Published in final edited form as:

Spat Demogr. 2016 July ; 4(2): 117–133. doi:10.1007/s40980-015-0015-z.

Influence of Demographic and Health Survey Point Displacements on Point-in-Polygon 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 use Demographic and Health Survey (DHS) data to evaluate the impact of random spatial displacements on analyses that involve assigning covariate values from ancillary areal and point feature data. We introduce a method to determine the maximum probability covariate (MPC), and compare this to the naive covariate (NC) selection method with respect to obtaining the true covariate of interest. The MPC selection method outperforms the NC selection method by increasing the probability that the correct covariate is chosen. Proposed guidelines also address how characteristics of ancillary areal and point features contribute to uncertainty in covariate assignment.

1 Background

The process of displacing Demographic Health Survey (DHS) cluster locations has been previously detailed in the literature (Burgert et al., 2013). In past studies utilizing DHS household survey data, the randomly displaced locations have been used by researchers in analyses which incorporate ancillary spatial polygon data (Balk et al., 2004; Simler, 2006; Pande et al., 2008; Feldacker et al., 2010). More specifically, researchers are often interested in assigning overlaid spatial polygon data to the DHS cluster locations in order to obtain spatially referenced covariates of interest. In recent work, these variables include poverty levels (Feldacker et al., 2010), socioeconomic status indicators (Feldacker et al., 2010), population density (Balk et al., 2004), climatic information (Simler, 2006), and ecological farming systems information (Balk et al., 2004). Most often, a spatial overlay of areal data is used to assign covariate values to DHS clusters for subsequent analyses. Hereon we will refer to this approach as point-in-polygon analyses. Due to the random displacement however, the resulting covariate is potentially incorrectly specified for a given DHS cluster. This occurs when the true DHS cluster location is displaced outside of the true polygon region.

Competing interests

The authors declare that they have no competing interests.

Past work examining the impact of locational inaccuracy on polygon-based analyses have focused on unintentionally and intentionally introduced location displacement. Strickland et al. (2007) examined the potential for address misclassification at the county level due to geocoding errors through use of a simulation study. Armstrong et al. (1999) showed that the general technique of intentionally randomly displacing locations to protect confidentiality leads to approximately correct subsequent spatial analyses. We further investigate when this approximation is adequate for polygon-based analyses under the specific random displacement method utilized by the DHS program. Similar to the results of Armstrong et al. (1999), we find that decreasing polygon sizes leads to increased potential for covariate misclassification and therefore, more error in subsequent analyses. Their work showed that the polygon region where the undisplaced point originated from could be estimated if certain properties of the random displacement method were known. They explored the special case when the displacement procedure allows for each point within a specified buffer distance to have an equal probability of selection (equal area displacement) and found that this estimate equates to calculating the area of each polygon region which intersects the displacement buffer. We extend this result for the DHS displacement procedure where this calculation is not valid since equal area displacement is not used. We show in Appendix A that the random angle and distance of displacement used by the DHS program must be accounted for in order to accurately estimate these probabilities. We also show that the presented probabilities can be estimated using Monte Carlo integration and provide R code to carry out the estimation for future research efforts.

In this paper, we investigate the impact that the displacement of DHS cluster locations in the DHS household survey has on point-in-polygon covariate analyses. We introduce a new method with respect to DHS displacement, maximum probability covariate (MPC) selection, that improves the process of assigning these point-in-polygon covariates by increasing the likelihood that the correct covariate is selected. The MPC selection method is compared to the naive covariate (NC) selection method, which simply chooses the polygon value associated with the displaced DHS cluster location, through the use of simulation and case studies.

We also investigate another related application of displaced DHS data in which ancillary point feature counts are calculated from within a specified neighborhood radius (see Messina et al. (2010) for an example of this application). Specifically, we refer to the number of point features within a pre-defined radius of a DHS cluster location as a neighborhood count, and consider this buffer count as the covariate of interest. Calculating this neighborhood count using the displaced DHS cluster location, however, can lead to an incorrect count due to the random displacement of corresponding neighborhood centroids. We show that the MPC selection method can also be used in this setting to determine the most probable buffer count, once again improving upon the NC selection method. In both scenarios, the problem of covariate misclassification due to random point displacement is addressed with respect to polygon-based analyses. In the first case, areal (i.e. polygon) units are fixed, and defined independently of point displacements. In the second case, however, polygons from which data are being extracted (i.e. neighborhood buffers) are variable, and dependent on the displaced points themselves. We show how both problems can be adequately addressed using the proposed MPC selection method.

Based on the simulation study results from both problems, we propose general guidelines for analyzing DHS data using overlaid spatial point-in-polygon covariate and for neighborhood counts. Two case studies are analyzed using the Uganda DHS 2011 data and the performance of the MPC selection method is investigated. Code is provided that calculates the probability that a DHS cluster location has been misclassified (moved outside of the original polygon region) and determines the MPC for both problem types.

2 Methods

2.1 Point-in-Polygon Analysis

When integrating ancillary spatial polygon data with point-referenced DHS cluster location data, the random displacement of DHS cluster locations can lead to misspecification of the polygon covariates needed in subsequent analyses. This happens when the true DHS cluster location is displaced outside of the its overlaid polygon region. The probability that a polygon covariate is incorrectly assigned depends on the size and shape of the considered polygon regions.

We refer to $x_i(s_t)$ as the polygon covariate assigned to DHS cluster i based on the true DHS location s_t . Similarly, $x_i(s_o)$ is the polygon covariate assigned to DHS cluster i based on the displaced DHS cluster location s_o . $x_i(s_t)$ is the main covariate of interest in the subsequent statistical analysis, however, it is unknown since s_t is not provided to researchers. Instead, only $x_i(s_o)$ is directly observed and available for analysis. Our introduced method aims to replace $x_i(s_t)$ with $\hat{x}_i(s_t)$ in the statistical analysis where $\hat{x}_i(s_t)$ represents an estimate of $x_i(s_t)$. This is similar to the process of regression calibration (Carroll and Stefanski, 1990). However, a model relating $x_i(s_t)$ and $x_i(s_o)$ is not directly available in this setting due to the discrete nature of the polygon covariates. As a result, a new method is required in order to estimate $x_i(s_t)$.

We introduce the maximum probability covariate (MPC) selection method which allows us to estimate $p_i(A) = P\{x_i(s_t) = A | s_o\}$, the probability that the true covariate of interest is equal to a polygon value A , given only the displaced DHS cluster location. Based on these estimated probabilities for all unique polygon values, the most probable polygon value for $x_i(s_t)$ is selected. The MPC is therefore defined as the polygon value which is most probable given the displaced DHS cluster location, $\hat{x}_i(s_t) = \max^{-1}\{p_i(A): A \text{ is any possible polygon value.}\}$.

In Appendix A, we describe how to estimate $p_i(A)$ in practice. This estimation is possible given the systematic nature of the DHS geographic displacement procedure (Burgert et al., 2013). Once the displaced DHS cluster location is observed, we know that the true DHS cluster location is within u km ($u=2$ for urban DHS clusters, 5 or 10 for rural DHS clusters) of the observed location. This result is due to the random angle and distance of the displacement. Therefore, a buffer can be placed around this displaced point (with a radius equal to u) and all possible locations where the true DHS cluster location originated from can be observed. Special considerations are required for rural locations due to potential variability in u (i.e., a random subset of 1% of rural points are displaced a maximum of 10 km rather than 5 km). These considerations are detailed in Appendix A. We note that due to

the systematic confinement of DHS cluster locations to their original administrative boundaries during the displacement, certain areas of the defined buffer may not be possible as an origin point for the true DHS cluster location.

Once the appropriate buffer is created and the probable region for the true DHS cluster location is defined based on administrative boundaries, we then observe that there are only a finite number of values that the true polygon covariate can assume due to the discrete nature of the areal dataset. The number of unique polygon values will depend on the size of the overlaid polygon regions in relation to the displacement buffer. For example, for very large areal units, the defined buffer may be entirely contained within a single polygon feature. In this case, we know that there is no possibility of covariate misclassification and therefore, the NC will match the true covariate. Mathematically, the MPC selection method results in $p_i(A) = 1$ in this setting where A is the true polygon value. In other words, we are 100% sure of the true covariate value. Therefore both the NC and MPC will match the true covariate. In the case where the areal units are very small with respect to the displacement buffer, many unique polygon values could be contained within the defined buffer. In such scenarios, covariate misclassification rates would generally be higher and the MPC selection method would outperform the NC selection method by reducing this error and choosing the correct covariate more often on average.

Figure 1 displays three examples of the covariate selection process for different polygon maps. In Figure 1A, the polygon regions are large with respect to the probable region within the buffer. Therefore, there is very little chance of covariate misclassification for either method. In Figure 1B, the polygon regions are smaller leading to an increased number of possible unique polygon values. Choosing the NC leads to misclassification in this example. The MPC selection process chooses the most probable value based on the size of each of the regions. Figure 1C has the smallest polygon regions and most unique polygon values. This leads to an even higher misclassification rate and once again choosing the NC leads to misclassification. Note that only the number of unique values are important in covariate selection. If all polygon features have the same value, misclassification is not possible, regardless of the number of regions.

In Appendix A, we show that the probabilities of interest can be estimated using Monte Carlo integration. We provide code to carry out the estimation in Appendix B. For each unique polygon value within the probable buffer region, we estimate the probability that the displaced point originated from a polygon region with that unique value ($p_i(A)$). These probabilities are estimated by simulating randomly distributed locations within the buffer and determining the proportion of these points that are associated with polygon value A . In order to generate a point within the buffer, we choose a random angle of displacement and random displacement distance and displace the observed DHS cluster location. This process is repeated many times which results in a large number of buffer points and improves the precision of the Monte Carlo integration. Simply calculating the proportion of areas of each unique polygon value within the buffer region leads to incorrect probability estimates due to the method of displacement used by the DHS Program, as detailed in Appendix A. The code in Appendix B implements the MPC selection method while taking into account the restricting administrative boundaries, and ultimately provides a list of possible polygon

values along with the corresponding estimated probabilities for each DHS cluster location. The probability that the displaced DHS cluster location is misclassified is also provided and is defined as $1 - p_j(A)$ where the displaced point is located in the polygon region with value A .

2.2 Neighborhood Count Analysis

In addition to the point-in-polygon analysis, we also consider the situation where the number or density of point features within a pre-defined distance of the true DHS cluster location is of interest as a covariate in a statistical analysis (Messina et al., 2010). The MPC selection method is shown to provide a solution in this setting as well. In Figure 2 we present this covariate selection process for an imaginary DHS cluster location. Figure 2A represents the correct assignment based on the true DHS cluster location. Figure 2B shows the potential for misclassification when using the NC selection method which counts the number of resources within the pre-defined distance of the displaced DHS cluster location. Figure 2C shows the MPC selection process using only three randomly created buffer points for display purposes. For each simulated DHS cluster location, the corresponding point feature count is collected and ultimately a distribution of possible neighborhood counts for the true but unknown DHS cluster location is created. In practice the number of simulated points is greatly increased. Without knowing the true DHS cluster location, the true covariate of interest in this example can only take the values 0, 1, or 2 due to the small number of point features in the area. This number of unique values will increase as the pre-defined distance gets larger or similarly as the density of point features increases. The MPC selection method estimates $P\{x_j(s_d) = z | s_o\}$ for all possible unique values z and chooses the number corresponding to the largest probability.

For this approach, we once again generate randomly distributed locations (random angle and direction of displacement) within the displacement buffer based on the location of the displaced DHS cluster location. For each generated point, we calculate the number of resources within a pre-defined distance. The MPC is chosen as the number of resources that was most often observed among all of the generated points. We present details in Appendix A that shows this results in a Monte Carlo integration of the probabilities of interest. Code is presented in Appendix C that can carry out this estimation for the neighborhood counts problem. Using techniques similar to the point-in-polygon analysis, we can calculate the covariate misclassification rate for a displaced DHS cluster location along with the MPC.

3 Simulation Studies

3.1 Point-in-Polygon

In the point-in-polygon simulation study, we investigate the covariate misclassification error in different settings and show how the MPC selection method reduces this error in each setting. Three different covariate selection models are discussed in the presented simulation study:

- True Covariate (TC): Choosing the polygon value associated with the true DHS cluster location s_p , $x_j(s_p)$.

- Method 1 (Naive Covariate (NC)): Choosing the polygon value associated with the displaced DHS cluster location $s_o, x_i(s_o)$.
- Method 2 (Maximum Probability Covariate (MPC)): The newly introduced method of choosing the polygon value associated with the highest probability of being the true covariate of interest, $\hat{x}_i(s_o)$.

TC selection represents the ideal covariate selection process of interest, where the true DHS cluster location is assigned to the appropriate polygon region and the corresponding polygon value is determined. This process however, is not possible since the true DHS cluster locations are not provided to researchers. Method 1 (NC selection) represents the most commonly implemented technique in practice whereby the displaced DHS cluster locations are assigned to the overlaid areal units and the corresponding polygon values are selected to define covariates. Method 1 can be problematic in cases where the misclassification rate for the displaced DHS cluster location is very high (small areal units). Method 2 (MPC selection) represents an improvement over Method 1 by ensuring that the most probable polygon value is selected for each DHS cluster. Method 2 maximizes the probability that the correct covariate is selected for each cluster (Appendix A). This leads to a set of covariates which more closely resemble the TC set on average.

We analyze the improvements that are seen when Method 2 is used instead of Method 1 and describe how both methods compare to TC selection. We investigate this while using three different areal datasets, each associated with a different level of misclassification error due to the sizes of the polygon regions (recall Figure 1). These datasets correspond to the first, third, and sixth level Uganda administrative units. Maps of the administrative boundaries are displayed in Figure 3. It is clear that the first level administrative boundary has the largest polygon regions while the sixth level administrative boundary has the smallest. Thus, we expect misclassification rates to be higher when using data associated with sixth-level compared to first-level administrative units, assuming unique values for all areal units.

We begin the simulation study by selecting one of the spatial polygon maps and randomly generating 5,000 locations within the extent of the administrative unit shapefile. We displace these locations based on the DHS geographic displacement procedure. This data generation and displacement process is repeated for each spatial polygon map (3 maps) resulting in a total simulation sample size of 15,000. The pre-displaced locations are treated as the true and unknown DHS cluster locations while the displaced locations represent those available to all researchers. From the true locations, we determine the covariate using the TC selection method for each DHS cluster. From the displaced locations, we similarly calculate the NC using Method 1. We then calculate the MPC using Method 2 by estimating the probabilities of each polygon value for each DHS cluster location. For each randomly generated location, we collect these three covariates of interest. Based on the collected data, we then determine whether Method 1 matches TC, whether Method 2 matches TC, whether Method 1 matches Method 2, and the misclassification rates (output from Method 2) associated with Method 1 and Method 2 for each DHS cluster.

Analyzing and summarizing these collected data, we are then able to estimate the probability that Method 1 matches TC, the probability that the Method 2 matches TC, and observe the

distribution of misclassification rates for both methods. Also, when there is disagreement between Method 1 and Method 2 with respect to the TC, we are able to estimate the probability that Method 2 is correct (and Method 1 is incorrect). We test to see if this probability is significantly larger than 0.50, indicating that Method 2 is outperforming Method 1, for each areal dataset considered.

3.2 Neighborhood Counts

A similar approach can be adapted to address potential bias associated with calculating ancillary point densities surrounding displaced DHS cluster locations. We once again attempt to quantify the improvement in Method 2 over Method 1 in terms of matching the TC. In this case, the TC represents the number of point features within r km of the true DHS cluster location, where r is the pre-defined distance of interest. Method 1 (NC) represents the number of resources within r km of the displaced DHS cluster location, and Method 2 (MPC) selects the most probable number of resources located within r km of the true DHS cluster location based only on the location of the displaced DHS cluster. The simulation study setup and analysis remains the same as in the point-in-polygon simulation study with a total sample size of 15,000. We consider different predefined distances of interest ($r=1, 2, 5, 10, 15$, and 30 km) around the true DHS cluster location. These different distances lead to different levels of misclassification rates in the same way that the different administrative boundary datasets lead to different levels of misclassification. Therefore, we generate 2,500 random locations for each buffer distance (15,000 in total) and displace them based on the DHS geographic displacement procedure.

For Method 1, we calculate the number of resources within the selected distance surrounding the displaced DHS cluster location. For Method 2, we estimate the probabilities associated with obtaining the observed number of point features located within a specified distance of the true DHS cluster location, and choose the most probable value. For TC selection, we calculate the neighborhood counts based on the true DHS cluster locations. The same simulation data outcomes are collected and analyzed as in the point-in-polygon study.

4 Case Studies

4.1 Data Description

In the presented case studies, 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 with seven points being gazetted to the nearest village. A total of 397 DHS clusters were available for analysis after four points were removed due to missingness and three points were removed for being displaced outside of the original administrative boundaries. We also use the Uganda administrative boundary shapefiles for the first, third, and sixth level administrative boundaries. Figure 3 displays the different administrative boundary maps. In the neighborhood count analysis, health center point resource locations from Uganda (World Health Organization, 2009) are used.

4.2 Point-In-Polygon: Neighborhood Determinants of HIV Knowledge

The goal of this case study is to determine whether comprehensive knowledge of HIV among women DHS respondents in rural Uganda is associated with the economic characteristics of neighborhoods (i.e., area-level poverty measures). We obtain area-level poverty data from the Uganda Bureau of Statistics (Emwanu et al., 2007) (available only for rural Uganda third level administrative units) that we link to the point-referenced DHS cluster locations using the different point-in-polygon covariate assignment methods. This results in 208 rural clusters available for analysis. We obtain the outcome of interest, the number of respondents interviewed in rural DHS clusters that had comprehensive knowledge of HIV, from the Uganda 2011 DHS. This indicator variable is constructed according to correct responses to all of the following DHS survey true/false questions (correct answer given in parentheses):

1. To reduce the risk of getting HIV one should always use condoms (True),
2. To reduce the risk of getting HIV one should have only one sex partner (True),
3. One can get HIV from mosquito bites (False),
4. One can get HIV by witchcraft or supernatural means (False), and
5. A healthy looking person can have HIV (True).

Area-level poverty is defined as the percentage of the population below the poverty line within an administrative unit. A separate analysis is also conducted using a discretized poverty measure. In that case, area-level poverty is redefined as an ordinal variable corresponding to low (< 10% below poverty line), moderate (10–50% below poverty line), and high poverty levels (> 50% below poverty line). For each covariate type (continuous and discrete), we determine the Method 1 and Method 2 covariates for each DHS cluster. We also use the TC selection method since we have access to the true DHS cluster locations.

Using these covariates, we analyze the association of poverty and HIV knowledge using a log-linear Poisson regression model of the form $Y_i | \lambda_i \overset{ind}{\sim} \text{Poisson}(\lambda_i)$ where

$\ln(\lambda_i) = \beta_0 + \beta_1 x_i^{(M)} + \log(\text{population}_i)$, $i = 1, \dots, n$. λ_i represents the expected number of women respondents with comprehensive HIV knowledge for DHS cluster i , Y_i represents the observed number of women with comprehensive HIV knowledge from DHS cluster i , $\log(\text{population}_i)$ is the population offset term for cluster i , $x_i^{(M)}$ is the poverty covariate for a selected method representing the main variable of interest in the study, and β_0, β_1 are

parameters of interest describing the relationship between $x_i^{(M)}$ and the expected cluster counts. Note that in the discrete case study, the form of the model changes due to the introduction of discrete levels of poverty such that

$\ln(\lambda_i) = \beta_0 + \beta_0 I(x_i^{(M)} = \text{Moderate}) + \beta_2 I(x_i^{(M)} = \text{High}) + \log(\text{population}_i)$, where $I(\cdot)$ is the indicator function taking value one if the input statement is true and zero otherwise.

Results are compared when $x_i^{(M)} = x_i(s_t)$ (TC), $x_i^{(M)} = x_i(s_o)$ (NC), and

$x_i^{(M)} = \hat{x}_i(s_t)$ (MPC). We hypothesize that rural DHS clusters in high poverty areas will

have fewer people who have comprehensive HIV knowledge on average, after controlling for the size of the population.

4.3 Neighborhood Counts: HIV Testing and Density of Health Centers

In this Uganda based buffer count case study, we attempt to determine whether a woman's access to nearby health resource centers 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 covariate used in the study. The outcome of interest, number of women respondents in DHS clusters who had ever been tested for HIV, is obtained from the Uganda 2011 DHS.

The number of health resources within 10 and 30 km of each DHS cluster location is calculated based on the true DHS cluster locations and displaced DHS cluster locations. We also obtain the MPC using Method 2. These covariates are then used in the statistical analysis that relates the number of women who had ever been tested for HIV at a DHS cluster and the number health centers surrounding the residence, with separate analyses for each pre-defined distance. The same form of the log-linear Poisson regression model presented for the point-in-polygon analysis is also used to analyze the association of interest. We hypothesize that DHS clusters with fewer health centers located nearby will have a fewer number of women who have ever been tested for HIV on average, controlling for the population of the DHS cluster. Once again we compare the results using the covariate constructed from each method, $x_j(s_\rho)$, $x_j(s_o)$, and $\hat{x}_i(s_\rho)$.

5 Results and Discussion

5.1 Simulation Study Results: Point-in-Polygon

The simulation study results for the point-in-polygon covariate assignment problem are displayed in Table 1 and Table 2. Table 1 summarizes the performance of each method with respect to the TC while Table 2 displays the distribution of the difference in misclassification rates between Method 1 and Method 2 (Method 1-Method2) when there is disagreement between the two methods.

For the first level administrative boundary dataset, which has the largest polygon regions as seen in Figure 3, the resulting misclassification rates for both Method 1 and Method 2 are zero. There is no chance that a DHS cluster is misclassified in this setting because DHS clusters are restricted to the first level administrative boundary during the displacement. Therefore, it is not possible for a cluster to be displaced outside of the true polygon region and as a result the correct covariate is always selected using either Method 1 or Method 2. Methods 1 and 2 perform identically in this trivial situation and exactly match the TC for each DHS cluster.

For the third level administrative boundary map, it becomes possible for the true DHS cluster location to be displaced outside of the original areal unit, resulting in an incorrectly assigned covariate. Table 1 shows that Method 1 and Method 2 produce the same covariate 98.13% of the time. Both Method 1 and Method 2 match the true covariate over 88% of the time. When the covariates from Method 1 and Method 2 disagree with respect to the TC

however, 66.29% of the time Method 2 correctly matches the TC. This proportion is statistically different from 0.50 at the $\alpha = 0.05$ level, and this suggests that Method 2 represents an improvement over Method 1 by increasing the likelihood that the correct polygon value is assigned to each DHS cluster. Table 2 shows that when Method 1 and Method 2 disagree, the difference in misclassification rates is substantial in favor of Method 2. The misclassification rate for Method 1 is 0.1959 larger on average with one observed difference greater than 0.92. These results suggest that when Method 1 and Method 2 disagree, Method 2 is favored due to reduced misclassification rates and better performance with respect to the TC.

For the sixth level administrative boundary, the performance of Method 1 and Method 2 with respect to the TC decreases with both matching the TC only 73.25% and 74.29% of the time respectively. Method 1 and Method 2 now match 96.52% of the time. Method 2 once again represents an improvement over Method 1 by increasing the likelihood that the correct polygon value is assigned to each DHS cluster. 64.94% of the time Method 2 is correct when Method 1 and Method 2 disagree with respect to the TC. This proportion is statistically different from 0.50 at the $\alpha = 0.05$ level. The difference in misclassification rates is similar to the third-level administrative unit data results and once again suggests that when there is disagreement in Method 1 and Method 2, Method 2 is favored.

5.2 Simulation Study Results: Neighborhood Counts

The simulation study results for the neighborhood count problem are displayed in Tables 3 and 4. It is clear that as the pre-defined distances increase, the proportion of times that Method 1 and Method 2 match the TC decreases. Similar to the point-in-polygon simulation study results, Method 2 matches the TC more often than Method 1 when the methods disagree with respect to the TC (Table 3). All presented proportions are statistically different from 0.50 at the $\alpha = 0.05$ level. Also, the difference in misclassification rates decreases as the distances increase, though they are strongly in favor of Method 2. Overall, the results are very similar to the point-in-polygon study and suggest that for larger distances the amount of misclassification error is increased. Method 2 should be used over Method 1 in order to ensure the highest probability of obtaining the TC.

5.3 Case Study Results: Neighborhood Determinants of HIV Knowledge

The covariate selected by Method 2 matches the TC of interest 175 out of the 208 available DHS clusters compared with only 173 for the Method 1 covariate, suggesting that Method 2 is the preferred predictor in this situation. The continuous poverty log-linear Poisson regression results in Table 5 suggest that as poverty increases, the expected number of women rural DHS cluster members with comprehensive HIV knowledge decreases on average, as expected. The results from each method are significant at the 0.05 level of significance. In order to reduce the misclassification error introduced by the use of the areal poverty data, we also consider discretization of the poverty covariate. After discretizing the poverty variable, the TC matches Method 2 199 out of 208 times and 198 times for Method 1. This suggests that both methods perform well with respect to the truth while Method 2 is still preferred. Discretizing the continuous values associated with the polygon regions reduces the misclassification rate by limiting the number of unique values and leads to an

improved predictor which more closely matches the truth. The discrete covariate results in Table 6 suggest that at the highest poverty category, the expected number of women rural DHS cluster members with substantial HIV knowledge is reduced when compared with the lowest category, though the results are not statistically significant at the 0.05 level. These results are similar to the continuous covariate results overall.

Note that this analysis of neighborhood determinants of HIV knowledge in Uganda is meant solely for purposes of illustration of the methods. It is likely that the true effects of neighborhood-level poverty rates on comprehensive knowledge of HIV by rural DHS respondents is moderated by other unaccounted for variables.

5.4 Case Study Results: HIV Testing and Proximity to Health Centers

Two distance buffers were evaluated with this case study: 10 and 30 km. For the 10 km distance, the covariate selected by Method 2 matches the true covariate of interest 196 out of the 397 available DHS clusters compared with only 193 for the Method 1 covariate, suggesting that Method 2 is the preferred predictor in this situation. For the 30 km distance, this drops to 160 and 158 for Method 2 and Method 1 respectively. The results in Tables 7 and 8 suggest that as the number of health resources within 10 and 30 km of the true DHS cluster increases, the expected number of women within the cluster who have ever been tested for HIV increases, 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 clusters.

As with the previous case study, this analysis of HIV testing and proximity to health centers in Uganda is meant solely for purposes of illustration of the methods. It is likely that the true effect of the density of closest available health resource centers and the likelihood of ever being tested for HIV is moderated by other unaccounted variables.

6 Conclusions and Proposed Guidelines

Our findings indicate that assigning spatial polygon values to displaced DHS cluster locations can sometimes lead to misrepresentation of respective covariate values. The probability that a covariate is misclassified is directly related to the size of the areal units and the maximum amount of potential displacement for a particular DHS cluster location. Similarly, the pre-defined buffer distance is related to misclassification in the neighborhood counts problem. We introduced a method to limit this misclassification error and showed that it could be used to obtain the TC of interest more often than the NC selection method. The results suggest that Method 1 (NC) and Method 2 (MPC) often match but when there is disagreement between the two, Method 2 is always preferred. Under certain scenarios, such as the spatial linkage of data from large areal units, covariate values obtained using either method will be similar or match exactly; however, under other circumstances, the two methods will yield different covariate values. In the latter case, we showed that the MPC approach outperformed the NC approach.

A simpler alternative to Method 2 could involve discretizing continuous measures associated with areal units. Discretizing the values associated with polygon features will reduce the

number of unique polygon values in a dataset as illustrated in the poverty and HIV knowledge case study. With regard to the neighborhood counts problem, smaller buffer distances are recommended since they lead to fewer unique count values, and therefore reduced covariate misclassification error. As the size of areal units become extremely small and the number of unique polygon values increases, the dataset begins to resemble a continuous raster dataset and the introduced MPC selection method will not be useful. In this setting, the probability of obtaining the true covariate will be very small and new techniques are required to improve covariate estimation. We explore the impact that DHS displacement has on raster based analyses in subsequent work.

We present two R functions (Appendix B and Appendix C) which can help researchers to make informed decisions regarding the type of analyses that are appropriate. We generally recommend that Method 2 be used in order to ensure that the TC is chosen with the highest probability for each DHS cluster. Method 2 always selects the maximum probability covariate value, unlike Method 1 which can sometimes choose a covariate with extreme misclassification error (column 6 of Table 2 and Table 4). We recommend the following steps for investigators interested in linking ancillary areal or point density data to DHS data: (1) Determine the distribution of covariate misclassification rates for their given application (i.e., point-in-polygon or neighborhood count variables; Appendix B), and (2) Calculate the most probable covariate for each DHS cluster for use in subsequent analysis (Appendix C). While Method 2 does improve upon the Method 1, it can still be associated with high levels of covariate misclassification for small polygon regions and larger buffer distances. In these cases, we recommend that the polygon region be discretized, leading to less unique polygon values and less covariate misclassification error, or the buffer distance be made smaller. It is also possible to identify certain DHS cluster locations which have particularly high covariate misclassification rates and remove these from the analyses if appropriate.

Point aggregation used by the DHS program can potentially lead to increased error in assigning polygon covariates to a cluster location. However, we expect this to be less of an issue in the polygon covariate setting, especially for larger polygon regions. We also note that GPS coordinates for individual villages are not available from the DHS program and as a result all spatial analyses occur at the cluster level. Therefore, we expect our results to be robust to this error source but do acknowledge this as a possible limitation. Future work could incorporate both error sources and determine their combined impact on subsequent spatial analyses.

By definition, Method 2 will perform at least as well as Method 1 regardless of the data setting (areal data overlay, density of point features, urban/rural DHS cluster status) because it selects the covariate that is most probable. An urban DHS cluster location will typically have less misclassification error than a rural DHS cluster location found in the same area due to the smaller maximum amount of potential displacement distance used for urban clusters (2 km vs. 5-10 km). This smaller distance results in less unique values within the probable buffer region and therefore limits the potential for incorrect covariate assignment. However, administrative boundary areas in urban locations are often smaller than in rural locations which can limit the misclassification error by decreasing the size of the probable buffer. While the misclassification rates will often differ between urban and rural DHS cluster

locations, the guideline remains the same for each type; misclassification rates for each location should be analyzed to determine the appropriate type of analysis and Method 2 should generally be used over Method 1. In Appendix A, we show how the misclassification rates and MPC are calculated separately for urban and rural DHS cluster locations due to the different maximum potential displacement distances. In this way, the location type is automatically considered in the MPC selection process.

The application of ancillary spatial data can enrich analyses that involve DHS data; however, the impact of random displacement of DHS cluster locations on such analyses warrants careful consideration. The approaches evaluated and recommended in this work are meant to optimize the usage of spatial DHS data for the point-in-polygon and neighborhood counts settings in light of this displacement protocol.

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]
- Balk D, Pullum T, Storeygard A, Greenwell F, Neuman M. A spatial analysis of childhood mortality in West Africa. *Population, Space and Place*. 2004; 10(3):175–216.
- Burgert, CR.; Colston, J.; Roy, T.; Zachary, B. Geographic displacement procedure and georeferenced data release policy for the Demographic and Health Surveys. ICF International; Calverton, Maryland, USA: 2013. DHS Spatial Analysis Report No. 7.
- Carroll RJ, Stefanski LA. Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*. 1990; 85(411):652–663.
- Emwanu, T.; Okwi, PO.; Hoogeveen, JG.; Kristjanson, P.; Henninger, N. Technical report. Uganda Bureau of Statistics and the International Livestock Research Institute (ILRI); 2007. Nature, distribution and evolution of poverty and inequality in Uganda..
- 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]
- Messina JP, Emch M, Muwonga J, Mwandagaliwa K, Edidi SB, Mama N, Okenge A, Meshnick SR. Spatial and socio-behavioral patterns of HIV prevalence in the Democratic Republic of Congo. *Social Science & Medicine*. 2010; 71(8):1428–1435. [PubMed: 20739108]
- Pande S, Keyzer MA, Arouna A, Sonneveld BG. Addressing diarrhea prevalence in the West African Middle Belt: social and geographic dimensions in a case study for Benin. *International Journal of Health Geographics*. 2008; 7(1):17. [PubMed: 18433488]
- Simler, KR. Nutrition mapping in Tanzania: an exploratory analysis. International food policy research institute (IFPRI). Food consumption and nutrition division (FCND); Washington DC, USA: 2006.
- Strickland MJ, Siffel C, Gardner BR, Berzen AK, Correa A. Quantifying geocode location error using GIS methods. *Environmental Health*. 2007; 6(10)

Uganda Bureau of Statistics (UBOS) and ICF International. Uganda Demographic and Health Survey 2011 [Datasets]. UBOS and Calverton; ICF International [Producers]. ICF International [Distributor]; Uganda: Maryland: 2012. UGIR60FL and UGGE61FL Kampala

World Health Organization (WHO). Uganda health facilities. 2009.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1. Point-in-Polygon Covariate Misclassification Rate Example. Figure 1A shows the true, s_t (*), and displaced, $s_o(\cdot)$, DHS cluster locations along with the probable displacement buffer and overlaid spatial polygon map. Similar plots are displayed in Figure 1B and Figure 1C with spatial polygon maps with smaller polygon regions. Plots not shown on the same scale.

Figure 2.

Buffer Count Covariate Creation Example. Each figure shows the true (*), s_b and displaced, s_o (.), DHS cluster locations along with the resource locations (triangles), probable displacement region (solid circle), and buffer count region (dashed circle). The first figure shows the true covariate calculation where the number of resources surrounding the true DHS cluster location are observed. The second figure shows the naive covariate calculation where the number of resources surrounding the displaced DHS cluster location are observed. The third figure shows an example of the maximum probability covariate calculation where random points are generated in the probable displacement region and a distribution of buffer counts is created.

Figure 3. Administrative Boundaries. First, third, and sixth level administrative boundaries for Uganda, leading to reduced, moderate, and high covariate misclassification rates.

Table 1

Point-in-Polygon Simulation Study Results.

Administrative Boundary	M2=Truth when $I(M1 = TC)$	$I(M2 = TC)$	M1=M2	M1=TC	M2=TC
1	NA		1.0000	1.0000	1.0000
3	0.6629 *		0.9813	0.8878	0.8939
6	0.6494 *		0.9652	0.7325	0.7429

TC: True covariate, M1: Naive covariate, M2: Maximum probability covariate, $I(.)$ is the indicator function equal to one if the statement is true and zero otherwise.

Proportions are presented in the table.

* indicates the estimate is statistically different from 0.50 at the 0.05 level of significance.

Table 2

Distribution of the difference in misclassification rates for Method 1: Naive covariate and Method 2: Maximum probability covariate when the methods disagree for the point-in-polygon problem (Method 1 - Method 2).

Administrative Boundary	Mean	SD	Min	Median	Max
1	NA	NA	NA	NA	NA
3	0.1959	0.2129	0.0003	0.1335	0.9274
6	0.1237	0.1476	0.0004	0.0798	0.8808

Note that by definition the difference in rates will be 0 since Method 2 maximizes the probability of success.

Table 3

Buffer Counts Simulation Study Results.

Buffer Size (km)	M2=Truth when I (M1 = Truth)	I (M2 = Truth)	M1=M2	M1=Truth	M2=Truth
1	0.7885 [*]		0.9788	0.9662	0.9784
2	0.7100 [*]		0.9593	0.9138	0.9309
5	0.5413 [*]		0.9554	0.8148	0.8185
10	0.6266 [*]		0.9354	0.7070	0.7234
15	0.5245 [*]		0.9167	0.6252	0.6292
30	0.6201 [*]		0.8859	0.4818	0.5092

TC: True covariate, M1: Naive covariate, M2: Maximum probability covariate, $I(\cdot)$ is the indicator function equal to one if the statement is true and zero otherwise.

Proportions are presented in the table.

^{*} indicates the estimate is statistically different from 0.50 at the 0.05 level of significance.

Table 4

Distribution of the difference in misclassification rates for Method 1: Naive covariate and Method 2: Maximum probability covariate when the methods disagree for the buffer counts problem (Method 1 - Method 2).

Buffer Size	Mean	SD	Min	Median	Max
1	0.4188	0.2124	0.0548	0.5026	0.7023
2	0.2021	0.1192	0.0003	0.1896	0.4477
5	0.1545	0.1395	0.0018	0.1093	0.6965
10	0.1170	0.1038	0.0001	0.0867	0.5956
15	0.1166	0.1148	0.0001	0.0828	0.6567
30	0.0969	0.0949	0.0001	0.0694	0.6933

Note that by definition the difference in rates will be 0 since Method 2 maximizes the probability of success.

Table 5

Point-in-Polygon Case Study Results: Continuous Poverty and HIV Knowledge Results.

Covariate	$\hat{\beta}$	$SE(\hat{\beta})$	# Correct
TC	-0.00235 *	0.00056	208
Method 1: NC	-0.00232 *	0.00057	173
Method 2: MPC	-0.00227 *	0.00059	175

SE: Standard error.

* indicates the estimate is statistically different from 0 at the 0.05 level of significance

Table 6

Point-in-Polygon Case Study Results: Discrete Poverty Results.

Covariate	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	# Correct
TC	-1.036 [*] (0.077)	0.111 (0.086)	-0.110 (0.085)	208
Method 1: NC	-1.011 [*] (0.076)	0.095 (0.085)	-0.154 (0.084)	198
Method 2: MPC	-1.011 [*] (0.076)	0.089 (0.085)	-0.152 (0.085)	199

Standard errors shown in parentheses.

^{*} indicates the estimate is statistically different from 0 at the 0.05 level of significance.

Table 7

Buffer Counts Case Study Results: 10 km Distance Buffer Count of Health Resources and HIV Testing.

Covariate	$\hat{\beta}$	$SE(\hat{\beta})$	# Correct
TC	0.000221 *	0.000074	397
Method 1: NC	0.000222 *	0.000075	193
Method 2: MPC	0.000224 *	0.000075	196

SE: Standard error.

* indicates the estimate is statistically different from 0 at the 0.05 level of significance

Table 8

Buffer Counts Case Study Results: 30 km Distance Buffer Count of Health Resources and HIV Testing.

Covariate	$\hat{\beta}$	$SE(\hat{\beta})$	# Correct
TC	0.000170 *	0.000058	397
Method 1: NC	0.000174 *	0.000058	158
Method 2: MPC	0.000173 *	0.000058	160

SE: Standard error.

* indicates the estimate is statistically different from 0 at the 0.05 level of significance