ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2020)

Adding measurement error to location data to protect subject confidentiality while allowing for consistent estimation of exposure effects

Mahesh Karra

Boston University, USA

and David Canning and Ryoko Sato

Harvard T.H. Chan School of Public Health, Boston, USA

[Received July 2018. Final revision July 2020]

Summary. In public use data sets, it is desirable not to report a respondent's location precisely to protect subject confidentiality. However, the direct use of perturbed location data to construct explanatory exposure variables for regression models will generally make naive estimates of all parameters biased and inconsistent. We propose an approach where a perturbation vector, consisting of a random distance at a random angle, is added to a respondent's reported geographic co-ordinates. We show that, as long as the distribution of the perturbation is public and there is an underlying prior population density map, external researchers can construct unbiased and consistent estimates of location-dependent exposure effects by using numerical integration techniques over all possible actual locations, although coefficient confidence intervals are wider than if the true location data were known. We examine our method by using a Monte Carlo simulation exercise and apply it to a real world example using data on perceived and actual distance to a health facility in Tanzania.

Keywords: Data privacy; Measurement error; Monte Carlo simulation; Numerical integration; Tanzania

1. Introduction

Many data sets involving human subjects contain sensitive information and require the protection of subject confidentiality. This protection of subject confidentiality can, however, limit access to the data, which in turn may limit the scope for valuable research. A common approach to addressing this issue is to make available public use data sets that mask, or perturb, information that would enable the direct identification of subjects. However, perturbing reported data introduces measurement error, and it is well known that measurement error in an explanatory variable in a regression usually biases the estimate of its coefficient towards zero (attenuation bias) and can render the coefficient estimates on all variables inconsistent (Aigner, 1973; Carroll *et al.*, 2006; Hausman, 2001). We address the issue of perturbing reported public use data in a way that protects subject confidentiality while also enabling external researchers to conduct consistent and unbiased estimations of effects.

Although the general issue is well known, we focus on the important special case of the geographic co-ordinates of the respondent's location. Reporting the exact co-ordinates of a

Address for correspondence: Mahesh Karra, Frederick S. Pardee School of Global Studies, Boston University, 152 Bay State Road, Boston, MA 02215, USA.
E-mail: mvkarra@bu.edu

 0035–9254/20/69000

respondent's household will clearly identify the individual; however, perturbing the reported location will introduce measurement error when exposures that depend on location are calculated. Monte Carlo simulation studies of the effect of using perturbed location rather than actual locations to measure the effect of distance to a facility have shown a large downward bias in parameter estimates (Arbia *et al.*, 2015). Comparable results were found in a simulation study where the actual household location data were known and results by using the actual locations were compared with those by using perturbed locations (Elkies *et al.*, 2015). Using data from South Africa, the study found that imposing a scrambling radius of up to 5 km on household co-ordinates yielded overestimates of average distances from households to the nearest clinic or school by about 5%, and corresponding regression coefficients on the mismeasured distance variables were up to 36% smaller than the true effects. In reconciling these results with the existing empirical literature, the extent to which the evidence of the effect of proximity to services on service utilization (e.g. healthcare seeking or school attendance) and other outcomes remains mixed may, to a large degree, be explained by the size of the potential attenuation bias that is induced by mismeasured distance. On the one hand, studies that have calculated distance to the nearest health facility by using perturbed location data from data sets like the demographic and health surveys (DHSs) have found that distance is not significantly associated with outcomes such as child mortality (Lohela *et al.*, 2012). In contrast, studies that used actual distances to the nearest health facility found positive and significant effects of increased proximity to care on both health service utilization (facility delivery or receipt of antenatal care) and mortality outcomes (Schoeps *et al.*, 2011; Karra *et al.*, 2017).

Existing approaches do not fully solve the problem of estimation of distance effects by using perturbed location data because of the complexity of the problem. For example, although it can be shown that the expected measurement error in distances has positive mean and is bounded when measuring the distance to a single fixed point (Elkies *et al.*, 2015), when measuring distance to the nearest of a set of facilities, it is possible that the 'nearest' facility may change when noise is added to the household location, and the distribution of the errors that are induced depends on the precise locations of the facilities. If we assume that the distributions of actual distances and of the measurement errors in distances that are induced by using perturbed locations are independent and normally distributed, then we can apply standard regression calibration methods (Warren *et al.*, 2016). However, none of these assumptions are true (Arbia *et al.*, 2015; Elkies *et al.*, 2015), which makes this method unlikely to solve the problem. We investigate how to perturb location data and undertake consistent estimation with this perturbed data for the case of measuring the relationship between measured distance to facility and subjective distance. However, our method can be more generally applied to any estimation problem where an exposure variable varies with the true location of the household.

The issue that we address in this study is very closely related to the general problem of estimation with measurement error in an explanatory variable. Various approaches have been proposed for dealing with such measurement error, including regression calibration (Hardin *et al.*, 2003; Spiegelman *et al.*, 1997) maximum likelihood methods (Rabe-Hesketh *et al.*, 2003a, b) and, most recently, Bayesian Markov chain Monte Carlo algorithms (Goldstein and Shlomo, 2020). Usually, measurement error is inadvertent and unknown, and the structure of the error is estimated from replication or validation data. In our approach, as in Goldstein and Shlomo (2020), we have the advantage that the structure of the measurement error that is added to protect subject confidentiality is known and can be shared with external researchers. This removes the need for replication or validation data, provided that the error generation process is replicable.

Perhaps the closest parallel in the literature to our approach is the use of randomized response or list experiment methods to elicit information on sensitive topics, such as sexual activity or

health status, where a respondent may not want to reveal information directly to the interviewer. For example, a respondent can be asked to throw a die privately and to respond 'no' on a 1 and 'yes' on a 6, but to answer the sensitive question truthfully on other numbers. This randomized response approach clearly introduces measurement error in the response but helps to maintain subject confidentiality with respect to their status. If people follow the protocol as described, the fact that the error structure in the reported variable is known can be used to construct consistent estimates of the effect of status as an explanatory variable in a regression (Blair *et al.*, 2015). Similarly, in list experiments, the respondent provides the aggregate number of 'yes' answers to a list of questions that may include a sensitive question rather than answer the sensitive question directly. Comparing this with responses by others to the same list, but without the sensitive question, provides noisy evidence on the status of the subject with respect to the sensitive question which can then be used to undertake consistent estimation (Imai *et al.*, 2015).

Our approach is based on the fact that we can consistently estimate the parameters of a linear regression so long as we can first obtain the expected value of the true variable of interest given the perturbed data for each observation. We show that we can construct this expected value by integrating over all the possible actual values of the true location, weighted by the conditional probability that the location is correct given the observed perturbed data and other observed variables in the data set. These conditional probabilities can be derived provided that we know both how the measurement error is generated, and we have a prior distribution of the true variable for observations in the data set. Essentially, we replace our perturbed variables with new variables that contain only Berkson errors that allow consistent estimation in linear models (Fuller, 2009). Although we obtain consistent estimates, we show that our parameter estimates are less precise (i.e. have larger confidence intervals) than those estimates that could be obtained if we knew the true values of the variable in the survey. In addition to knowing the distribution of the perturbation, our approach requires an independent source for obtaining the underlying true distribution of the household location data. We require a population density map for the unconditional probability that an individual resides at a particular location.

In Section 2, we derive our theory that enables consistent estimation based on replacing the unknown true exposure with the expected value of the exposure given the reported perturbed location. We describe a method for calculating the expected value of the explanatory exposure variable based on numerical integration over all possible true locations. We show how the method can be applied to the case of perturbing household location data by a random distance at a random angle, and we conduct a Monte Carlo simulation study of the effect of distance to nearest facility on healthcare utilization to show that our method overcomes the bias in the estimates that results from using perturbed location data directly.

In Section 3, we apply our method by using survey data from Arusha, Tanzania. In this data set, we have household locations as well as the locations of family planning health facilities, and we estimate the association between the measured distance to the family planning facility that is visited by the respondent (our explanatory variable) and the respondent's perceived (subjective) distance to the facility. We first estimate the association by using measured distances that are calculated with the actual household location data and then using measured distances that are calculated with perturbed household location data, where we add a uniformly distributed random distance at a uniformly distributed random angle to the measured household co-ordinates. We compare these results with the results from our numerical integration method, which derives consistent estimates with the perturbed data. We also investigate the effect of changing the maximum possible location perturbation and of changing the grid mesh that is used for the numerical integration on the estimates.

In an on-line appendix, we provide three examples that illustrate how our method can be used in other cases where data are masked to protect confidentiality. We begin with the example where the respondent's location is not reported as a point but is coarsened to an area, such as a zip code, state or region. We also give the similar case where a single continuous variable is reported as lying within an interval. Our final example is the case in which a normally distributed noise term is added to a normally distributed underlying variable, and we show that our method gives results that are the same as using the standard regression calibration approach but with a known measurement error structure.

The programs that were used to analyse the data can be obtained from

`https://rss.onlinelibrary.wiley.com/hub/journal/14679876/series-c-datasets`.

2. Theory

We begin with a simple case showing how perturbed data can be used to generate consistent estimates in general. Suppose that a variable that is reported with error has a true value that is a real number x drawn from the set X and we wish to estimate a relationship of the form

$$y_i = \alpha + \beta g(x_i) + \gamma z_i + \varepsilon_i \tag{1}$$

where y_i is the outcome, the exposure $g(x_i)$ is a known function g of the variable x_i, z_i is a covariate (generalization to many covariates is straightforward) and ε_i is a random-error term with mean 0 and which is uncorrelated with both x_i and z_i. In this model, x_i and z_i are realizations of random variables. In what follows, we assume, for simplicity, that the errors ε_i in equation (1) are independent and identically distributed and follow a normal distribution. In the data, x_i is not observed, but we do observe m_i, which is a perturbation of x_i and potentially conditional on z_i, where we assume that the probability density function of the error generation process, which is denoted $p(m|x, z)$, is known.

It is well known that simply replacing x_i with m_i will typically make the results estimates of equation (1) inconsistent and biased. However, suppose that we can calculate the expected value

$$E[g(x_i)|m_i, z_i] = \int_X g(x)\, p(x|m_i, z_i)\mathrm{d}x.$$

Now set

$$u_i = g(x_i) - E[g(x_i)|m_i, z_i].$$

Here, u_i is the gap between the true value $g(x_i)$ and our calculated expected value.

Since $g(x_i) = E[g(x_i)|m_i, z_i] + u_i$, we can rewrite the estimating equation as

$$y_i = \alpha + \beta E[g(x_i)|m_i, z_i] + \gamma z_i + v_i, \qquad v_i = \beta u_i + \varepsilon_i. \tag{2}$$

For the linear model this regression calibration approach, replacing the perturbed exposure with the expected value of the true exposure given the perturbed measure, is known to produce unbiased and consistent estimates (Carroll *et al.*, 2006). To see this, we note that, by the law of iterated expectations, we have for any m_i and z_i

$$E[g(x_i)|m_i, z_i] = E[\{E[g(x_i)|m_i, z_i] + u_i\}|m_i, z_i] = E[g(x_i)|m_i, z_i] + E[u_i|m_i, z_i].$$

Hence, we have $E[u_i|m_i, z_i] = 0$ for all m_i and z_i, and we see that u_i is an error term with expected value 0 for all values of m_i and z_i and is therefore uncorrelated with both m_i and z_i. Similarly for all m_i and z_i,

$$E[u_i|E[g(x_i)|m_i,z_i]] = E[\{g(x_i) - E[g(x_i)|m_i,z_i]\}|E[g(x_i)|m_i,z_i]]$$
$$= E[g(x_i)|m_i,z_i] - E[g(x_i)|m_i,z_i] = 0.$$

Hence, the error term in equation (2), $v_i = \beta u_i + \varepsilon_i$, is mean 0 and uncorrelated with either of the explanatory variables, $E[g(x_i)|m_i,z_i]$ or z_i. It follows that, by replacing the unknown $g(x_i)$ in the regression with $E[g(x_i)|m_i,z_i]$, we can estimate equation (2) by using standard ordinary least squares methods to obtain consistent and unbiased estimates of α, β and γ. Moreover, since all of the classical assumptions of ordinary least squares are satisfied following the correction, the standard errors of the parameter estimates will also be unbiased and consistent.

Although we have unbiased and consistent estimates, the variance of the error term v_i, using our calculated expected values of the explanatory variable, will be larger than the variance of the error term ε_i when estimating using the true values. This will make our estimates less precise, and we would expect to obtain larger standard errors than if we used the true explanatory variable. However, although our estimates are less precise, the standard errors will be correctly estimated, and hypothesis tests will be correctly sized and will have the correct type 1 error rate.

Our results depend on the fact that the underlying relationship that we want to estimate is linear. There is a large literature applying this regression calibration approach in non-linear models when the consistency and unbiasedness results do not hold, though in practice the bias in non-linearity is often found to be small (Buonaccorsi, 2010; Carroll *et al.*, 2006).

To calculate $E[g(x_i)|m_i,z_i]$, we can use Bayes rule to give

$$E[g(x_i)|m_i,z_i] = \int_X g(x)\,p(x|m_i,z_i)\mathrm{d}x = \int_X g(x)\frac{p(m_i|x,z_i)\,p(x|z_i)}{\int_X p(m_i|x,z_i)\,p(x|z_i)\mathrm{d}x}\mathrm{d}x. \tag{3}$$

We can calculate this expectation, provided that we know the exposure at each possible true value x given by $g(x)$, the mechanism that was used to induce the error structure, $p(m_i|x,z_i)$, and the underlying probability density function of the true values x given z_i, given by $p(x|z_i)$. It is important to condition the expectation on covariates when they are present since they may affect how the perturbed variables are generated. Again, by Bayes rule, we have

$$p(x|z_i) = \frac{p(x)\,p(z_i|x)}{p(z_i)}$$

and hence

$$E[g(x_i)|m_i,z_i] = \int_X g(x)\frac{p(m_i|x,z_i)\,p(z_i|x)\,p(x)}{\int_X p(m_i|x,z_i)\,p(z_i|x)\,p(x)\mathrm{d}x}\mathrm{d}x. \tag{4}$$

This reduces the issue of knowing $p(x|z_i)$ to knowing the unconditional population density at each point x, and knowing the probability of observing the covariate z_i conditionally on being at that location.

In what follows, we assume for notational simplicity that we estimate a model without covariates so that we have

$$E[g(x_i)|m_i] = \int_X g(x)\,p(x|m_i)\mathrm{d}x = \int_X g(x)\frac{p(m_i|x)\,p(x)}{\int_X p(m_i|x)\,p(x)\mathrm{d}x}\mathrm{d}x. \tag{5}$$

Although this can be calculated in principle, it may be difficult to calculate analytically if the functions are complex. For complex cases, suppose that we divide the range of x, given by

the interval $[x_{\min}, x_{\max}]$, into an evenly spaced grid with $S+1$ points and S intervals, at x_s for $s=0,\ldots,S$, where $x_0 = x_{\min}$ and $x_S = x_{\max}$. Let the mesh of the grid be denoted as

$$h = |x_{s+1} - x_s| = \frac{x_{\max} - x_{\min}}{S}.$$

Then, we have that, for all $\varepsilon > 0$, there is an $h_0 > 0$ such that, for $h < h_0$, we have

$$\left| \int_X g(x) \frac{p(m_i|x)\,p(x)}{\int_X p(m_i|x)\,p(x)\mathrm{d}x} \mathrm{d}x - \sum_{s=0}^{S-1} g(x_s) \frac{p(m_i|x_s)\,p(x_s)h}{\sum_{s=0}^{S-1} p(m_i|x_s)\,p(x_s)h} \right| < \varepsilon$$

by the definition of the Riemann integral, provided that the functions $g(x)$, $p(m_i|x)$ and $p(x)$ are continuous almost everywhere, i.e. the set of points at which there are discontinuities is of Lebesgue measure zero. For example, functions with any finite set of discontinuities are continuous almost everywhere (Rudin, 1976). If the range of x is infinite, then we can replace the summation with fixed limits $[x_{\min}, x_{\max}]$ and we can sum over a wider range as S increases, i.e. $[x_{\min}(S), x_{\max}(S)]$. Provided that this range goes to ∞, while the mesh converges to 0, then, as S increases, we shall obtain the same result. It therefore follows that we can approximate the continuous integral arbitrarily closely through numerical integration by taking a large number of grid points S, with sufficiently small mesh h, over all possible values of x. There are many algorithms, such as importance sampling and those available through R functions `stats::integrate` and `cubature::hcubature`, to approximate the integral by numerical integration in equation (5). We code our approach in Stata and apply it to a spatial data example below.

This approach generalizes straightforwardly to the case of multiple variables measured with error. In the case of perturbed location data, there are two spatial co-ordinates (x_{i1}, x_{i2}) measured with error where we observe only the reported co-ordinate pair (m_{i1}, m_{i2}). To make things explicit, suppose that the actual exposure at the true location is $g(x_{i1}, x_{i2})$ and

$$y_i = \alpha + \beta g(x_{i1}, x_{i2}) + \varepsilon_i. \tag{6}$$

Let $p\{(m_{i1}, m_{i2})|(x_1, x_2)\}$ be the known distribution of the reported measurement given the true co-ordinate pair (x_{i1}, x_{i2}). For each observed pair (m_{i1}, m_{i2}), we wish to calculate

$$\begin{aligned} E[g(x_1,x_2)|(m_{i1},m_{i2})] &= \int_{X_2}\int_{X_1} g(x_1,x_2)\,p\{(x_1,x_2)|(m_{i1},m_{i2})\}\mathrm{d}x_1\,\mathrm{d}x_2 \qquad (7)\\ &= \int_{X_2}\int_{X_1} g(x_1,x_2)\frac{p\{(m_{i1},m_{i2})|(x_1,x_2)\}\,p(x_1,x_2)}{\int_{X_2}\int_{X_1} p\{(m_{i1},m_{i2})|(x_1,x_2)\}\,p(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_2}\mathrm{d}x_1\,\mathrm{d}x_2. \end{aligned}$$

We need to know the underlying joint distribution $p(x_1, x_2)$, a population density map, to calculate this expectation. Given these expectations, we can then estimate

$$y_i = \alpha + \beta E[g(x_1, x_2)|(m_{i1}, m_{i2})] + v_i \tag{8}$$

where, as before, the expectation is uncorrelated with the error term v_i, so this equation can be estimated by standard methods.

We propose to perturb the location data by adding a random distance at a random angle. Suppose that we have true location data for an individual i given by a pair of co-ordinates (x_{i1}, x_{i2}). These co-ordinates will exactly identify the location of the individual. To protect

respondent confidentiality, we perturb these co-ordinates by a random angle that is uniformly distributed over 0 to 2π rad and by a random distance that is uniformly chosen between 0 and δ km. The resulting displaced co-ordinates for this respondent are given by (m_{i1}, m_{i2}), which are then shared with the external researcher. It is straightforward to see that the probability density function of the perturbed data given the true data is

$$p\{(m_{i1}, m_{i2})|(x_{i1}, x_{i2})\} = \begin{cases} 0 & \text{if } \sqrt{\{(m_{i1}-x_{i1})^2+(m_{i2}-x_{i2})^2\}} > \delta, \\ \dfrac{1}{2\pi\delta\sqrt{\{(m_{i1}-x_{i1})^2+(m_{i2}-x_{i2})^2\}}} & \text{if } \sqrt{\{(m_{i1}-x_{i1})^2+(m_{i2}-x_{i2})^2\}} \leqslant \delta. \end{cases} \tag{9}$$

Now suppose that we have a set of W facilities at locations $r_w = (r_{w1}, r_{w2})$ for $w = 1, \ldots, W$. These may represent W health facilities or other locations that may affect the outcome of interest. The function g is given by the Euclidean distance from the individual to the nearest exposure unit (e.g. the nearest health facility), which is defined as follows:

$$g(x_{i1}, x_{i2}) = \min_w D[(x_{i1}, x_{i2}), (r_{w1}, r_{w2})] = \min_w \sqrt{\{(x_{i1}-r_{w1})^2+(x_{i2}-r_{w2})^2\}}. \tag{10}$$

We have chosen the function $g(x_{i1}, x_{i2})$ as the Euclidean distance between the household location and the nearest health facility. However, the exposure $g(x_{i1}, x_{i2})$ can be defined by the researcher as any function of the true location, e.g. the shortest distance by road to a facility, or a measure of air pollution at the location (x_{i1}, x_{i2}) if the researcher were investigating environmental effects.

We wish to estimate the relationship between the true minimum distance to a facility and some outcome y_i given by equation (6). If we have the true location data, we can run this regression and produce the coefficient estimates $\hat{\alpha}_x$ and $\hat{\beta}_x$. A naive estimation of the equation above would use the reported location data (m_{i1}, m_{i2}) and run the following regression:

$$y_i = \alpha_m + \beta_m g(m_{i1}, m_{i2}) + v_i \tag{11}$$

where the function $g(m_{i1}, m_{i2})$ is the minimum distance from a reported location (m_{i1}, m_{i2}) to a facility. The estimates $\hat{\alpha}_m$ and $\hat{\beta}_m$ based on this approach have been shown to be biased and inconsistent. Using our method, we take each possible true location for the individual and calculate the minimum distance from this location to a facility. We then calculate the expected minimum distance on the basis of the probability that each location is the true location given the observed perturbed location.

Exact calculation of this integral is not possible. However, we can approximate this expectation by using our numerical integration method by taking a grid of points (x'_{j1}, x'_{k2}) for $j = 0, \ldots, S$ and $k = 0, \ldots, S$ that covers the complete (x_{i1}, x_{i2}) space. Let values of each co-ordinate lie in the range $[0, X]$; then we have $\{x'_{01} = 0, \ldots, x'_{S1} = X\}$ and $\{x'_{02} = 0, \ldots, x'_{S2} = X\}$. With equal spacing h on each axis, that covers all possible locations of the individual, we can calculate

$$E[g(x_1, x_2)|(m_{i1}, m_{i2})] \approx \sum_{j=0}^{S-1}\sum_{k=0}^{S-1} g(x'_{j1}, x'_{k2}) \frac{p\{(m_{i1}, m_{i2})|(x'_{j1}, x'_{k2})\}p(x'_{j1}, x'_{k2})h^2}{\sum_{j=0}^{S-1}\sum_{k=0}^{S-1} p\{(m_{i1}, m_{i2})|(x'_{j1}, x'_{k2})\}p(x'_{j1}, x'_{k2})h^2} \tag{12}$$

where the approximation becomes arbitrarily good as the mesh size h declines to 0. Given this expectation for each individual, we can calculate the effect of the expected minimum distance on the outcome given by equation (8). Our theory implies that these corrected estimates $\hat{\alpha}_c$ and $\hat{\beta}_c$ will be consistent.

We demonstrate the advantage of this approach by using a Monte Carlo simulation. We first generate a 100×100 grid space over which 100 health facilities and 1000 respondents are randomly located. In particular, we generate 100 facilities at locations $r_w = (r_{w1}, r_{w2})$ for $w = 1, \ldots, 100$ where each co-ordinate is randomly and uniformly distributed in the range $(0, 100)$, i.e. $r_{w1}, r_{w2} \sim U(0, 100)$, and we similarly generate 1000 respondents to be at their true locations $x_i = (x_{i1}, x_{i2})$, where $x_{i1}, x_{i2} \sim U(0, 100)$.

We calculate the minimum distance from each facility to the true respondent location $g(x_{i1}, x_{i2})$, and we then generate an outcome variable y_i, which we define by equation (6), by the following relationship:

$$y_i = 1 + 1 \times g(x_{i1}, x_{i2}) + \varepsilon_i \tag{13}$$

where $\varepsilon_i \sim N(0, 1)$ is a randomly drawn error term, i.e. we generate the data on the assumption that the true values of the parameters are $\alpha = 1$ and $\beta = 1$.

We then simulate perturbed location co-ordinates to their new location $m_i = (m_{i1}, m_{i2})$. We displace the respondent co-ordinates by a random distance d that is uniform in the interval $[0, 5]$ and by a random angle that is uniform on the interval $[0, 2\pi]$. This displacement algorithm implies that the probability density function of the perturbed data given the true data is defined by equation (9), with $\delta = 5$. Since our population is uniformly distributed on the grid and the mesh is uniform across the 100×100 space, the prior probability $p(x'_{j1}, x'_{k2})$ and mesh terms cancel and our expectation of the minimum distance (equation (12)) simplifies to

$$E[g(x_1, x_2)|(m_{i1}, m_{i2})] \approx \sum_{j=0}^{S-1} \sum_{k=0}^{S-1} g(x'_{j1}, x'_{k2}) \frac{p\{(m_{i1}, m_{i2})|(x'_{j1}, x'_{k2})\}}{\sum_{j=0}^{S-1} \sum_{k=0}^{S-1} p\{(m_{i1}, m_{i2})|(x'_{j1}, x'_{k2})\}}.$$

Table 1 presents summary statistics of the estimated parameters from this simulation exercise, in which we run 1000 iterations over a grid mesh length of $h = 1$ (a 100×100 mesh space) to generate empirical distributions for the parameter estimates. The distributions of the estimated values of β for the three approaches are shown in Fig. 1. We begin with regressions using the true location data. As expected, the mean of the estimates $\hat{\alpha}_x$ and $\hat{\beta}_x$ based on the actual minimum distance data are very close to the true values of $\alpha = 1$ and $\beta = 1$. Variation in the estimates in each

Table 1. Distance to nearest facility effect estimates: Monte Carlo simulation results†

Parameter	*Mean*	*Standard deviation*	*Minimum*	*Maximum*	*Type 1 error*	*RMSE*
$\hat{\beta}_x$	0.9997	0.0094	0.9703	1.0301	0.052	0.0094
$\hat{\alpha}_x$	1.0004	0.0587	0.8193	1.1965	0.042	0.0587
$\hat{\beta}_m$	0.8604	0.0151	0.8112	0.9085	1.00	0.1404
$\hat{\alpha}_m$	1.7238	0.0951	1.4458	2.0546	1.00	0.7300
$\hat{\beta}_c$	0.9920	0.0170	0.9427	1.0460	0.054	0.0188
$\hat{\alpha}_c$	1.0524	0.0945	0.7785	1.3634	0.029	0.1080
N	1000					

†Based on 1000 replications. The x-subscript refers to estimates based on the actual location data; the m-subscript to estimates based on perturbed location data. The c-subscript gives estimates based on the corrected data, the expected distance to the nearest facility. The column type 1 error reports the proportion of simulated iterations for which a hypothesis test rejects the true parameter value ($\alpha = 1$ or $\beta = 1$) at the nominal 5% confidence level.

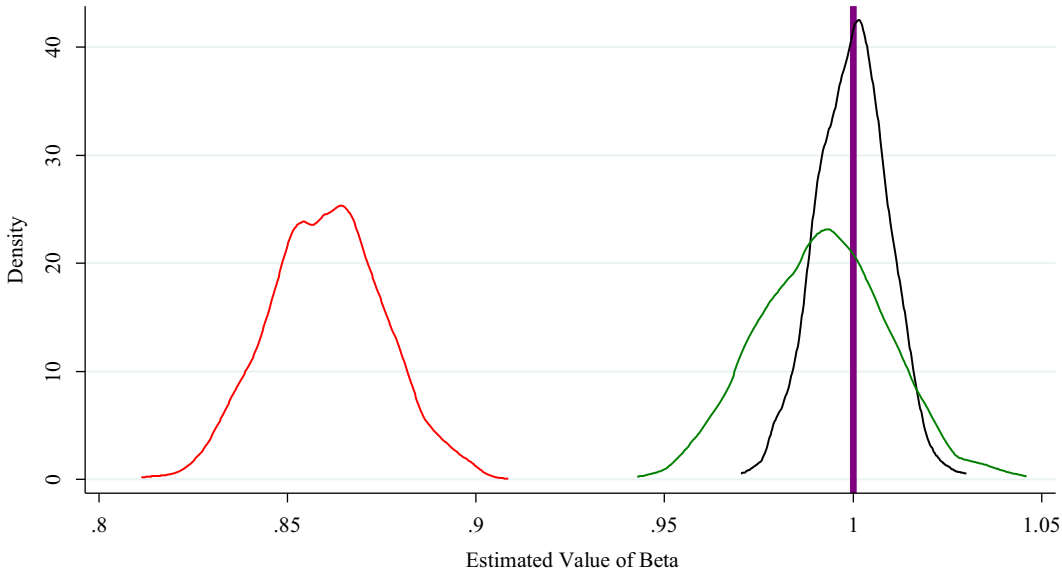

Fig. 1. Empirical distributions of parameter estimates for β from Monte Carlo simulation, 1000 iterations: ———, distribution of β based on the true explanatory variable; ———, distribution of β based on the perturbed explanatory data; ———, distribution of β based on the expected value of the explanatory data; |, true parameter value, $\beta = 1$

simulation is because we are estimating with a finite sample of 1000 household observations. The standard deviation of the estimates is small and Fig. 1 shows that the range of the estimates is small and is fairly symmetrical around the true value. We reject the null hypothesis that $\beta = 1$ at a 5% confidence level close to 5% of the time, which indicates that the hypothesis tests are correctly sized.

We then use simulated data in which the reported locations are perturbed by a 5-km random distance at a random angle. When we run regressions using the perturbed location data to construct minimum distances, we find that the estimate $\hat{\beta}_m$ is biased downwards. The mean of the estimate is well below the true value of $\beta = 1$; in fact, the estimated value is below 1 in all 1000 simulations that we run. The estimate of the intercept $\hat{\alpha}_m$ is correspondingly biased upwards. We always reject the null hypothesis $\beta = 1$ at a 5% confidence level in all 1000 simulations in spite of the fact that it is true, thereby indicating that this naive model may be very misleading for inference.

Finally, we apply our numerical integration method in each simulation to construct the expected distance to the nearest facility given the perturbed household location. When we use this expected value as a regressor, the mean of the estimates by using these corrected data, $\hat{\alpha}_c$ and $\hat{\beta}_c$ are again close to the true values of $\alpha = 1$ and $\beta = 1$. However, the standard deviations and the ranges of the estimates are somewhat larger than what are observed for the true minimum distance data, which are expected given the greater noise in the regression error when using the corrected explanatory variable. We again reject $\beta = 1$ at a 5% confidence level close to 5% of the time, which suggests that we have valid confidence intervals and correctly sized tests when using our corrected data.

The final column of Table 1 presents the root-mean-square error (RMSE) for each estimator, showing that using the true data gives the smallest RMSE whereas our corrected estimator has a somewhat larger RMSE. Moreover, estimates with the naive estimator with the perturbed data have a much larger RMSE. Fig. 1 shows the empirical distributions of the estimated values of β

for our three estimators. Fig. 1 shows graphically that the estimates based on the actual data are centred on the true value $\beta = 1$, whereas those estimates that are based on the perturbed locations give systematically lower estimates of the effect size. When using our corrected data, however, the estimates are again, on average, close to the true value $\beta = 1$, although the distribution of the coefficient estimates is wider than with the true data.

3. Application: respondents' perceived (subjective) distance to facility and measured distance to facility in Tanzania

We apply our method to real spatial data. Our location data on respondents were gathered as part of an evaluation project that was conducted with women of childbearing age in two areas of the Arusha region of Tanzania in 2017 and 2018. As part of the project, a complete household census of each area was taken, and then a random sample of households with at least one woman aged 16–44 years was selected for the survey. As part of the data collection process for the evaluation, interviewers asked respondents to estimate the distance from their household to the family planning health facility that they used. In addition, interviewers recorded the Global Positioning System co-ordinates of each sampled household by using hand-held tablets. We match these household location data with facility census data that were collected by the Ministry of Health in Tanzania, which included Global Positioning System co-ordinates of all the family planning health facilities in Tanzania in 2016 (Tanzania Ministry of Health, Community Development, Gender, Elderly, and Children, 2018). Using the household survey data and the health facility census data, we measure the distance from respondents' households to the family planning facility that they report using.

Fig. 2 shows the location of health facilities and respondents' distances to their facilities on a 100 m × 100 m grid of locations in our two areas of Arusha. Descriptive statistics of our sample are presented in Table 2. The average perceived distance to a facility is 3.6 km, though the range of perceived distances to the family planning facility varies considerably, with some respondents reporting perceived distances of up to 45 km. The average measured distance from a respondent's household to her visited family planning facility is 2.5 km. We again observe

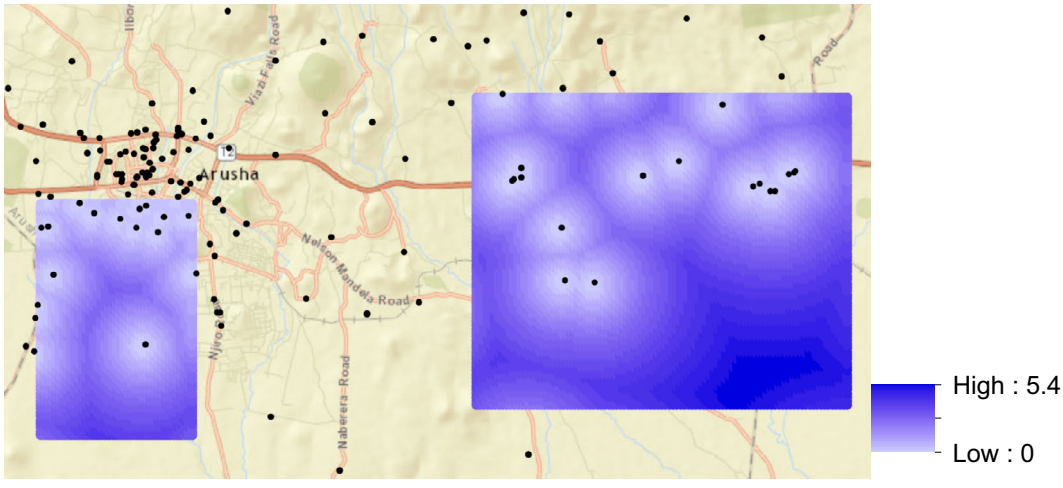

Fig. 2. Map of distances to health facilities in Arusha, Tanzania: the dots represent family planning health facilities, and the heat map presents the distance from a household to the family planning facility (in kilometres) on a 100 m × 100 m grid

Table 2. Descriptive statistics on perceived and measured distance to a family planning facility, Tanzania†

Variable	*N*	*Mean*	*Standard deviation*	*Minimum*	*Maximum*
Subjective distance to family planning facility (km)	820	3.561	5.119	0.250	45.000
Measured distance to family planning facility (km)	820	2.437	2.060	0.042	9.956
log(subjective distance to family planning facility) (km)	820	0.609	1.165	−1.386	3.807
log(measured distance to family planning facility) (km)	820	0.533	0.902	−3.161	2.298

†The unit of observation is a woman's household.

variation in the range of measured distances, but to a lesser degree than with perceived distance, with a maximum measured distance of almost 10 km.

We estimate a linear regression model of the relationship between the logarithm of the measured distance from a respondent's household to her visited family planning facility, our explanatory variable, and the logarithm of the respondent's perceived distance to the facility (equation (6)) as follows:

$$y_i = \alpha + \beta d_i + \varepsilon_i$$

where y_i is respondent i's logarithmic reported (subjective) distance to her family planning facility (in kilometres), and d_i is the logarithm of measured distance to the reported facility (in kilometres). We specify the function as logarithmic in both perceived and actual distances, which we find to be a better fit than a simple linear model. Both actual and subjective distances are always positive, and taking logarithms is therefore always feasible. If people were correct, on average, in their subjective expectations, we would expect the coefficient α to be 0 and β to be 1.

To demonstrate the validity of our approach, we compare our estimation results from calculating measured distances to the family planning facility that use the true (unperturbed) household location data, the perturbed household location data and our correction approach that estimates the expected measured distance to the family planning facility given the perturbed household locations. Unlike the simulation experiment above, we do not know the true parameter value even when using the actual household location data since we obtain only an estimate of the true parameter that includes sampling variation. To account for this variation, we run 1000 iterations of a parametric bootstrap to obtain an empirical distribution of the estimated regression coefficient β given the actual location data. The first row of Table 3 presents these bootstrapped results based on the actual location data. We estimate a value for β of $\beta_x = 0.503$, which indicates that the perceived distance increases with measured distance but much less than proportionately. The 95% confidence interval, which is constructed by using the 0.025- and 0.975-quantiles of the bootstrap replicates, is 0.429–0.576 and the coefficient is significantly different from 0 at the 5% confidence level.

We then displace household locations by adding a perturbation vector that is generated by selecting a random angle uniformly between 0 and 2π rad and a random distance δ uniformly between 0 and either 2 km or 5 km. We then re-estimate the regression coefficients by using these perturbed household location data to calculate mismeasured distances. The second and third rows of Table 3 present our results with a perturbed household location of up to 2 km or 5 km respectively. Again, we conducted 1000 replications for each set of perturbations. For each bootstrap iteration, we constructed the 95% confidence interval and then counted the number

Table 3. Coefficient estimates on distance to a family planning facility: an application in Tanzania†

Coefficient on d_i, β	*Mean*	*Standard deviation*	*Bootstrapped standard error*	*95% confidence interval*	
				Lower bound	*Upper bound*
With actual locations, β_x	0.503	0.046	0.045	0.429	0.576
With perturbed locations: $\delta = 2$ km error, β_m	0.426	0.031	0.049	0.372	0.477
With perturbed locations: $\delta = 5$ km error, β_m	0.275	0.045	0.056	0.198	0.348
With correction, $\delta = 2$ km error: $h = 100$ m, β_c	0.528	0.023	0.053	0.491	0.565
With correction, $\delta = 5$ km error: $h = 100$ m, β_c	0.481	0.047	0.075	0.403	0.556
With correction, $\delta = 2$ km error: $h = 500$ m, β_c	0.544	0.029	0.060	0.492	0.590
With correction, $\delta = 5$ km error: $h = 500$ m, β_c	0.428	0.078	0.099	0.296	0.552

†Estimates for β_x by using actual location data are generated by using a parametric bootstrap with 1000 replications. Estimates for β_m by using perturbed location data are generated with 1000 replications with the location perturbed at a random distance and at a random angle. Estimates for β_c by using the numerical integration method are calculated for the expected logarithm of the measured distance d_i given the perturbed location data for a 500 m × 500 m grid and a 100 m × 100 m grid and with 1000 replications.

of coefficient estimates that are significantly different from 0. In the current application, all measures of distance are significant predictors of subjective distance, and hence we do not report this measure. As expected, the coefficient on the distance to the family planning health facility by using the perturbed location data is, on average, attenuated towards 0, and the size of the attenuation increases with larger perturbations. These results are shown graphically in Fig. 3, which plots the distribution of estimates obtained with 1000 replications of the perturbed data under the two levels of displacement (2 km and 5 km) compared with the estimate based on using the actual household location data.

We finally implement our correction to the perturbed data by using our numerical integration method. Given a perturbed data point, we calculate the expected value of the logged distance d_i over a grid of $S \times S$ points on a mesh of size h. In calculating the expectation, we define the exposure function (equation (10)) to be $g(x'_{j1}, x'_{k2}) = D[(x'_{j1}, x'_{k2}), (r_1, r_2)]$, which is the logarithm of the distance from a respondent's household location (x'_{j1}, x'_{k2}) to the respondent's family planning health facility, where facilities are at locations (r_1, r_2). We define the probability density function to be given by equation (9), with the respondent's household co-ordinates given by (x'_{j1}, x'_{k2}). For this probability density function, we also require the term $p(x'_{j1}, x'_{k2})$: a prior population density function at the point (x'_{j1}, x'_{k2}). We obtain this for the WorldPop database (World Bank, 2015), which produces population density maps for each country. The data that we use are calculated by using census, survey, satellite and geographical information system data sets in a flexible machine learning framework that produces population density estimates on a 100 m × 100 m grid for Tanzania in 2015. Fig. 4 presents this population density map over the two areas of Arusha, Tanzania, where we conducted our survey. We note that the health facility locations are clustered in areas of high population density; the expected distance to the nearest facility is therefore lower by using the population density map than if we simply assumed a uniform distribution of population over the area.

The final four rows of Table 3 present estimates of the coefficient on the expectation of the logged distance d_i-variable, β, based on a maximum perturbation distance of 2 km and 5 km by using our numerical integration approach for 1000 replications. We report results for mesh sizes of $h = 100$ m (the fourth and fifth rows of Table 3) and $h = 500$ m (the sixth and seventh

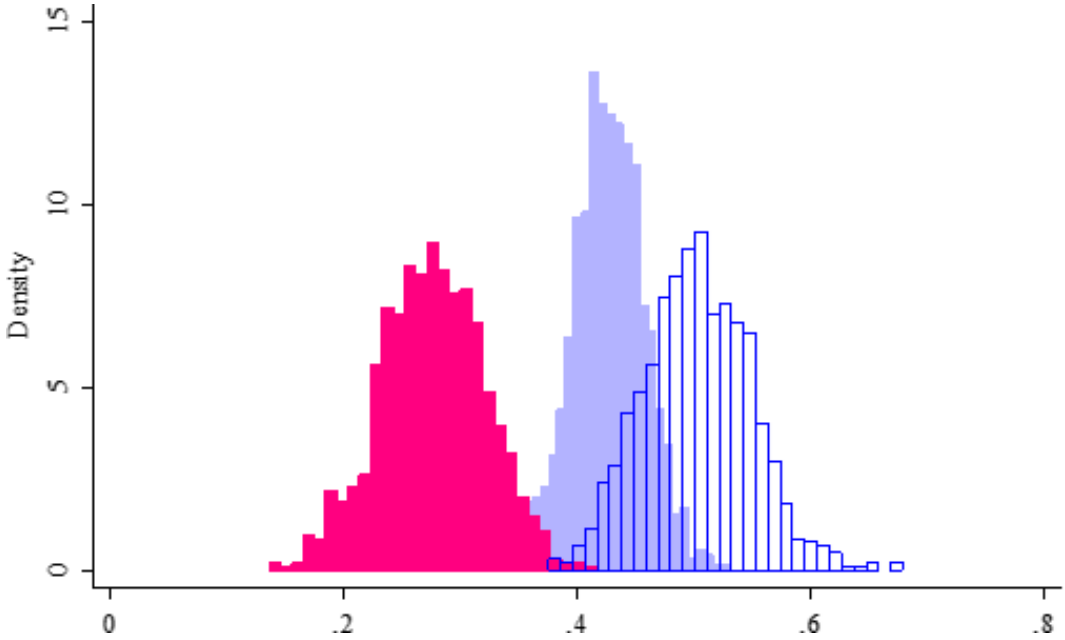

Fig. 3. Empirical distributions of parameter estimates for β by using the perturbed Tanzanian location data, 1000 iterations: comparison of the estimate for β with actual location data (□) with estimates for β based on perturbed location data with perturbations of up to 2 km (■) or 5 km (■); simulations are based on 1000 iterations

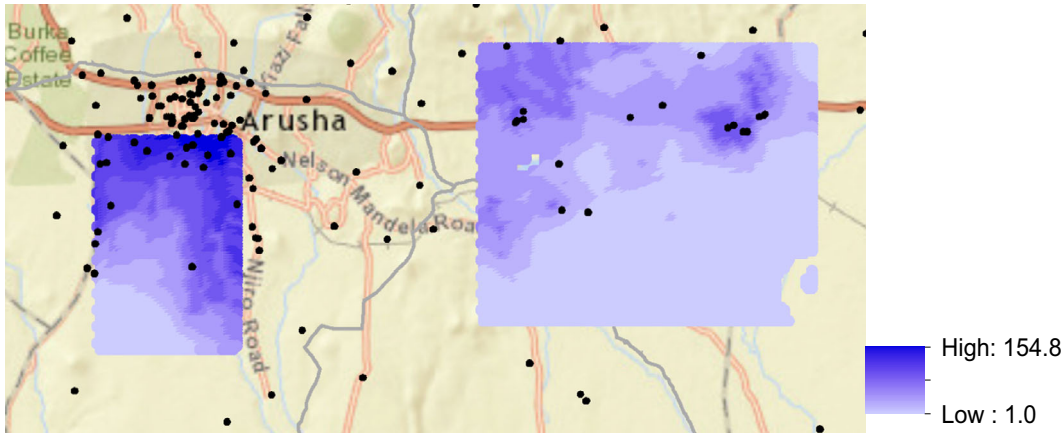

Fig. 4. Population density map in Arusha, Tanzania: the heat map presents the population density in people per 100 m × 100 m grid square (the data on population density were obtained from the WorldPop database for Tanzania)

rows of Table 3) for the numerical integration. Using a grid mesh of 100 m × 100 m, the mean of the estimates with our correction are closer to the estimate with the actual location data of $\beta_x = 0.503$. With a larger mesh, the correction appears to be close, but not quite as accurate. Our results of the empirical distributions of the parameter estimates for β are presented graphically in Fig. 5, which plots the distributions of the coefficient estimate under 1000 replications by using the actual location data, perturbed data for a perturbation of up to 5 km and corrected

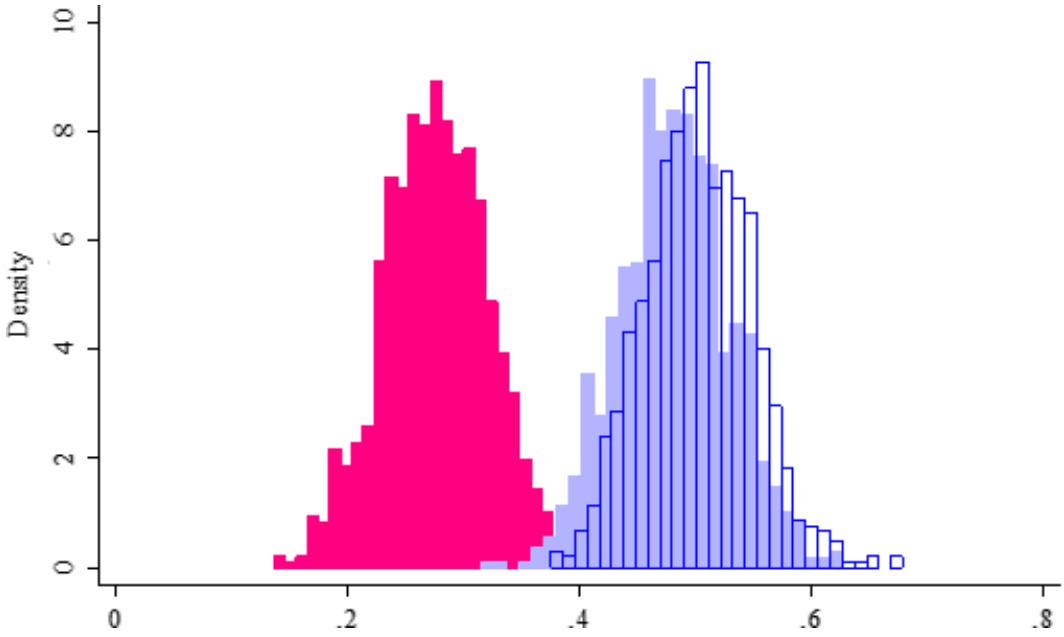

Fig. 5. Empirical distributions of parameter estimates for β from the Tanzania analysis, 1000 iterations: comparison of estimates of the unadjusted correlation between the logarithm of the measured distance to the family planning facility, d_i, and the logarithm of perceived distance to the facility with actual location data (□) with estimates based on perturbed location data (■) and with corrected distance estimates (■); replications with actual data are generated by using a parametric bootstrap; replications with perturbed data use a random angle and random distance up to 5 km; replications with corrected distance estimates are calculated for the expected logarithm of measured distance d_i by using numerical integration over a 100 m × 100 m grid; all results are generated by using 1000 iterations

estimation through numerical integration using a 100 m × 100 m mesh. Whereas the averages of the estimates with corrected data are close to the estimates by using the actual location, showing unbiasedness, the standard errors and confidence intervals are wide because of the large size of perturbation in the location data. The results show that our correction can produce unbiased results; however, with a large perturbation, the confidence interval is also very large. With an average distance to a family planning health facility of around 2.4 km, and a maximum distance of just under 10 km, a perturbation of up to 5 km in the household location considerably reduces the signal-to-noise ratio in the data and generates wide confidence intervals for our corrected estimates.

Our results indicate that our numerical integration method works well in a real world example by producing unbiased estimates with perturbed location data that are close to estimates that are obtained with the actual location data, although the confidence intervals that we estimate are wider than if the actual location data were used. For simplicity, we report results without covariates, although we find very similar results when we adjust for covariates such as women's age and education level.

The Stata code that we use to run our numerical integration approach for the Tanzania data and for the Monte Carlo simulation exercise is publicly available to researchers from `https://www.hsph.harvard.edu/david-canning/data-sets/`. To run the code, researchers are required to specify a grid mesh, to provide the values for the exposure function $g(x'_{j1}, x'_{k2})$ over this grid space, to specify the maximum location perturbation distance δ

and to provide a prior population density distribution over the grid space, $p(x'_{j1}, x'_{k2})$. Given these inputs, the expected value $E[g(x_1, x_2)|(m_{i1}, m_{i2})]$ will be calculated for the set of observed perturbed locations (m_{i1}, m_{i2}).

4. Recommendations and conclusions

In this study, we propose an approach for consistent inference under circumstances where location data are deliberately reported with error. Our approach is based on calculating the expected value of the true exposure variable given reported perturbed data. Our approach relies on knowing how the measurement error is constructed and also requires knowledge of the underlying probability density function (e.g. a population density map). We conduct a Monte Carlo simulation and use spatial data from Tanzania to demonstrate that replacing exposure data that are based on perturbed locations with the expected exposure yields improved coefficient estimates. Our study contributes to the literature on estimation and inference with perturbed location data and improves on existing regression calibration methods that assume normality of the induced distance errors, which is unlikely to be true.

Although our method is an improvement on existing approaches, it has some weaknesses. First, our two-stage approach in calculating the expected exposure and using this in the regression of interest is consistent only in linear regressions. Consistent estimation in non-linear models will require a maximum likelihood approach as suggested by Fuller (1993) and Little (1993). Secondly, our estimation approach also requires that researchers have access to a prior distribution (in the case of our examples, a population density map), which may not always be so.

Our approach does, however, enable us to make a series of recommendations for holders of data sets who wish to create pubic use versions that mask subject locations. There is a key conflict between fulfilling our dual responsibilities as researchers to making high quality inferences on data that are meaningful for policy and social welfare while simultaneously ensuring that the data are protected to preserve the confidentiality of the respondents who have consented to our use of their (potentially sensitive) information. These two objectives are intricately linked, given that the preservation of respondent confidentiality by the researcher, in addition to being a legal requirement, is also essential for the preservation of trust between the researcher and respondent, which in turn is linked to data accuracy (Charest, 2010; Reiter, 2012). As the amount of microdata continues to increase, there is a commensurate increase in the need for privacy protecting methods of data analysis that simultaneously minimize loss of information (Armstrong *et al.*, 1999; Brand, 2002). It is therefore important that we, as researchers, take into account the extent to which our proposed method may be used to make inferences on the data such that the original privacy guarantees to respondents are preserved; this criterion speaks to the extent to which differential privacy, which is considered to be the *de facto* standard for privacy preservation, is upheld following the application of our method to masked data (Dwork and Lei, 2009; Dwork and Smith, 2009).

With these considerations in mind, we make five recommendations to data holders when perturbing location data for public use data sets:

(a) to make the perturbation process simple and replicable,
(b) to make transparent the methods that were used to generate and add the perturbations to location data,
(c) to ensure access to, or to make available, a population density map for the population from which the sample is drawn,

(d) to make available a covariate intensity map for key covariates that were used in the perturbation process, particularly if the range of location perturbations is large, and
(e) to assess the trade-off between larger measurement error, and more masking of subjects, against the larger confidence intervals in estimates that result.

On recommendation (a) we suggest a perturbation based on adding a random distance at a random angle; however, any process that is simple and replicable would be acceptable. In the online appendix, we examine the case of reporting an administrative area of residence, e.g. zip code. Provided that this process is made transparent in line with recommendation (b), the researcher can use the process to perform the correction that we suggest. Population density maps for the world are widely and publicly available; for example, a gridded map of the world's population can be found through the Center for International Earth Science Information Network (Center for International Earth Science Information Network—CIESIN—Columbia University, 2018), whereas country level population density data can be obtained through the World Bank's WorldPop database (World Bank, 2015). As long as the data holder clearly defines the geographical area from where the sample was selected, the researcher should be able to construct a suitable prior population density. Alternatively, the data holder could construct their own smoothed population density map based on their sample and release this map in accordance with recommendation (c). Our model suggests that knowing the geographical distribution of a covariate can improve the calculated expected exposure. For example, if we know the ethnicity of the respondent, and the distribution of ethnic groups varies geographically, we can improve the location estimate by adjusting for this variation. In principle, the data holder could follow recommendation (d) by producing maps showing the relative frequency of different ethnic groups in different areas based on their sampling strategy. However, it seems overly onerous to do this for every possible covariate and, for small perturbations, the assumption that the relative frequency of covariates across the range of possible locations is uniform might be a reasonable approximation. Finally, recommendation (e) illustrates the key trade-off that is faced by data holders; increases in the size of the geographic perturbation provides for greater masking and increased confidentiality but at the cost of increased noise in the corrected estimation and less precise confidence intervals for estimates. We present some evidence of this trade-off in our examples with Tanzanian data.

The choice of perturbation process and the amount of noise to add raises the issue of the extent to which it is possible to reverse the perturbation process and to identify subject locations (Xiao and Xiong, 2015; Zandbergen, 2014). The random perturbation mask that we advocate for location data has been shown to be a robust type of mask in protecting respondent confidentiality (Qardaji *et al.*, 2013; Rushton *et al.*, 2006; Zandbergen, 2014), although the issue of designing an optimal mask that maximizes confidentiality while minimizing induced errors in estimation has not been fully resolved. Current studies of this issue have often focused on the issue of subject identification based on masked location data alone; however, a particularly concerning development is that a combination of several masked variables may jointly enable subject reidentification. For example, recent studies with US data have shown that knowledge of an individual's age in months, gender and five-digit zip code were sufficient for the reidentification of many individuals in anonymized data (Sweeney, 2015; Sweeney *et al.*, 2017, 2018). These findings illustrate that ensuring confidentiality in public use data sets may be more complex than simply ensuring lack of identification based on location data alone.

Recommendation (a) may not be as easy to implement as it appears. For example, the DHSs are nationally representative cross-sectional surveys that cover a range of demographic, social and health topics (USAID and ICF Macro International, 2014). To protect the identity of interviewed households, the DHS masks the precise location data of the cluster that are collected

as part of the survey (Burgert *et al.*, 2013; Perez-Heydrich *et al.*, 2013). The process that is used to mask the location data is publicly available. Reported location co-ordinates in the DHS are perturbed by adding a randomly generated distance, at a randomly generated angle, to the true location. Urban locations are perturbed by up to 2 km at a random angle, whereas most rural locations are perturbed by up to 5 km, a randomly selected 1% of rural locations are perturbed by up to 10 km. These conditional perturbations can be controlled for, since the urban or rural nature of the cluster is reported. However, if the displaced location is outside the administrative region that the cluster is in, the DHS takes a new draw of the displacement. This redrawing of the perturbation is repeated until a location within the region is obtained. This final condition may involve an arbitrarily large number of redraws, conditionally on the outcome of the draws made and exact shape of the regional border near the location. This potentially infinite process can be shown to have a convergent probability limit and, in theory, can therefore be modelled by researchers; however, its complexity makes practical applications difficult.

Our approach can resolve a key issue of conducting consistent inference with perturbed location data. However, the procedure requires that the holder of the data implement the perturbation in a way that is transparent and replicable while continuing to maintain respondent confidentiality. In addition, the researcher would require auxiliary information to be able to form a prior distribution of the true locations as well as any additional information of the geographical distribution of covariates that were used as part of the perturbation process. For this, we recommend that this auxiliary distributional information be made available along with the perturbed data. Further investigation is needed to assess the extent to which our method enables the preservation of differential privacy for data protection while simultaneously enabling the researcher to conduct consistent inference.

Acknowledgements

The authors thank Günther Fink, Noah Haber, Carlos Riumallo-Herl, session participants at the 2017 International Health Economics Association Congress, the 2017 International Union for the Scientific Study of Population International Population Conference, and the referees for their reviews and comments.

The authors received no specific funding for this research.

References

Aigner, D. J. (1973) Regression with a binary independent variable subject to errors of observation. *J. Econmetr.*, **1**, 49–59.

Arbia, G., Espa, G. and Giuliani, D. (2015) Measurement errors arising when using distances in microeconometric modelling and the individuals' position is geo-masked for confidentiality. *Econometrics*, **3**, 709–718.

Armstrong, M. P., Rushton, G. and Zimmerman, D. L. (1999) Geographically masking health data to preserve confidentiality. *Statist. Med.*, **18**, 497–525.

Blair, G., Imai, K. and Zhou, Y.-Y. (2015) Design and analysis of the randomized response technique. *J. Am. Statist. Ass.*, **110**, 1304–1319.

Brand, R. (2002) Microdata protection through noise addition. In *Inference Control in Statistical Databases* (ed. J. Domingo-Ferrer), pp. 97–116. Berlin: Springer.

Buonaccorsi, J. P. (2010) *Measurement Error: Models, Methods, and Applications*, 1st edn. Boca Raton: Chapman and Hall–CRC.

Burgert, C. R., Colston, J., Roy, T. and Zachary, B. (2013) Geographic displacement procedure and georeferenced data release policy for the Demographic and Health Surveys. *Spatial Analysis Report 7*. ICF International, Calverton.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006) *Measurement Error in Nonlinear Models*, 2nd edn. Boca Raton: Chapman and Hall–CRC.

Center for International Earth Science Information Network—CIESIN—Columbia University (2018) Gridded population of the world, version 4 (GPWv4): population density, revision 11. *Map*. Nasa Socioeconomic Data and Applications Center, Palisades. (Available from `https://doi.org/10.7927/H49C6VHW`.)

Charest, A.-S. (2010) How can we analyze differentially-private synthetic datasets? *J. Privcy Confidentlty*, **2**, no. 2, 21–33.
Dwork, C. and Lei, J. (2009) Differential privacy and robust statistics. In *Proc. 41st A. Symp. Theory of Computing*. Bethesda: Association for Computing Machinery Press.
Dwork, C. and Smith, A. (2009) Differential privacy for statistics: what we know and what we want to learn. *J. Privcy Confidentlty*, **1**, no. 2, 135–154.
Elkies, N., Fink, G. and Bärnighausen, T. (2015) "Scrambling" geo-referenced data to protect privacy induces bias in distance estimation. *Popln Environ.*, **37**, 1–16.
Fuller, W. A. (1993) Masking procedures for microdata disclosure limitation. *J. Off. Statist.*, **9**, 383–406.
Fuller, W. A. (2009) *Measurement Error Models*. New York: Wiley.
Goldstein, H. and Shlomo, N. (2020) A probabilistic procedure for anonymisation, for assessing the risk of re-identification and for the analysis of perturbed data sets. *J. Off. Statist.*, **36**, 89–115.
Hardin, J. W., Schmiediche, H. and Carroll, R. J. (2003) The regression-calibration method for fitting generalized linear models with additive measurement error. *Stata J.*, **3**, 361–372.
Hausman, J. (2001) Mismeasured variables in econometric analysis: problems from the right and problems from the left. *J. Econ. Perspect.*, **15**, 57–67.
Imai, K., Park, B. and Greene, K. F. (2015) Using the predicted responses from list experiments as explanatory variables in regression models. *Polit. Anal.*, **23**, 180–196.
Karra, M., Fink, G. and Canning, D. (2017) Facility distance and child mortality: a multi-country study of health facility access, service utilization, and child health outcomes. *Int. J. Epidem.*, **46**, 817–826.
Little, R. J. (1993) Statistical analysis of masked data. *J. Off. Statist.*, **9**, 407–426.
Lohela, T. J., Campbell, O. M. R. and Gabrysch, S. (2012) Distance to care, facility delivery and early neonatal mortality in Malawi and Zambia. *PLOS One*, **7**, no. 12, article e52110.
Perez-Heydrich, C., Bragg-Gresham, J. L., Burgert, C. R. and Emch, M. E. (2013) Guidelines on the use of DHS GPS data. *Spatial Analysis Report 8*, pp. 626–627. ICF International, Calverton.
Qardaji, W., Yang, W. and Li, N. (2013) Differentially private grids for geospatial data. In *Proc. 29th Int. Conf. Data Engineering, Brisbane*, pp. 757–768. New York: Institute of Electrical and Electronics Engineers.
Rabe-Hesketh, S., Pickles, A. and Skrondal, A. (2003a) Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statist. Modllng*, **3**, 215–232.
Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2003b) Maximum likelihood estimation of generalized linear models with covariate measurement error. *Stata J.*, **3**, 386–411.
Reiter, J. P. (2012) Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences. *Publ. Opin. Q.*, **76**, 163–181.
Rudin, W. (1976) *Principles of Mathematical Analysis*, vol. 3. New York: McGraw-Hill.
Rushton, G., Armstrong, M. P., Gittler, J., Greene, B. R., Pavlik, C. E., West, M. M. and Zimmerman, D. L. (2006) Geocoding in cancer research: a review. *Am. J. Prev. Med.*, **30**, no. 2, suppl., S16–S24.
Schoeps, A., Gabrysch, S., Niamba, L., Sié, A. and Becher, H. (2011) The effect of distance to health-care facilities on childhood mortality in rural Burkina Faso. *Am. J. Epidem.*, **173**, 492–498.
Spiegelman, D., McDermott, A. and Rosner, B. (1997) Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *Am. J. Clin. Nutrn*, **65**, suppl., 1179S–1186S.
Sweeney, L. (2015) Only you, your doctor, and many others may know. *Technol. Sci.*, article /a/2015092903/.
Sweeney, L., von Loewenfeldt, M. and Perry, M. (2018) Saying it's anonymous doesn't make it so: re-identifications of "anonymized" law school data. *Technol. Sci.*, article /a/2018111301/.
Sweeney, L., Yoo, J. S., Perovich, L., Boronow, K. E., Brown, P. and Brody, J. G. (2017) Re-identification risks in HIPAA safe harbor data: a study of data from one environmental health study. *Technol. Sci.*, article /a/2017082801/.
Tanzania Ministry of Health, Community Development, Gender, Elderly, and Children (2018) *Health Facility Registry*. Tanzania Ministry of Health, Community Development, Gender, Elderly, and Children, Dodoma.
USAID and ICF Macro International (2014) The DHS program. US Agency for International Development, Rockville. (Available from `http://dhsprogram.com/`.)
Warren, J. L., Perez-Heydrich, C., Burgert, C. R. and Emch, M. E. (2016) Influence of Demographic and Health Survey point displacements on distance-based analyses. *Spatl Demog.*, **4**, 155–173.
World Bank (2015) EnergyData.Info: Tanzania—population density (2015). World Bank, Washington DC.
Xiao, Y. and Xiong, L. (2015) Protecting locations with differential privacy under temporal correlations. In *Proc. 22nd Conf. Computer and Communications Security*, pp. 1298–1309. New York: Association for Computing Machinery.
Zandbergen, P. A. (2014) Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. *Adv. Med.*, article 2014.567049.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Appendix: additional examples'.