopting a truncation procedure, or making the noise variance a monotonically increasing function of the true value. These options will be pursued in future research. The results show also that the disclosure risks increase for the case of only perturbing the predictors and not the response variable.

9. Discussion

We have shown that the disclosure protection provided by additive random noise addition increases with sample size and this will generally be of concern for sample surveys with relatively small sample sizes. For example, we see that with a sample size of only 1000, in the case most favourable to an attacker that we have explored in our simulation, the probability that the nearest record is the correct record is less than half. If an attacker has access to the noise parameters and utilises this information to obtain an estimate of the covariance matrix for the true values, our simulation suggests that there is only a small increase in the probability of selecting the correct record, and that this is of little practical importance. If the number of identifiers available to the attacker increases then this can enhance the chance of a successful attack. We have shown that we require rather larger amounts of noise as the number of identifiers available to the attacker increases so that a realistic assessment is needed on the number of identifiers that may be available to an attacker. On the basis of our simulation it appears that the disclosure risks are relatively insensitive to correlations between the identifiers. We would, however, caution that our simulations are limited and based upon an assumption of multivariate normality. Further research needs to explore more general cases and we would suggest that one responsibility of a data provider is to provide estimates for disclosure probabilities for their own data, based upon the distributions observed in the data.

Our results are based on the assumption that the attacker has exact knowledge of who is in the dataset and some target individual's exact data points. In the absence of any other information available to the attacker, this is a worst case scenario and indeed in a sampling context, response knowledge is not assumed known. Therefore, disclosure risks would be considerably less than our reported findings. If an attacker has some random data points from the population she will first have to check if the target individual is in the dataset and that depends on the sampling fraction which are generally small. Often, a data attacker will have no pre-existing individual data and may be concerned to trawl the dataset to discover an 'interesting' record, for example an individual with an unusual combination of values.

Having identified such an individual they may then attempt to identify the real person in the population using other variables in the data record. Our procedure is also relevant to such an attack so long as the noise has been applied to the variables in question.

For the extremes of the distribution it would be useful for disclosure purposes, to apply measurement errors with larger variances. For example, if noise is added to a variable such as income, we might wish to make the variance of the noise a function of income itself. Such a function could be non-linear or a step function where all true value greater than a specified (absolute) value have additional noise added. Nevertheless, the release of the information describing such a functional relationship will generally be informative for individual records and so such information could be disclosed only to the accredited data analyst. In terms of the measurement error algorithm described in the appendix the value of the variance would simply need to be updated at each MCMC iteration with the current value for the true value of the variable. This is an area of further research to be pursued.

There is an interesting contrast with the *k*-anonymity criterion that is often used as a measure of disclosiveness. If we have 2-anonymity this implies that an attacker is able to identify two individual records matching her own information, so that choosing either of them at random means that there is a probability of 0.5 that it is the correct one. The *h*-index, however, as quoted in the previous paragraph, only yields a single individual with a probability about 0.5 and thus provides less information to the attacker than in the case of 2-anonymity. Indeed, an attacker may be quite content with the information that they can access 2 or perhaps even 5 records containing the one that is sought. By contrast, with the *h*-index procedure, in our most favourable case the probability of the sought-for individual being one of the two nearest is just over 60% and one of the five nearest just under 80%, so that it could be argued that this is sufficient to deter an attacker and hence suitable in terms of protecting against disclosure. In practice careful attention needs to be paid to the amount of noise required to satisfy disclosure concerns and this is an area for further research.

The generality of our procedure is that it makes no assumptions about the final model to be fitted, which we point out in the appendix can be a generalised linear model or a multilevel model with some or all of the variables perturbed, and the procedure allows a full range of exploratory analyses. Likewise, in contrast to previous work, it does not assume any particular distribution for the true values, at least for those used as covariates in the substantive model of interest.

Our procedure can be contrasted with procedures based upon the production of fully synthetic data simulated from estimates of the structure of the real data where exploratory analyses are recommended prior to the choice of a small number of models to be fitted to the real data within a secure environment. Such procedures not only rely upon good estimates of the real data structure, they also rely upon exploratory analyses converging on the appropriate set of final models, and this is by no means guaranteed. We note, however, that in some cases where we wish only to fit a linear regression model a more tailored procedure such as that using correlated noise (Shlomo, 2010) may provide superior estimates. With our proposed procedure exploratory analyses would generally be carried out using the measurement error methods proposed.

For a response variable with added noise, the situation is more complex but for normally distributed responses we can carry out an analysis using the observed values to obtain consistent estimates. In this case we can obtain a consistent estimate for the residual variance by subtracting the (known) variance for the measurement errors in the response from the final estimated residual variance. The drawback, as illustrated in our second example in Section 9 on the sugar cane farms dataset, is that where the proportion of variance explained is high, this will lead to large standard errors. In this case we can either add less noise to any variables that a given data analyst wishes to use as responses, or no noise at all. This would of course, require a close collaboration between data provider and data analyst in deciding which variables to perturb. In our example this has little effect on the disclosiveness and in practice therefore will often be acceptable. For categorical responses as with continuous responses, when data are released it may be acceptable to provide any variable(s) to be treated as a response with its true values or with just a small amount of added noise. Where a categorical response has added noise it is rounded to the nearest category value, for example 0 or 1 in the case of binary data, and these values are then used as responses as described in the appendix. Since in general different data analysts may wish to treat different variables as responses, this implies that there should be a close liaison between the data provider and the analyst so that appropriate amounts of noise can be attached to each variable. One aspect of this to explore would be the implementation of automatic procedures for noisy dataset generation according to given specifications of disclosure risk and statistical estimation efficiency.

In some cases we can reduce disclosure risk by first transforming one or more variables. This will often arise with skewed distributions with long tails where, for example, a logarithmic transformation will create fewer extreme values. In such cases the noise will be added to the transformed variable. In an analysis where the original variable is required in the model of interest, then for the likelihood term associated with the model of interest the back-transformed value of the proposed value will be used (see appendix).

A key issue, of course, is the requirement that the data provider supplies to the data analyst the necessary parameters used to generate the noise. Since the degree of privacy established needs to be published and the degree of privacy established is a function of these parameters, in a weak sense aspects of the noise will become publicly available. This, however is not seriously disclosive since a guarantee of *h*-level disclosure is only weakly informative about the noise parameter values themselves. Furthermore, we have also shown in Table 3 that even where the attacker has access to the noise parameters this does not materially improve the probability of disclosure. In addition, a further precaution is to release the noise information only to accredited data analysts under secure conditions, and thus it would be unavailable to an external malicious attacker.

Whilst our proposed procedure can provide general protection against attack, a data provider may wish to guard against specific aspects of disclosiveness in the data, such as the presence of one or two very large outliers where the use of truncation on the perturbed data may be adequate. As long as the relevant information is made available, the analyst typically will be able to take account of these additional constraints in the analysis. As discussed in the appendix, it is possible to incorporate judicious groupings of data values within the estimation algorithm, and this allows the data provider some freedom in deciding where to coarsen particular data ranges. Further research exploring such possibilities, and in particular investigating the trade-off between increased security and reduced efficiency, would be welcome.

Data sets are often supplied with weights that may incorporate aspects of sample design or bias correction procedures. This poses particular problems for our procedure as it does in general for models utilising Bayesian methods. In a recent paper Goldstein, Carpenter and Kenward (2018) showed how weights could be incorporated in a Bayesian model for handling missing data. Goldstein, Browne and Charlton (2017) extend the model for missing data to handle measurement errors and the procedures for handling weights given by Goldstein et al. (2018) can be extended in a straightforward fashion to this extended model. It should also be noted that the ability of the extended model to handle both measurement errors and missing data allows our procedures to deal with the case where there are missing values.

The protection method of perturbing a few variables in released microdata and other common approaches such as truncation and grouping that we have presented here are all standard statistical disclosure control (SDC) methods implemented by statistical agencies (Hundepool et al, 2012 and references therein). The computer science definition of Differential Privacy (DP) also assumes a worst case scenario that the attacker knows who is in the dataset and does not take into account of any protection afforded by sampling. The perturbation mechanism in the DP setting is also additive random (Laplace) noise where the parameters of the noise distribution depend on a privacy budget and the ‘sensitivity’ defined as the maximum distance of two neighbouring datasets that differ in only one data subject. For more details on DP, see Dwork (2006) and Dwork and Roth (2014). However, DP is related to output perturbation where every query is perturbed and it is not relevant in our case where we release microdata with only a few variables perturbed, coarsened or truncated as is the norm in SDC practices at statistical agencies. One advantage of the DP framework is that the noise distribution does not need to be secret thus removing one potential threat, although this does not appear to be an insurmountable problem for our proposed method. See Charest (2010) and Rinott et al. (2017) for examples of inference under DP.

There are practical considerations to be taken into account if our procedures are to be implemented. Not least of these is the need to provide easy-to-use software to perform the appropriate analysis on the noisy data and accompanying training materials. The software routines written in MATLAB(2017), used for the present paper are not optimised for either speed or user accessibility. They are, however, available by request from the first author.

Finally, as we pointed out in the introduction, it is important to recognise that there is always a trade-off between reducing disclosure risk and increasing the complexity and efficiency of any resulting analysis. The more noise that is added, the lower the statistical efficiency. In practice the balance between exposure risk and analytical efficiency can be tailored to individual data users through a secure environment. The safer the environment of the data analyst, in general the less noise will be needed, and likewise the level of noise could be tailored to the sensitivity of the data.

Acknowledgements

The authors would like to thank the Isaac Newton institute for mathematical sciences for its hospitality during the record linkage programme supported by EPSRC grant EP/K032208/1, where useful feedback was obtained.

References

Chambers, R.L. and Dunstan, R. (1986). Estimating Distribution Functions from Survey Data. Biometrika, 73, 597-604.

Charest, A.-S. (2010). How Can we Analyse Differentially-private Synthetic Datasets? Journal of Privacy and Confidentiality 2 21-33.

Cox, L., Karr, A.F. and Kinney, S.K. (2011). Risk-Utility Paradigms for Statistical Disclosure Limitation: How to Think, But Not How to Act. International Statistical Review, Vol. 79, Issue 2, 160-183.

Delaigle A, Hall P (2008). Using SIMEX for Smoothing-Parameter Choice in Errors-in-Variables Problems." Journal of the American Statistical Association, 103(481), 280{287.

Dwork, C. (2006). Differential Privacy. In ICALP 2006 (M. Bugliesi, B. Preneel, V. Sassone and I. Wegener, eds.). Lecture Notes in Computer Science 4052 1-12. Springer, Heidelberg.

Dwork, C. and Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. Foundations and Trends in Theoretical Computer Science 9, 211-407.

Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. Journal of Official Statistics, 9, 383-406.

Fuller, WA. (2006). Measurement error models. Chichester: John Wiley and Sons.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D. and Thomas, S. (1993). A Multilevel Analysis of School Examination Results. *Oxford Review of Education*, Vol. 19, No. 4, 425-433.

Goldstein, H., Carpenter, J., Kenward, M. and Levin, K. (2009). Multilevel models with multivariate mixed response types. Statistical Modelling, 9,3, 173-197.

Goldstein, H. and Browne, W. (2016). Multilevel models: current developments. Wiley Online Library, www.wileyonlinelibrary.com/ref/stats.

Goldstein, H., Browne, WJ., and Charlton, C. (2017). An MCMC procedure for handling measurement and misclassification errors alongside missing data in multilevel multivariate generalised linear models with an application to a study of Australian youth. Journal of Applied Statistics, DOI: 10.1080/02664763.2017.1322558

Goldstein, H., Carpenter, J., and Kenward, M. (2018). Bayesian models for weighted data with missing values: a bootstrap approach. J. Royal Statistical Society, series C., DOI: 10.1111/rssc.12259

Gouweleeuw, J., Kooiman, P., Willenborg, L.C.R.J., and De Wolf, P.P. (1998). Post Randomisation for Statistical Disclosure Control: Theory and Implementation. Journal of Official Statistics, 14, 463-478.

Hundepool, A., Domingo-Ferrer, J., Francono, L., Giessing, S., Schulte-Nordholt, E., Spicer, K. and De Wolf, P.P (2012). Statistical Disclosure Control. Wiley series in survey methodology. Chichester: John Wiley and Sons.

Kim, J.J. (1986). A Method for Limiting Disclosure in Micro-data Based on Random Noise and Transformation. ASA Proceedings of the Section on SRM, 370-374.

Little, RJA (1993). Statistical analysis of masked data. J. Official Statistics, 9, 407-426.

MATLAB (2017). https://www.mathworks.com/products/matlab.html (accessed October 18 2017)

Polettini, S. and Arima, S. (2015). Small area estimation with covariates perturbed for disclosure limitation. Downloaded from https://rivista-statistica.unibo.it/article/view/5823 , December 13, 2016.

Reiter, J.P. (2005). Releasing Multiply Imputed, Synthetic Public-Use Microdata: An Illustration and Empirical Study. Journal of the Royal Statistical Society, A, Vol. 168, No. 1, 185-205.

Reiter, J. P. and Mitra, R. (2009). Estimating Risks of Identification Disclosure in Partially Synthetic Data. Journal of Privacy and Confidentiality, 1 (1), 99 - 110.

Richardson, S. and Gilks, W.R. (1993). Conditional independence models for epidemiological studies with covariate measurement error. Statistics in Medicine, 12, 1703-1722.

Rinott, Y., O'Keefe, C., Shlomo, N., and Skinner, C. (2018). Confidentiality and Differential Privacy in the Dissemination of Frequency Tables. Statistical Sciences (to be published).

Rubin, D.B. (1993). Satisfying Confidentiality Constraints through the Use of Synthetic Multiply-imputed Microdata. Journal of Official Statistics, 91, 461-468.

Shlomo, N. (2010). Measurement Error and Statistical Disclosure Control. In *PSD'2010 Privacy in Statistical Databases*, (Eds. J. Domingo-Ferrer and E. Magkos), Springer LNCS 6344, pp. 118-126.

Shlomo, N. and Skinner, C.J. (2010). Assessing the Protection Provided by Misclassification-Based Disclosure Limitation Methods for Survey Microdata. Annals of Applied Statistics, Vol. 4, No. 3 pp. 1291-1310.

Ting, D., Fienberg, S. and Trottini, M. (2008). Random orthogonal matrix masking methodology for microdata release, International Journal on Information and Computer Security, vol. 2, no. 1, 86–105.

Willenborg, L. and De Waal, T. (2001). Elements of Statistical Disclosure Control in Practice. Lecture Notes in Statistics, 155. New York: Springer-Verlag.

Winkler, W.E. (1998). Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata, Research in Official Statistics, 1, 87-104, https://www.census.gov/srd/papers/pdf/rrs2005-09.pdf

Woo, Y.M.J. and Slavkovic, A. (2014). Generalized Linear Models with Variables Subject to Post Randomization Method. Italian Journal of Applied Statistics, Vol 24 (1), 29-56.

Appendix. Model estimation with random noise or measurement errors.

The following exposition, as introduced in section 6, is for a single level linear model with a single predictor variable that contains noise (measurement error) with known parameters. The case of several predictors with added noise, the case where we have a generalised linear model and the multilevel case follow straightforwardly. We assume multivariate normality for the added noise and where we have categorical or count variables or continuous variables for which a normalising transformation exists (Goldstein et al., 2009), then the appropriate extra steps are inserted into the MCMC algorithm to enable a random draw from underlying normal distributions. The following estimation steps are based upon those described by Goldstein et al. (2016).

Define the true values of the variable with measurement error (noise) as X_1 and those without measurement error as X_2, and $X = [X_1\ X_2]$.

Define the joint model - the measurement error model (MEM) in two parts, (1a) and (1b) and the model of interest (MOI) (1c) – see derivation below:

$$x_1 = X_1 + \gamma_1 \qquad \text{(A1a)}$$

$$X_1 = X_2^T\alpha + \gamma_2 \qquad \text{(A1b)}$$

$$Y = X\beta + e \qquad \text{(A1c)}$$

where $\gamma_1 \sim N\left(0, \sigma^2_{\_\gamma_1}\right)$, $\gamma_2 \sim N\left(0, \sigma^2_{\_\gamma_2}\right)$, $e \sim N(0, \sigma_e^2)$. We note that the MOI (A1c) may contain functions of the X_1, such as interaction or power terms. Lower case variables define observed and upper case true values, and we assume the residual terms in (A1a)-(A1c) are independent. A Metropolis step is used for record i, where a current value is proposed. If we denote this by X_{1i}, the joint log likelihood for (A1a), (A1b) and (A1c) is

$$-\{1.5log2\pi + \log\left(\sigma_{\gamma_1}\sigma_{\gamma_2}\sigma_e\right) + \frac{0.5(x_{1i}-X_{1i})^2}{\sigma^2_{\gamma_1}} + \frac{0.5\left(X_{1i}^T - X_{2i}^T\alpha\right)^2}{\sigma^2_{\gamma_2}} + \frac{0.5(\tilde{y}_i)^2}{\sigma_e^2}\} \qquad \text{(A2)}$$

where $\tilde{y}_i = y_i - X_i\beta$ and only the final three terms in (A2) are required in the Metropolis step.

For a proposal distribution we can use

$$p(X_1|x_1) \sim N(x_1R, R(1-R)\sigma^2_{x_1}) \qquad \text{(A3)}$$

where R is the reliability $= \frac{var(X_1)}{var(x_1)}$. Model (A1) is similar to the formulation by Richardson and Gilks (1993) where they have a ‘gold standard’ validation sample that provides the information contained in (A1a). In the present case, of course, the values of the noise variances are known to the data analyst.

For the case where we have > 1 variables with measurement error we can propose the set of values defined independently for each variable or look at the joint proposal distribution in the case where correlated noise has been used, namely $f(X_1|x_1) \sim MVN(X_1\Omega_{x_1}^{-1}\Omega_{X_1}, \Omega_{X_1} - \Omega_{X_1}\Omega_{x_1}^{-1}\Omega_{X_1})$, although the use of correlated noise, generally would seem to be unnecessary and serves only to complicate the analysis. Further details can be found in Goldstein, Browne and Charlton (2017).

For discrete variables where we have misclassification errors, there is an analogous procedure (Goldstein and Browne, 2016), but this becomes complicated when there are multiple categories. Instead we apply the procedure given by (A3) as follows.

For the discrete variables in X_1 in (A1a), we now have the set of category codes (0*,…,p-1*), as discussed in Section 5. Associated with such a variable we will have p-1 dummy variables D_1. The proposal distribution for the Metropolis step is conveniently chosen as the observed distribution across the categories, based on choosing the nearest integer, or alternatively the 'true' distribution could be made available by the data provider for use as the proposal distribution. The log-likelihood contribution is then given by one of the following depending on the observed value m_{ij}

$$-\{0.5log2\pi + \log(\sigma_m) + \frac{0.5\left(m_{ij}-j^*\right)^2}{\sigma_m^2}\} \quad \text{if } 0 < m_{ij} < p-1$$

$$\log(\int_{-\infty}^{0} \phi(f)df), \quad f \sim N(j^*, \sigma_m^2) \text{ if } m_{ij} \leq 0$$

$$\log(\int_{p-1}^{\infty} \phi(f)df), \quad f \sim N(j^*, \sigma_m^2) \text{ if } m_{ij} \geq p-1$$

where $j*$ is the proposed value and m_{ij} is the observed 'noisy' value truncated at zero and $p-1$.

In some cases we may wish to present perturbed categorical data to a data analyst only as a set of discrete values, for example as the nearest integer to the 'noisy' value. Thus, for example for a perturbed value in the range $(-\infty, 1.5)$ we would report the value as 1 and generally as j if it is in the interval $(j-0.5, j+0.5)$. For the likelihood contribution for a proposed value j^* we now have the likelihood contributions

$$\int_{-\infty}^{0.5} \phi(f)df \quad \text{if } m_{ij} = 0$$

$$\int_{m_{ij}-0.5}^{m_{ij}+0.5} \phi(f)df \quad \text{if } 0 < m_{ij} < p-1$$

$$\int_{p-1.5}^{\infty} \phi(f)df \quad \text{if } m_{ij} = p-1$$

For each proposed category for variable X_1 we will have a corresponding entry of '1' for the dummy variable in the model of interest, i.e. in D_1. This set of dummy variables will enter the MOI as predictors with the response vector corresponding to (A1b), where the default link function is the multivariate probit as described in Goldstein et al. (2009). If we wish to allow actual measurement errors for continuous predictors as well as the imposed anonymization categorical measurement errors, it will be convenient to propose true values for the former in a separate step, conditional on all the current categorical predictor values. Where we have imposed anonymization measurement errors for continuous variables as well as actual measurement errors, we may simply add the variances for the former to the corresponding diagonal terms of the actual measurement error covariance matrix.

Standard errors as quoted in the tables are the standard deviations computed from the MCMC chains in the usual way.

As mentioned in Section 6, in some cases we may have additional constraints on the data values. For example, if the perturbed value x_1 is constrained to be no larger than a chosen value, we may have

$if(x_1 > C_1), set\ x_1 = C_1$

where the value C_1 does not occur in the dataset. Then, whenever $x_1 = C_1$, we would sample a value from the upper tail of the normal distribution defined by the current values from (A1b) and use this in the Metropolis step. Likewise, where we apply truncation to a continuous response variable, an extra step will be introduced into the algorithm that samples from the tail area of the normal distribution conditional on current parameter values. A similar procedure could be used more generally where a grouping of values takes place. Care will be needed, however, to ensure that there is not too much loss of efficiency associated with this.

For a response variable with added noise, the situation is more complex but for normally distributed responses we can carry out an analysis using the observed values to obtain consistent estimates of the fixed coefficients. We can obtain a consistent estimate for the residual variance by using the observed variance during the chain sampling and subtracting the (known) variance for the measurement error in the response, say σ_δ^2, from the final estimated residual variance. As we show in our example this may result in a large increase in the standard errors when the measurement error variance is large relative to the residual variance. When data are released it may be acceptable to provide any variable(s) to be treated as a response with the true values or just a small amount of noise, and as we show in our example in Section 9, this may not be too disclosive.

For categorical responses with misclassification or measurement errors a further modification is required. Thus, for example, in the case of a binary response where normally distributed noise has been added as in (3), if we choose, as above, to round the observed value to the nearest integer (0,1), denoted by y_i, then for the likelihood for the model of interest we can write

$$\delta_1 = \Pr(obs = 1|true = 0) = \int_{0.5}^{\infty} \phi_\delta(t)dt) \quad , \quad \Pr(obs = 1|true = 1) = 1 - \int_{0.5}^{\infty} \phi_\delta(t)dt)$$

and this leads to

$$\Pr(y_i = 1|X\beta) = \delta_1 + (1 - 2\delta_1) \int_{-X\beta}^{\infty} \phi(t)dt, \ \ \phi_\delta \sim N(0, \sigma_\delta^2) \qquad \text{(A4)}$$

Thus for the β parameters, since σ_δ^2 is known, we will have a Metropolis step for each one in turn using the observed (0,1) values, subject to $\delta_1 + (1 - 2\delta_1) \int_{-X\beta}^{\infty} \phi(t)dt < 1$.

Routines to implement the models described in this appendix have been written in MATLAB (2017) and details can be obtained from the first author.