Published in final edited form as:

Spat Demogr. 2016 July ; 4(2): 155–173. doi:10.1007/s40980-015-0014-0.

Influence of Demographic and Health Survey Point Displacements on Distance-Based Analyses

Joshua L. Warren¹, Carolina Perez-Heydrich², Clara R. Burgert³, and Michael E. Emch⁴

¹Department of Biostatistics, Yale School of Public Health, Yale University, New Haven, CT, USA

²Department of Biological Sciences, Meredith College, Raleigh, NC, USA

³ICF International, Rockville, MD, USA

⁴Department of Geography, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Abstract

We evaluate the impacts of random spatial displacements on analyses that involve distance measures from displaced Demographic and Health Survey (DHS) clusters to nearest ancillary point or line features, such as health resources or roads. We use simulation and case studies to address the effects of this introduced error, and propose use of regression calibration (RC) to reduce its impact. Results suggest that RC outperforms analyses involving naive distance-based covariate assignments by reducing the bias and MSE of the main estimator in most settings. Proposed guidelines also address the effect of the spatial density of destination features on observed bias.

1 Background

The Demographic and Health Survey (DHS) Program data include geo-located survey cluster locations dating back to 1986 for household surveys including the DHS, AIDS Indicator Survey, and Malaria Indicator Survey. The early data were focused in West Africa and were gathered through the use of gazetteers. The first Global Positioning System (GPS) data were collected in Benin in 1996 and the collection of GPS locations for surveys has become fairly standard practice since the early 2000s. Currently, there are over 130 surveys with GPS or geo-located data. In order to protect the confidentiality of DHS survey respondents, the geo-located data are displaced before being provided to researchers (Burgert et al., 2013).

Competing interests The authors declare that they have no competing interests.

Additional file 1 — Appendix A

Appendix A is a .pdf file which shows that the random displacement of DHS cluster locations leads to a classical measurement error problem when a distance-based covariate is used.

Additional file 2 — Appendix B

Appendix B is a .txt file which provides R code to randomly displace DHS survey cluster locations based on the DHS Program geographic displacement procedure.

The displacement process used by the DHS Program shifts the true latitude and longitude of a location using a random simulation procedure with fixed settings. DHS cluster locations in urban areas are randomly displaced differently than those in rural locations such that

- Urban Locations: Randomly displaced 0–2 kilometers (km),
 - Rural Locations: Randomly displaced 0–5 km, and
- Every 100th rural location: Randomly displaced 0–10 km.

The displacement process includes a random direction and random distance component. First, a random angle of direction between 0 and 360 degrees is simulated with each value in the range having an equal probability of selection. Next, a random distance according to the urban and rural settings is chosen uniformly over the specified ranges. The new location is created by combining both random components, resulting in a shifted latitude and longitude for the original DHS cluster location. Finally, the new location is checked to ensure that it falls within the designated administrative boundaries and is re-displaced if necessary. These randomly displaced locations are then provided to researchers in a number of fields along with the corresponding survey data. Researchers do not have access to the true DHS survey cluster location information.

In past studies, the randomly displaced locations have been used by researchers in analyses which depend on distance calculations. More specifically, researchers are often interested in determining how the distance from a DHS cluster location to a particular resource location affects the outcome of interest. For example, a previous study examined the association between HIV infection in rural Malawi and distance to health clinics, major roads, and major cities (Feldacker et al., 2010). Another study explored distances to similar locations with respect to early neonatal mortality in Malawi and Zambia (Lohela et al. 2012). These types of analyses can potentially be problematic since calculating distances for displaced locations results in incorrect measurements. There are actually two problems associated with use of this direct distance measurement technique. First, researchers must determine the closest resource to a DHS cluster location in order to calculate the appropriate distance of interest. Using the displaced locations can potentially lead to incorrectly assigning a DHS cluster location to its actual closest resource location. The potential for this error varies based on the number of resources in the study area. Second, even if the true closest resource is correctly determined, the calculated distance is almost guaranteed to be incorrect due to the random displacement of the DHS cluster location.

In this locational inaccuracy setting, much of the previous literature has explored properties of the introduced error (Bonner et al., 2003; Cayo and Talbot, 2003; Ward et al., 2005; Rushton et al., 2006; Whitsel et al., 2006; Zhan et al., 2006; Goldberg et al., 2007; Kravets and Hadden, 2007; Zandbergen, 2007, 2008, 2011; Zimmerman et al., 2007, 2010; Zandbergen and Hart, 2009; Zandbergen et al., 2011) as well as its impact on subsequent spatially-based analyses (Burra et al., 2002; DeLuca and Kanaroglou, 2008; Mazumdar et al., 2008; Jacquez and Rommel, 2009; Zimmerman et al., 2010; Zinszer et al., 2010), where the errors are unintentionally introduced due to incorrect geocoding. A number of other studies have investigated the locational error that is intentionally introduced (Armstrong et al., 1999; Leitner and Curtis, 2004, 2006; Kamel Boulos et al., 2006; Olson et al., 2006;

Wieland et al., 2008), such as with DHS point displacement. Armstrong et al. (1999) investigated the impact that a number of geomasking techniques had on creating and analyzing spatially constructed variables. They found that the random point displacement geomasking technique, similar to that used by the DHS program, approximately preserved all considered spatial analyses and concluded “this mask appears satisfactory from a comprehensive information-preservation standpoint”. In this work, we further investigate when this approximation is adequate for distance-based analyses under the specific random displacement technique utilized by the DHS program. We also place more focus on the subsequent statistical analysis of interest that incorporates these distance covariates and the potential impact the displacement has on the bias and mean squared error (MSE) of the resulting estimators.

We present a simulation study which analyzes how the random displacement of DHS cluster locations affects the resulting statistical analysis and conclusions drawn from the analysis when the displacement is ignored. We also attempt to quantify the size of this error as well as propose the use of an established statistical method to account for the uncertainty introduced by the DHS cluster location displacement. The proposed method of interest is regression calibration (RC).

Specifically, we focus on the situation in which distance from the DHS cluster location to the closest resource location is used as the main continuous covariate of interest in a study. These resource locations include points, such as health facility resources, and line segments, such as road networks. We show that the random displacement of DHS cluster locations results in a classical measurement error model for the distance covariate. In the presented simulation study, we investigate the impact that this measurement error has on the statistical inference across different spatial density settings of the resources. These five spatial density settings range from dense (numerous resources located across the study area) to sparse (very few resources located across the study area). We also determine if RC can be used in this setting to improve estimation and inference.

In this study, we also investigate how the random displacement affects a cluster’s linkage to the true closest resource location when links are made using the displaced DHS cluster locations. It is common to use health facility resource-level treatment information to determine the level of care residents of a particular cluster likely received (Lohela et al., 2012). For displaced clusters however, the observed closest resource location may not be the true closest resource location of interest, leading to incorrect treatment assignment information. We explore how the displacement affects this assignment across the different spatial density settings and location types.

We carry out each of the proposed analyses for studies which consider only rural DHS cluster locations, only urban DHS cluster locations, and a mixture of location types since studies are sometimes carried out separately for urban/rural locations in practice. Separate analyses are required due to the different level of measurement error introduced to the different location types under the DHS Program geographic displacement procedure. The true Uganda DHS 2011 cluster locations (represented by points), health facility resources (points), and road networks (line segments) are used in the simulation study to ensure that

the results are realistic and useful for future research efforts. Based on the simulation study results, we propose general guidelines for analyzing DHS survey cluster location data using a distance-based covariate. We then apply these guidelines to a case study using the Uganda DHS 2011 data and results are discussed.

2 Methods

2.1 Data Description

We use the Uganda DHS 2011 data, including the GPS data (non-displaced and displaced) for the 404 clusters included in the survey (Uganda Bureau of Statistics and ICF International, 2012). GPS points represent the approximate center of a cluster of households. A total of 400 clusters were verified by DHS Program staff to be in the proper administrative areas with seven points gazetted to the nearest village and four points marked as missing which were subsequently removed from the analyses. These GPS data are used in the simulation study and in the presented case study.

2.2 Random Displacement as Classical Measurement Error Problem

In Appendix A, we show that randomly displacing the true DHS cluster locations and then calculating the distance to the closest resource location from the new displaced location results in a distance measured with error. The amount of introduced measurement error is shown to depend on the location type of the DHS survey cluster, with rural clusters potentially having a larger amount than urban clusters based on the DHS Program geographic displacement procedure. We also show that the observed distance covariate follows the classical measurement error model form (Carroll et al., 2010) such that $x_i(s_o) = x_i(s_t) + u_i$ where $x_i(s_o)$ is the distance from DHS cluster location i to closest resource location based on the displaced cluster location s_o , $x_i(s_t)$ is the distance from DHS cluster location i to closest resource location based on the true cluster location s_t , and u_i is the random measurement error component with $E(u_i) = 0$, $\text{Var}(u_i) = \sigma_u^2$, and u_i independent and identically distributed for each cluster location type (urban and rural). Note that $E(\cdot)$ and $\text{Var}(\cdot)$ represent the expected value and variance, respectively, of a random variable. We also verify these model assumptions through the use of validation data (See Figure 1). In Appendix A, we show that the amount of measurement error for each distance is bounded such that $u_i \in [-5, 5]$ km for most rural DHS cluster locations ($u_i \in [-10, 10]$ km for every 100th rural location) and $u_i \in [-2, 2]$ km for urban locations. Note that σ_u^2 changes depending on location type of DHS cluster location as a result. Given the classical measurement error form of the model, we are able to apply statistical methods in order to correct for the bias introduced by this measurement error.

2.3 Simulation Study

In the simulation study, we generate datasets for analysis, collecting information during each analysis which aid in answering the proposed questions of interest. We attempt to investigate the effect that the random displacement of DHS cluster locations has on statistical analyses involving the distance-based covariate in terms of bias and MSE of the estimator. We also

propose use of RC and attempt to determine if it is able to improve estimation in this measurement error setting.

In order to generate a single dataset for analysis, we begin by using the true (non-displaced) Uganda DHS 2011 cluster locations along with the resource locations of interest, in this case, health facility resources (points) and road networks (line segments) across Uganda. Next, we randomly displace these true DHS cluster locations based on the DHS Program geographic displacement procedure. These displaced DHS cluster locations represent the information available to all researchers. We then calculate distances from the true DHS cluster locations to the nearest resource locations, $x_i(s_d)$, and calculate similar distances using the observed (displaced) DHS cluster locations, $x_i(s_o)$.

We focus on a common statistical analysis setting used by researchers working with DHS cluster location data in the simulation study (Nhacolo et al., 2006; Byass et al., 2007; Kalemli-Ozcan and Turan, 2011; Mekonnen and Worku, 2011). The log-linear Poisson regression model has the form $Y_i | \lambda_i \stackrel{\text{ind}}{\sim} \text{poisson}(\lambda_i)$ where $\ln(\lambda_i) = \beta_0 + \beta_1 x_i(s_d)$, $i = 1, \dots, n$. λ_i represents the expected count for DHS cluster location i , Y_i represents the observed count from DHS cluster location i , $x_i(s_d)$ is the true distance covariate representing the main variable of interest in the study, and β_0, β_1 are parameters of interest describing the relationship between $x_i(s_d)$ and the expected cluster counts. This model allows researchers to examine the association between a selected cluster outcome and the distance to closest point or line segment resource. We choose regression coefficient values for the Poisson regression model considered in the study such that $\beta_0 = 2.50$ and $\beta_1 = -0.07$. These values represent similar values estimated using DHS household survey data and distance-based covariates (Feldacker et al., 2010; Lohela et al., 2012). Using these fixed β_0, β_1 parameters and true distances of interest, we are able to generate a dataset of Poisson distributed data, one data point for each DHS cluster location.

In the study, we consider five spatial density settings for the resource locations ranging from very dense to very sparse. In order to create a new spatial density setting, we randomly remove a portion of the available resource locations. Spatial density 1 begins with the full set of 2,596 health facility resources and 1,585 road segments in the Uganda country data. We then remove 25%, 50%, and 75% of the resources for spatial density settings 2, 3, and 4, respectively. Spatial density setting 5 has only 1% of the original resource locations remaining. Repeating the process of generating a single dataset for analysis, we create 100 independent datasets for each spatial density setting (five total), resulting in 500 total datasets available for analysis. We carry out these steps for generating 500 analysis datasets a total of three times, once for urban only DHS cluster locations, once for rural only DHS cluster locations, and once for a mixture of DHS cluster location types. This results in three separate simulation studies, one for each location type.

We propose two methods to analyze the generated datasets of interest and compare them with the true covariate analysis (TCA) such that

- True Covariate Analysis (TCA): Maximum likelihood estimation of the parameters using the true distances to the closest resource location, $x_i(s_d)$ (unknown to researcher),
- Method 1 (Naive): Maximum likelihood estimation of the parameters using the observed distances to the closest resource location, $x_i(s_o)$ (available to researcher), and
- Method 2 (Regression Calibration (RC)): Maximum likelihood estimation of the parameters using the estimated true distances to the closest resource location, $\hat{x}_i(s_d)$, obtained from the fitting of the classical measurement error model.

TCA represents the optimal analysis, based on the true distances of interest. Use of TCA is not possible in the current studies since access to the true DHS cluster locations that are needed to calculate the true distances is not available to researchers. Method 1 represents the naive analysis which is most commonly used in previous studies. This analysis fits the correct statistical model, similar to TCA, with the incorrect distance covariate, $x_i(s_o)$.

Method 2 fits the same statistical model as the first two analyses, while using estimates of the true distances, $\hat{x}_i(s_d)$, in place of the observed distances, $x_i(s_o)$. Recall the form of the classical measurement error model $x_i(s_o) = x_i(s_d) + u_i$, $E(u_i) = 0$, and $\text{Var}(u_i) = \sigma_u^2$. In RC (Carroll and Stefanski, 1990), our interest is in replacing $x_i(s_o)$ with $E\{x_i(s_d) | x_i(s_o)\}$ in the first stage statistical model. Assuming that $E\{x_i(s_d)\} = \mu_t$ and

$\text{Var}\{x_i(s_t)\} = \sigma_t^2$, $E\{x_i(s_t) | x_i(s_o)\} = \mu_t + \frac{\sigma_t^2}{\sigma_t^2 + \sigma_u^2} \{x_i(s_o) - \mu_t\}$. This quantity must be estimated since μ_t , σ_t^2 , and σ_u^2 are unknown, leading to

$$\hat{x}_i(s_t) = \hat{E}\{x_i(s_t) | x_i(s_o)\} = \hat{\mu}_t + \frac{\hat{\sigma}_t^2}{\hat{\sigma}_t^2 + \hat{\sigma}_u^2} \{x_i(s_o) - \hat{\mu}_t\}.$$

Assuming σ_u^2 is known, we use $\hat{\mu}_t = \bar{x}(s_o) = \frac{1}{n} \sum_{i=1}^n x_i(s_o)$ and

$\hat{\sigma}_t^2 = \frac{1}{n-1} \sum_{i=1}^n \{x_i(s_o) - \bar{x}(s_o)\}^2 - \sigma_u^2$ where both estimators are unbiased for μ_t and σ_t^2 , respectively. We use validation data to estimate σ_u^2 by displacing the observed DHS cluster locations, which themselves are already displaced, based on the DHS geographic displacement procedure m times. Each time we displace, we calculate a new set of observed distances $z_{ij}(s_o)$, treating the original set, $x_i(s_o)$, as the true distances of interest. We then calculate $w_{ij}(s_o) = \{z_{ij}(s_o) - x_i(s_o)\}$ and use these values to calculate

$\hat{\sigma}_u^2 = \frac{1}{m} \sum_{j=1}^m \left[\frac{1}{n-1} \sum_{i=1}^n \{w_{ij}(s_o) - \bar{w}_{.j}(s_o)\}^2 \right]$ where $\bar{w}_{.j}(s_o) = \frac{1}{n} \sum_{i=1}^n w_{ij}(s_o)$. $\hat{\sigma}_u^2$ is an unbiased estimator of σ_u^2 and we can improve the precision of $\hat{\sigma}_u^2$ simply by increasing m . In cases where $\hat{\sigma}_t^2 < 0$, we set $\hat{\sigma}_t^2 = 0$. Note that this process is repeated separately for urban and rural DHS cluster locations in order for μ_t , σ_t^2 , and σ_u^2 to vary by location type.

In Appendix B, we present R Statistical Software (R Core Team, 2012) code to randomly displace the observed DHS cluster locations based on the DHS Program geographic displacement procedure. Researchers can use this code to assist in calculating $\hat{\sigma}_u^2$ for urban and rural locations separately. If analyses are carried out by urban/rural status, the *rcaI* function in Stata (Hardin et al., 2003) can then be used along with $\hat{\sigma}_u^2$ to implement Method 2. However, if a mixture of location types is considered, then Stata can not be used since it does not allow for multiple measurement error models. In this case, the presented formulas must be used to calculate $\hat{x}_j(s_d) = \hat{E}\{x_j(s_d) | x_j(s_o)\}$ for each DHS cluster location with respect to location type (urban or rural). Since the main distance covariate is now estimated, there is increased uncertainty in the parameter estimates for the regression coefficients of the log-linear Poisson regression model. In order to account for this extra variability, bootstrapping is often used to calculate standard errors. Stata has the ability to provide correct standard errors in the RC setting if analyses are carried out separately for urban and rural DHS cluster locations.

We fit each of the proposed models to a single generated dataset and collect information from each of the model fits. First, we collect the estimate of β_1 , $\hat{\beta}_1^{(j)}$, for each model $j = 1$ (TCA), $j = 2$ (Method 1), and $j = 3$ (Method 2). β_1 represents the main parameter of interest in the study as it describes the association between the distance to closest resource and the expected cluster count for a selected outcome. Collection of $\hat{\beta}_1^{(j)}$ allows us to estimate the

bias of the estimator from each model, $E\{\hat{\beta}_1^{(j)}\} - \beta_1$, and the MSE, $E\left\{\left(\hat{\beta}_1^{(j)} - \beta_1\right)^2\right\}$. We prefer estimators with less bias and smaller MSEs. For each displaced DHS cluster location, we also determine if it was correctly assigned to the true closest resource location. We then average these binary responses to estimate the proportion of DHS cluster locations that are correctly assigned for each displaced dataset. This average is collected for each generated dataset with proportions closer to one being preferred.

We use standard mixed statistical models to analyze the bias and MSE outcomes which account for the fact that each model is applied to the same dataset through use of random effects. We then test whether the bias of the estimator associated with each model is significantly different from zero and investigate the change in bias across the spatial density setting of the resources. We also formally compare the MSE values from Method 1 and Method 2 to TCA at each spatial density setting. These analyses are repeated separately for the rural, urban, and mixed location studies as well as for point and line segment resource locations. Bonferroni corrections are used in order to account for the multiple testing.

2.4 Case Study: HIV Testing and Proximity to Health Centers

In this Uganda point resource case study, we attempt to determine whether the distance a woman has to travel to the closest available health facility resource is associated with the likelihood of her ever being tested for HIV. Health center resource locations from Uganda (World Health Organization, 2009) are used to define the distance-based covariate used in the study. The outcome of interest, number of women respondents 15–49 years in DHS cluster locations who had ever been tested for HIV, is obtained from the Uganda DHS 2011.

Distances from each DHS cluster location to the closest respective health center are calculated for the available clusters across Uganda based on true (undisplaced) cluster locations and displaced cluster locations. The distances are then used in the statistical analysis that relates the number of women who had ever been tested for HIV at a DHS cluster location and the closest distance to a health center. We hypothesize that DHS cluster locations further from health centers will have a lower number of women who have ever been tested for HIV on average, controlling for the population of the DHS cluster. A Poisson regression model is used to analyze the association of interest while controlling for the population of women 15–49 years at each DHS cluster using a population offset term. All three of the presented models are applied to the dataset and results are compared. Uganda DHS 2011 clusters from both urban and rural areas are included in the analysis.

3 Results and Discussion

3.1 Simulation Study Results

The results displayed in the presented figures are often not shown on the same scale for the different location types (all, rural, urban) due to large differences in the magnitudes of the estimated bias and MSE values.

3.1.1 Point Resource Locations—Figure 2 displays the bias analysis results for each of the models and location specific analyses. Open circles represent estimates which are statistically different from zero. As expected, TCA provides statistically significant unbiased estimates of the true β_1 parameter at each spatial density setting for each analysis. Method 2 results in more statistically significantly unbiased estimates across spatial density settings than Method 1 for analyses subset to all locations and to rural only locations. For urban locations alone, Method 1 and Method 2 perform very similarly due to the decrease in measurement error in the distance covariate. As the spatial density of the point resources decreases, the bias decreases for Method 1 and Method 2. This result is fairly consistent across each of the location specific analyses. Method 2 is clearly preferred over Method 1 in terms of bias reduction for all locations and rural only locations.

Figure 3 displays the MSE analysis results for each of the models and location specific analyses. An open circles represents an estimate which is statistically different than the estimate from TCA at the same spatial density setting. The MSE of the Method 1 estimator is statistically larger than that from TCA for spatial density settings 1–4 for analyses subset to all locations and to rural only locations. However, for the urban only locations, TCA and Method 1 produce estimators with similar MSE values across all spatial density settings. The MSEs from Method 2 are never statistically different from those of TCA. Once again, as the spatial density of the point resources decreases, the MSE from Method 1 and Method 2 become more similar to the MSE of the TCA estimator. Overall, Method 2 performs as well or better than Method 1 at all considered settings. This once again suggests that Method 2 outperforms Method 1 in estimating the main association of interest.

Table 1 displays the proportion of displaced DHS cluster locations that are correctly linked to their actual closest point resources (based on non-displaced locations) across each spatial density setting. Clearly, as the density of point resources decreases, the probability that a

DHS cluster is linked to the correct closest point resource increases for all location types. This makes intuitive sense when the extreme case of a single point resource is considered. In this scenario, each DHS cluster would be correctly assigned, regardless of the level of displacement. These results suggest that linking a DHS cluster with closest point resource attributes should only be done when the number of point resources in a study area is relatively small.

3.1.2 Line Segment Resource Locations—Figures 4 and 5 display the bias and MSE results, respectively, for the line segment resource locations. Table 1 also displays the proportion of correctly linked DHS cluster location and line segment resource results. The line segment resource results are very similar to the point resource results and overall suggest that Method 2 is preferred over Method 1 in most settings. For urban only DHS cluster locations at spatial density 1 however, Method 1 is preferred over Method 2. This result is due to the increased proportion of total variability in the observed distance covariate that is associated with the measurement error. In this setting, the total amount of variability is small due to the large number of line segment resources in the study area and the nature of the observed distances at urban only locations. This can be seen in Table 3 as the standard deviation of observed distances is $0.79 (\sqrt{\hat{\sigma}_t^2 + \hat{\sigma}_u^2})$, much smaller than in other settings. However, the amount of this variability that is due to the measurement error does not decrease accordingly. Instead, σ_t^2 is decreased, resulting in an increased proportion of variability that is due to the measurement error. This suggests that the observed distance covariate is measured with a relatively large amount of error and explains why RC results in highly variable estimates. As the spatial density of line segment resources decreases however, the total variability increases (as seen in Table 3) while the amount of measurement error is shown to be bounded regardless of the density of the line segment resources. This decreases the relative importance of the measurement error at each location type, resulting in improved analyses.

3.2 Proposed Guidelines

The presented simulation study results indicate that as the spatial density of point or line segment resources decreases, the amount of measurement error in the observed distance covariate becomes negligible due to the increased variability in the true distances to closest point or line segment resource. The RC technique of Method 2 accounts for the introduced measurement error by attempting to estimate the true distance covariate of interest. However, when only urban locations are considered, the naive distance model of Method 1 performs similarly to Method 2 due to the decreased level of measurement error in urban locations. Method 2 is not preferred in urban only locations for spatial density 1. The proportion of DHS clusters correctly assigned to their true closest point or line segment resource location increases as the spatial density of point or line segment resources decreases. These results are consistent for point and line segment based analyses.

Overall, we suggest that Method 2 should be used for distance-based covariate analyses, particularly as the spatial density of point or line segment resources increases. Information from the closest point or line segment resource location should only be used as a covariate in

situations with less dense point or line segment resource locations. Tables 2 and 3 display the general guidelines for the point and line segment resource locations, respectively. These results can help researchers determine the appropriate spatial density setting of an analysis and the recommended guideline. To use the presented table, researchers simply calculate the distances from the available DHS cluster locations (displaced) to the closest resource locations of interest. These distances are summarized using the sample mean, median, and standard deviation and compared with the appropriate tables based on the type of resources being considered (point or line segment).

3.3 Data Application Results

The presented case study is for purposes of illustrating the methods and guidelines previously discussed. Therefore, the results should only be used to assess the performance of each of the models in the given data setting.

3.3.1 Case Study: HIV Testing and Proximity to Health Centers—The observed distances from the displaced DHS cluster locations to the closest health center are 4.03 km (mean) and 3.24 km (median) with a sample standard deviation of 3.97 km. These values are similar to those of spatial density 1 in Table 2 (mean 4.15 km, median 3.01 km, and SD 3.97 km). Based on these results, we choose the RC method (Method 2) due to the reduction in bias and MSE seen at this spatial density setting, suggesting that it is a superior estimator when compared with the naive version (Method 1). The results in Table 4 suggest that as the distance from a DHS cluster location to the closest health facility resource increases, the expected number of women within the cluster who have ever been tested for HIV decreases, on average, as expected. Access to health facilities appears to be an important factor in the probability of HIV testing among women within the Uganda DHS 2011 clusters. Overall, all models performed similarly in terms of parameter estimates and standard errors. Note that we can not calculate bias and MSE in this setting because we do not know the true value of the main parameter of interest. This parameter was not defined as in the simulation study and can only be estimated in the case study.

This analysis of HIV testing and proximity to health centers in Uganda was carried out for purposes of illustration. It is likely that the true effect of the distance to the closest available health facility resource and the likelihood of ever being tested for HIV will be moderated by other unaccounted variables. The purpose of this case study is to demonstrate how well the presented guidelines perform in a realistic application of the DHS Program household survey GPS data. The results suggest that the proposed guidelines perform well in practice, although all models performed similarly in terms of statistical parameter estimation.

4 Conclusions

Our findings indicate that calculating distances from the randomly displaced DHS cluster locations to closest resource location results in a distance-based covariate which follows the classical measurement error model. We were able to quantify this error and provide bounds for the maximum amount of potential measurement error, which varies based on the location type of the DHS cluster (rural vs. urban). These findings were then used to create a simulation study which examined the effect of the random displacement on statistical

estimation when the measurement error is ignored. We proposed use of RC to account for the error and found that it was preferred over the naive method in almost all location and spatial density settings.

With this study, we also attempted to offer a technique less rigorous than RC in order to deal with the introduced measurement error. We examined the idea of discretizing the observed distance covariate, using a number of definitions to determine the appropriate categories, in order to determine if these categorical predictor variables could be used to improve model fit. The discretization was based on quantiles of observed distances as well as set cutpoints with a varying number of categories. We wanted to determine if using a discretized distance covariate based on the observed distances could match the same analysis done using the true distances. If successful, this would have provided an easy-to-use alternative to RC. The results however, suggested that these observed discretized variables only worked well in the urban only setting and in general, at the larger spatial density settings. Essentially, discretizing the observed distances was adequate at the same settings as when the naive continuous covariate was successful. This was an intuitive result since we know that if the naive continuous distances are similar to the true continuous distances, the discretized versions are also likely similar. If RC is not going to be used in future analyses, we suggest that the research question of interest needs to be modified. For example, instead of using the distance to closest resource in a study, the researcher could identify the number of resources within a certain distance of the displaced location. This too will result in an error-prone covariate, but we further explore the impact of similar analyses in subsequent work dealing with point-in-polygon studies under displaced DHS cluster locations.

A potential limitation of the presented analyses is failing to account for the fact that point aggregation is occurring in the DHS data which can lead to a separate error. However, the DHS program does not release the individual village GPS coordinates and as a result analyses can not be carried out at this level. It is also important to note that the point displacement occurs at the cluster level, not the individual village level. All spatial analyses are implemented at the cluster level, therefore our work focused on examining the impact that cluster displacement had on statistical inference. Future work in the Bayesian setting could allow for both errors to be integrated simultaneously through the use of well-developed prior distributions for distance covariates.

Based on the simulation study results, general guidelines were developed to aid future research with DHS household survey data and distance-based covariates. The guidelines were then applied to a case study and results were analyzed. Overall, use of distance-based covariates with displaced DHS cluster locations is problematic when the measurement error is not accounted for, resulting in biased estimators with inflated MSEs. We suggest that RC be used to correctly model the error and generally improve estimation of the association of interest. The proposed guidelines cover a number of settings and are general enough to be useful for future research in multiple data scenarios.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This research was supported in part by grants from the National Institute of Environmental Health Sciences (T32ES007018, P30ES010126) and the United States Agency for International Development (USAID) through the MEASURE DHS project (Contract No. GPO-C-00-08-00008-00).

References

- Armstrong MP, Rushton G, Zimmerman DL. Geographically masking health data to preserve confidentiality. *Statistics in Medicine*. 1999; 18(5):497–525. [PubMed: 10209808]
- Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL. Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology*. 2003; 14(4):408–412. [PubMed: 12843763]
- Burgert, CR.; Colston, J.; Roy, T.; Zachary, B. Geographic displacement procedure and georeferenced data release policy for the Demographic and Health Surveys. Calverton, Maryland, USA: ICF International; 2013. DHS Spatial Analysis Report No. 7.
- Burra T, Jerrett M, Burnett RT, Anderson M. Conceptual and practical issues in the detection of local disease clusters: a study of mortality in Hamilton, Ontario. *The Canadian Geographer/Le Géographe Canadien*. 2002; 46(2):160–171.
- Byass P, Worku A, Emmelin A, Berhane Y. DSS and DHS: longitudinal and cross-sectional viewpoints on child and adolescent mortality in Ethiopia. *Population Health Metrics*. 2007; 5(1):12. [PubMed: 18162133]
- Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. Measurement error in nonlinear models: a modern perspective. Boca Raton, FL: CRC press; 2010.
- Carroll RJ, Stefanski LA. Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*. 1990; 85(411):652–663.
- Cayo MR, Talbot TO. Positional error in automated geocoding of residential addresses. *International Journal of Health Geographics*. 2003; 2(1):10. [PubMed: 14687425]
- DeLuca P, Kanaroglou P. Effects of alternative point pattern geocoding procedures on first and second order statistical measures. *Journal of Spatial Science*. 2008; 53(1):131–141.
- Feldacker C, Emch M, Ennett S. The who and where of HIV in rural Malawi: Exploring the effects of person and place on individual HIV status. *Health & Place*. 2010; 16(5):996–1006. [PubMed: 20598623]
- Goldberg DW, Wilson JP, Knoblock CA. From text to geographic coordinates: the current state of geocoding. *URISA Journal*. 2007; 19(1):33–46.
- Hardin JW, Schmiediche H, Carroll RJ. The regression-calibration method for fitting generalized linear models with additive measurement error. *The Stata Journal*. 2003; 3(4):373–385.
- Jacquez GM, Rommel R. Local indicators of geocoding accuracy (LIGA): theory and application. *International Journal of Health Geographics*. 2009; 8(1):60. [PubMed: 19863795]
- Kalemli-Ozcan S, Turan B. HIV and fertility revisited. *Journal of Development Economics*. 2011; 96(1):61–65.
- Kamel Boulos MN, Cai Q, Padget JA, Rushton G. Using software agents to preserve individual health data confidentiality in micro-scale geographical analyses. *Journal of Biomedical Informatics*. 2006; 39(2):160–170. [PubMed: 16098819]
- Kravets N, Hadden WC. The accuracy of address coding and the effects of coding errors. *Health & Place*. 2007; 13(1):293–298. [PubMed: 16162420]
- Leitner M, Curtis A. Cartographic guidelines for geographically masking the locations of confidential point data. *Cartographic Perspectives*. 2004; (49):22–39.
- Leitner M, Curtis A. A first step towards a framework for presenting the location of confidential point data on maps - results of an empirical perceptual study. *International Journal of Geographical Information Science*. 2006; 20(7):813–822.
- Lohela TJ, Campbell OM, Gabrysch S. Distance to care, facility delivery and early neonatal mortality in malawi and zambia. *PloS One*. 2012; 7(12):e52110. [PubMed: 23300599]

- Mazumdar S, Rushton G, Smith BJ, Zimmerman DL, Donham KJ. Geocoding accuracy and the recovery of relationships between environmental exposures and health. *International Journal of Health Geographics*. 2008; 7(1):13. [PubMed: 18387189]
- Mekonnen W, Worku A. Determinants of fertility in rural Ethiopia: the case of Butajira Demographic Surveillance System (DSS). *BMC Public Health*. 2011; 11(1):782. [PubMed: 21985493]
- Nhacolo AQ, Nhalungo DA, Saco CN, Aponte JJ, Thompson R, Alonso P. Levels and trends of demographic indices in southern rural Mozambique: evidence from demographic surveillance in Manhica district. *BMC Public Health*. 2006; 6(1):291. [PubMed: 17137494]
- Olson KL, Grannis SJ, Mandl KD. Privacy protection versus cluster detection in spatial epidemiology. *American Journal of Public Health*. 2006; 96(11):2002–2008. [PubMed: 17018828]
- R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2012. ISBN 3-900051-07-0.
- Rushton G, Armstrong MP, Gittler J, Greene BR, Pavlik CE, West MM, Zimmerman DL. Geocoding in cancer research: a review. *American Journal of Preventive Medicine*. 2006; 30(2):S16–S24. [PubMed: 16458786]
- Uganda Bureau of Statistics (UBOS) and ICF International. Uganda Demographic and Health Survey 2011 [Datasets]. UGIR60FL and UGGE61FL Kampala, Uganda: UBOS and Calverton, Maryland: ICF International [Producers]. ICF International [Distributor]. 2012
- Ward MH, Nuckols JR, Giglierano J, Bonner MR, Wolter C, Airola M, Mix W, Colt JS, Hartge P. Positional accuracy of two methods of geocoding. *Epidemiology*. 2005; 16(4):542–547. [PubMed: 15951673]
- Whitsel EA, Quibrera PM, Smith RL, Catellier DJ, Liao D, Henley AC, Heiss G. Accuracy of commercial geocoding: assessment and implications. *Epidemiologic Perspectives & Innovations*. 2006; 3(1):8. [PubMed: 16857050]
- Wieland SC, Cassa CA, Mandl KD, Berger B. Revealing the spatial distribution of a disease while preserving privacy. *Proceedings of the National Academy of Sciences*. 2008; 105(46):17608–17613.
- World Health Organization (WHO). Uganda health facilities. 2009.
- Zandbergen PA. Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. *BMC Public Health*. 2007; 7(1):37. [PubMed: 17367533]
- Zandbergen PA. A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems*. 2008; 32(3):214–232.
- Zandbergen PA. Influence of street reference data on geocoding quality. *Geocarto International*. 2011; 26(1):35–47.
- Zandbergen PA, Hart TC. Geocoding accuracy considerations in determining residency restrictions for sex offenders. *Criminal Justice Policy Review*. 2009; 20(1):62–90.
- Zandbergen PA, Ignizio DA, Lenzer KE. Positional accuracy of TIGER 2000 and 2009 road networks. *Transactions in GIS*. 2011; 15(4):495–519.
- Zhan FB, Brender JD, De Lima I, Suarez L, Langlois PH. Match rate and positional accuracy of two geocoding methods for epidemiologic research. *Annals of Epidemiology*. 2006; 16(11):842–849. [PubMed: 17027286]
- Zimmerman DL, Fang X, Mazumdar S, Rushton G. Modeling the probability distribution of positional errors incurred by residential address geocoding. *International Journal of Health Geographics*. 2007; 6:1. [PubMed: 17214903]
- Zimmerman DL, Li J, Fang X. Spatial autocorrelation among automated geocoding errors and its effects on testing for disease clustering. *Statistics in Medicine*. 2010; 29(9):1025–1036. [PubMed: 20087879]
- Zinszer K, Jauvin C, Verma A, Bedard L, Allard R, Schwartzman K, de Montigny L, Charland K, Buckridge DL. Residential address errors in public health surveillance data: A description and analysis of the impact on geocoding. *Spatial and Spatio-temporal Epidemiology*. 2010; 1(2):163–168. [PubMed: 22749471]

(a) Point Resources, Rural DHS Cluster Location.

(b) Point Resources, Urban DHS Cluster Location.

(c) Line Resources, Rural DHS Cluster Location.

(d) Line Resources, Urban DHS Cluster Location.

Figure 1.

Validation of Measurement Error Model Through Simulation. Histograms display the differences between the observed distance from a DHS cluster location (first column: rural, second column: urban) to closest resource location (first row: points, second row: lines) and similarly defined true distance ($x_j(s_o) - x_j(s_d)$). A single DHS cluster location is displaced 1,000 times and the differences are calculated to create each plot. The mean of each presented distribution does not differ statistically from zero based on t-tests, suggesting that the presented classical measurement error model is valid.

(a) All Locations.

(b) Rural Locations.

(c) Urban Locations.

Figure 2.

Bias Results: Point Resource Locations. Bias results for the point resource locations simulation study by DHS cluster location type. Open circles indicate bias which is significantly different from zero while closed circles represent bias not statistically different from zero.

Figure 3.

MSE Results: Point Resource Locations. MSE results for the point resource locations simulation study by DHS cluster location type. At each spatial density setting, the MSEs of Method 1 and Method 2 are compared with the MSE of TCA. Open circles indicate statistically significant differences in MSEs.

(a) All Locations.

(b) Rural Locations.

(c) Urban Locations.

Figure 4.

Bias Results: Line Segment Resource Locations. Bias results for the line segment resource locations simulation study by DHS cluster location type. Open circles indicate bias which is significantly different from zero while closed circles represent bias not statistically different from zero.

(a) All Locations.

(b) Rural Locations.

(c) Urban Locations.

Figure 5.

MSE Results: Line Segment Resource Locations. MSE Results for the line segment resource locations simulation study by DHS cluster location type. At each spatial density setting, the MSEs of Method 1 and Method 2 are compared with the MSE of TCA. Open circles indicate statistically significant differences in MSEs.

Table 1

Correctly Assigned Proportions. Proportion of DHS cluster locations that are correctly assigned to their actual closest point or line segment resource.

Spatial Density	All Locations	Rural Locations	Urban Locations
Point Resources			
1	0.68	0.70	0.62
2	0.72	0.73	0.67
3	0.75	0.77	0.69
4	0.80	0.83	0.72
5	0.96	0.98	0.93
Line Segment Resources			
1	0.42	0.38	0.49
2	0.48	0.43	0.58
3	0.53	0.49	0.66
4	0.68	0.61	0.76
5	0.90	0.81	0.94

Table 2

Point Location Guidelines. Guidelines to determine the correct method for use based on the spatial density of point resources for each DHS cluster location type.

Observed Distances to Closest Point Resource (km)				
Spatial Density	Mean	Median	SD	Selected Method
All Locations				
1	4.15	3.01	3.97	2: RC
2	4.77	3.84	4.50	2: RC
3	5.76	4.47	5.58	2: RC
4	8.69	6.56	8.22	2: RC
5	64.27	48.96	59.72	1: Naive or 2: RC
Rural Locations				
1	5.13	4.04	4.10	2: RC
2	6.17	4.70	5.24	2: RC
3	7.83	6.30	6.32	2: RC
4	11.01	8.32	9.82	2: RC
5	47.73	39.61	31.34	1: Naive or 2: RC
Urban Locations				
1	1.58	1.15	1.42	1: Naive
2	2.24	1.35	2.25	1: Naive or 2: RC
3	2.23	1.46	2.13	1: Naive or 2: RC
4	3.43	1.65	4.34	1: Naive or 2: RC
5	26.50	11.89	32.16	1: Naive or 2: RC

Table 3

Line Segment Location Guidelines. Guidelines to determine the correct method for use based on the spatial density of line segment resources for each DHS cluster location type.

Observed Distances to Closest Line Segment Resource (km)				
Spatial Density	Mean	Median	SD	Selected Method
All Locations				
1	2.73	1.46	3.90	2: RC
2	3.36	1.87	4.56	2: RC
3	4.66	2.40	6.44	2: RC
4	8.06	5.37	8.57	2: RC
5	57.56	58.93	30.13	1: Naive or 2: RC
Rural Locations				
1	3.68	2.32	4.66	2: RC
2	4.91	3.29	5.90	2: RC
3	6.29	4.03	7.22	2: RC
4	9.39	6.44	9.84	2: RC
5	89.26	60.92	73.08	1: Naive or 2: RC
Urban Locations				
1	0.90	0.69	0.79	1:Naive
2	1.36	0.96	1.76	1: Naive or 2: RC
3	1.99	1.55	2.32	1: Naive or 2: RC
4	3.18	1.82	4.35	1: Naive or 2: RC
5	32.68	15.93	31.78	1: Naive or 2: RC

Table 4

HIV Testing and Proximity to Health Center Case Study Results. Parameter estimates and standard errors are presented. The standard error for the RC method is calculated using the bootstrapping technique.

Model	Estimate (SE)	Guidelines Choice
TCA	−0.007 (0.003)	NA
Method 1 (Naive)	−0.009 (0.003)	
Method 2 (RC)	−0.012 (0.005)	***

chosen method based on the presented guidelines.