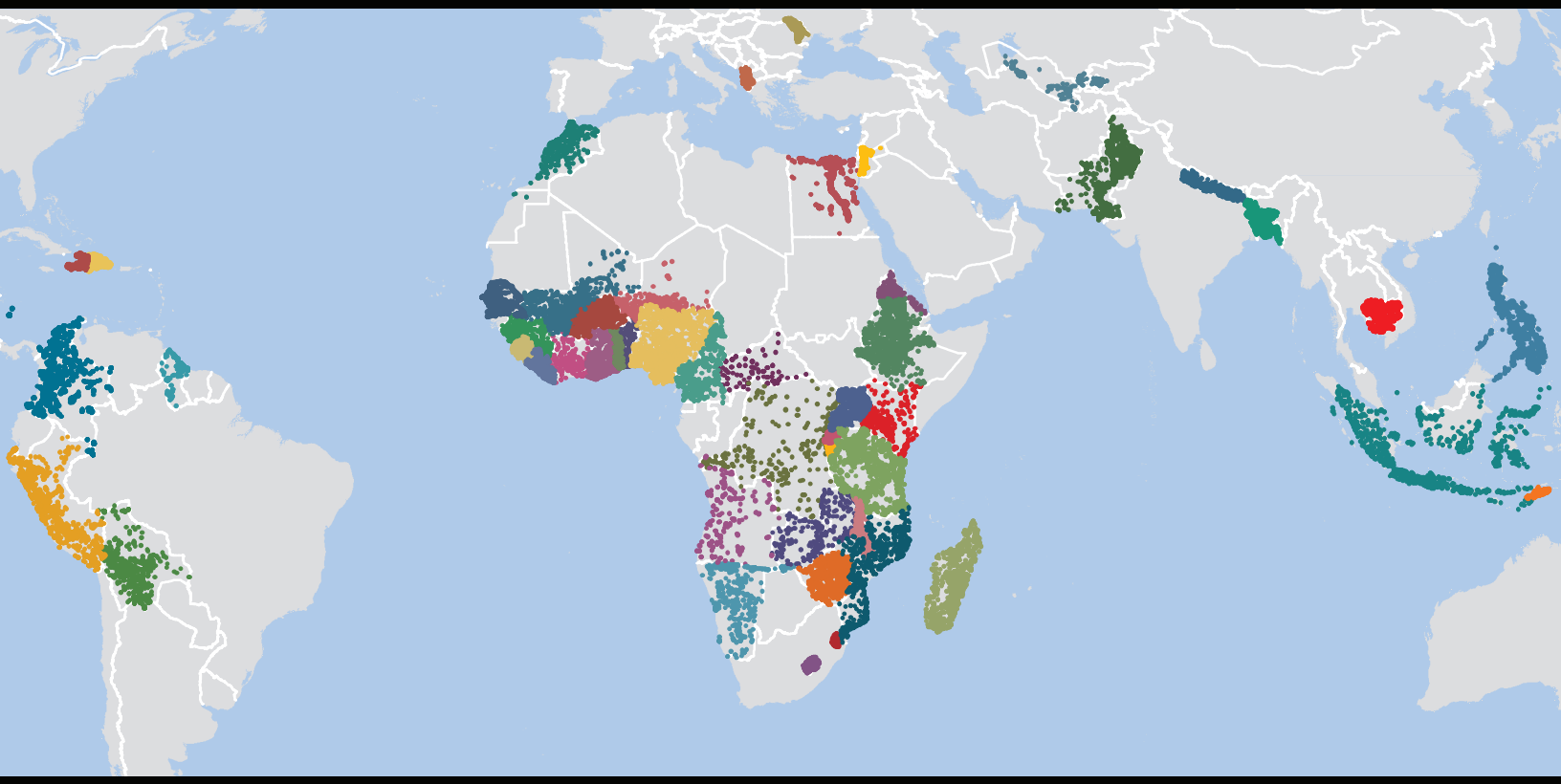
USAID
FROM THE AMERICAN PEOPLE

GEOGRAPHIC DISPLACEMENT PROCEDURE AND GEOREFERENCED DATA RELEASE POLICY FOR THE DEMOGRAPHIC AND HEALTH SURVEYS

DHS SPATIAL ANALYSIS REPORTS 7



SEPTEMBER 2013

This publication was produced for review by the United States Agency for International Development (USAID). The report was prepared by Clara R. Burgert, Josh Colston, Thea Roy, and Blake Zachary of ICF International, Calverton, Maryland, USA.

MEASURE DHS assists countries worldwide in the collection and use of data to monitor and evaluate population, health, and nutrition programs. Additional information about the MEASURE DHS project can be obtained by contacting MEASURE DHS, ICF International, 11785 Beltsville Drive, Suite 300, Calverton, MD 20705 (telephone: 301-572-0200; fax: 301-572-0999; e-mail: reports@measuredhs.com; internet: www.measuredhs.com).

The main objectives of the MEASURE DHS project are:

- to provide decision makers in survey countries with information useful for informed policy choices;
- to expand the international population and health database;
- to advance survey methodology; and
- to develop in participating countries the skills and resources necessary to conduct high-quality demographic and health surveys.

DHS Spatial Analysis Report No. 7

Geographic Displacement Procedure and Georeferenced Data Release Policy for the Demographic and Health Surveys

Clara R. Burgert

Josh Colston

Thea Roy

Blake Zachary

ICF International
Calverton, Maryland, USA

September 2013

Corresponding author: Clara R. Burgert, International Health and Development, ICF International, 11785 Beltsville Drive, Calverton, Maryland 20705, USA; Phone +1 301-572-0446; Fax +1 301-572-099; Email: clara.burgert@icfi.com

Acknowledgment: The authors would like to acknowledge the following people for their assistance: Livia Montana, David Johnson, Andrew Inglis, and Fred Arnold.

Editor: Sidney Moore
Document Production: Chris Gramer

This study was carried out with support provided by the United States Agency for International Development (USAID) through the MEASURE DHS project (#GPO-C-00-08-00008-00). The views expressed are those of the authors and do not necessarily reflect the views of USAID or the United States Government.

Recommended citation:

Burgert, Clara R., Josh Colston, Thea Roy, and Blake Zachary. 2013. *Geographic displacement procedure and georeferenced data release policy for the Demographic and Health Surveys.* DHS Spatial Analysis Reports No. 7. Calverton, Maryland, USA: ICF International.

Contents

List of Tables iv

List of Figures iv

List of Abbreviations iv

Preface v

Executive Summary vii

1 Background 1

1.1 Confidentiality of Geographic Data 1

1.1.1 Aggregated Data Disclosure 3

1.1.2 Geographic Masking 4

1.2 Demographic and Health Survey Georeferenced Data 5

2 DHS Georeferenced Data-release Policy 6

2.1 GPS Coordinate Data Collection, Aggregation, and Validation 7

2.2 GPS Coordinate Displacement Process 9

3 Case Studies 11

3.1 Case Study 1: Distribution of Displaced Coordinates 11

3.2 Case Study 2: Administrative Unit Displacement Restriction 16

3.3 Case Study 3: Enumeration Area Disclosure 18

4 Discussion and Conclusion 20

Appendix 23

Appendix A 23

Appendix B 25

References 35

List of Tables

Table 1: Summary of geographic data release methods with examples 2
Table 2: DHS household survey displacement distances 14
Table 3: Displacement Restriction Case Study 17
Table 4: Enumeration Area Case Study 19

List of Figures

Figure 1: DHS household survey displacement process 10
Figure 2: Dartboard Displacement Method Illustrated 10
Figure 3: Urban Displaced Distance Distribution Simulation 12
Figure 4: Rural Displacement Distance Distribution 13
Figure 5: Cluster and Enumeration Area Illustration 18

List of Abbreviations

AIS	AIDS Indicator Survey
CDC	Centers for Disease Control and Prevention
DHS	Demographic and Health Surveys
EA	Enumeration Areas
FIA	Forest Inventory Analysis
GADM	Global Administrative Areas
GIS	Geographic Information System
GPS	Global Positioning System
HHS	U.S. Department of Health and Human Service
IHSN	International Household Survey Network
MICS	Multiple Indicator Cluster Survey
MIS	Malaria Indicator Surveys
NASS	National Agricultural Statistics Service
SALB	Second Administrative Level Boundaries
USDA	United States Department of Agriculture
WASAP	West Africa Spatial Analysis Project

Preface

One of the most significant contributions of the Demographic and Health Surveys (DHS) program since its initiation in 1984 is the creation of an internationally comparable body of data on the demographic and health characteristics of populations in developing countries. These data have been augmented in recent years by the addition of more spatial data in the datasets.

The *DHS Spatial Analysis* series joins the existing DHS comparative and analytical report series to meet the growing interest and use of demographic and health data in a spatial realm. The principal objectives of all DHS report series are to provide information for policy formulation at the international level and to examine individual country results in an international context.

Studies in the DHS Spatial Analysis series are based on a variable number of data sets, depending on the topic being examined. A range of methodologies are used in these studies, including geostatistical and multivariate statistical techniques. The topics covered are selected by DHS staff in consultation with the U.S. Agency for International Development.

It is anticipated that the *DHS Spatial Analysis* studies will enhance the understanding of analysts and policymakers regarding significant issues in the fields of international population and health and spatial analysis.

Sunita Kishor
Project Director

Executive Summary

Georeferencing population-based surveys such as the Demographic and Health Surveys (DHS) have many benefits. Most important, researchers can analyze respondent locations spatially to identify geographical patterns associated with specific demographic and health outcomes and programs. Second, the proximity of survey communities to geographic locations such as health centers, roads, and cities can serve as a proxy for access to services; and third, data from sampled locations can be aggregated to form new units of analysis such as climatic zones or program intervention areas, rather than being constrained to administrative units. However, while it is important to make available to researchers, analysts, and policymakers the georeferenced data from population-based surveys, it is also important to maintain the confidentiality of survey respondents.

This report describes the geographic displacement procedures and georeferenced data release policy developed by the DHS project to protect the identity of survey respondents. The georeferenced data release policy applies specifically to the release of georeferenced data from DHS household surveys. It aims to balance the need to protect respondent confidentiality with the need to make available to the public analytically useful data. The policy incorporates two levels of protection: first, data from the same enumeration area (EA) are aggregated to a single point coordinate; then the coordinate is geomasked through use of the Global Positioning System (GPS) coordinate displacement process. In DHS household surveys the GPS coordinate displacement process is carried out as follows: urban clusters are displaced a distance up to two kilometers (0-2 km) and rural clusters are displaced a distance up to five kilometers (0-5 km), with a further, randomly-selected 1% (every 100th) of rural clusters displaced a distance up to 10 kilometers (0-10 km).

Analysis of both simulated and real DHS household survey data shows that the GPS coordinate displacement process produces data with displaced distances that are uniformly distributed. Furthermore, the addition of 1% of rural points that are displaced up to 10 km—for purposes of reducing disclosure risk in rural areas—affects very few points and does not change the overall average distribution of rural coordinates. Analysis of the effect of adding restrictions to the displacement process to prevent points from being displaced across administrative boundaries shows that, in most cases, adding restrictions does not change the average displacement, as long as the units used for the

restrictions are not too small. Comparing the number of households in sampled enumerations areas (EA) and the total number of households in all the EAs that fall within the displacement buffer shows an increase of between 2 to 18 times as many households with the displacement.

The geographic displacement procedure and parameters used by the MEASURE DHS project are supported by real data applications. As DHS survey countries transition from limited spatial data infrastructure to consistent production of reliable EA shapefiles and accurate population density layers, other approaches will need to be considered in the ongoing effort to maintain respondent confidentiality while providing public access to DHS household survey georeferenced datasets.

1 Background

There are many potential benefits to georeferencing population-based surveys. First, researchers can analyze respondent locations spatially to identify geographical patterns associated with specific demographic and health outcomes. Second, the proximity of survey communities to geographic locations of interest such as health centers, roads, and cities can serve as a proxy for access to services. Third, data from the sampled locations can be aggregated to form new units of analysis such as climatic zones or program intervention areas, rather than being constrained to administrative units. Fourth, overlaying coordinates with gridded surface layers in a Geographic Information System (GIS) allows for the extraction of values for remotely-sensed indicators such as altitude. However, releasing georeferenced data for use by researchers is not without its drawbacks, particularly with regard to the issue of respondent confidentiality.

1.1 Confidentiality of Geographic Data

To ensure that ethical standards are applied to population research, protection of respondent confidentiality is one of the fundamental guiding principles for administrators of population data (VanWey et al., 2005). The benefits of georeferencing data must be weighed against the risk of identity exposure for individual survey respondent. If the structure of the geographic information does not prevent identification of individual respondents. Therefore, when releasing such information, any obvious individual, household, or cluster identifiers such as names, ID numbers or addresses, are suppressed from the final dataset, preventing data users from linking that information to the particular individuals. When a data user discovers something about a person from a released dataset, it is called "disclosure" (Hundepool et al., 2010); "disclosure risk" is the likelihood of this happening for a given survey. The information disclosed may be the identity of the individual (identity disclosure) or an associated attribute value (attribute disclosure) (Hundepool et al., 2010).

As GIS methods become more sophisticated, it is increasingly possible to link published health information back to individuals using their geographic location (Hampton et al., 2010). Several published articles have drawn attention to the ease with which, using standard GIS techniques, maps displaying coordinates representing case locations in the United States can be georeferenced, reverse geocoded, or otherwise reengineered to identify individual residence addresses, based on very little spatial reference information (Brownstein et al., 2006a; Brownstein et al., 2006b; Curtis et al., 2006). A

respondent's geographical location should therefore be treated as an indirect identifier (Dupriez and Boyko, 2010) and should be protected.

A balance needs to be struck between the requirement to make data public and the desire to link it to geographic coordinates and the ethical obligation to safeguard the confidentiality of individual survey respondents (VanWey et al., 2005). There is however little agreement on the level of disclosure risk considered "tolerable" for a given survey. Moreover, there is little consensus among experts—as well as a lack of accepted best practices—regarding how geographic data should be disseminated in order to minimize disclosure risk (Brownstein et al., 2006b; National Academies, 2005). Following a brief flurry of interest in the mid-2000s little has been published in the academic literature on safeguards to respondent confidentiality. Because of the lack of standardized guidance on this issue, there are some household surveys for which no geographic identifiers are publically released, other surveys have adopted aggregation, or geomasking approaches so that data can still be shared publically (Table 1). All have implications for confidentiality and the ability to analyze the data at low spatial resolution.

Table 1: Summary of geographic data release methods with examples

Geographic data release method		Example surveys	Agency
Aggregation			
Public-access	National	Vital Statistics Birth Certificates (2005-present)	CDC
	Region	Multiple Indicator Cluster Surveys (MICS)	UNICEF
		International Reproductive Health Survey	CDC
		National Health Interview Survey (NHIS)	CDC
	Estimation IAP Area of residence	National Immunization Survey (NIS)	CDC
	Metropolitan/micropolitan statistical area (MMSA)	Behavioral Risk Factor Surveillance System (BFRSS)	CDC
		Vital Statistics Birth Certificates (1994-2004)	CDC
	Census tract	US Census	US Census Bureau
	Cluster	Demographic Surveillance System (DSS)	INDEPTH Network
Restricted-use	Zip-code of residence	National Immunization Survey (NIS)	CDC
		National Survey of Family Growth (NSFG)	CDC
		National Survey of Children's Health (NSCH) (Derived best zip-code)	HHS
	Block of residence (census)	National Health and Nutrition Examination Survey (NHANES)	CDC
		National Health Interview Survey (NHIS)	CDC
Geographic-masking			
Public-access	Swapping	Forest Inventory Analysis Program	USDA
	Displacement	Forest Inventory Analysis Program	USDA
		Living Standard Indicator Survey (LSMS)	World Bank
		Demographic and Health Surveys (DHS)	ICF Int.

1.1.1 Aggregated Data Disclosure

One way that data administrators attempt to eliminate disclosure risk is by releasing data only after it has been tabulated or aggregated from the individual or household level to a higher level, that of a larger administrative unit. This is the approach endorsed by the International Household Survey Network (IHSN). In its data dissemination guidelines, the IHSN recommends stripping data records of all geographical identifiers below the stratum level—the lowest level at which the sample design is representative (Dupriez and Boyko, 2010). For example, UNICEF's Multiple Indicator Cluster Survey (MICS) collects cluster-level Global Positioning System (GPS) data but the administrators do not release this information, or any other geographical identifier, below the region level (MICS Team, 2013). The Centers for Disease Control and Prevention (CDC) publically releases geographic data from various national surveys—in most cases, aggregated datasets available down to the level of the county, zip code, or block— but these are only available through a Research Data Center. A Research Data Center is a secure location that serves as a repository for sensitive data. Members of the public may be given permission to enter a Research Data Center and use the datasets, subject to the submission and approval of a research proposal (CDC, 2013). For other CDC surveys, some higher-levels of aggregation (states) are available for unrestricted public use. The U.S. Census Bureau, on the other hand, makes GIS data available from the U.S. Census for download in a variety of formats (including polygon shapefiles) down to the level of the census tract (U.S. Census Bureau, 2013).

Under the aggregation approach, data may be aggregated to coordinate points instead of areal units (polygons). The National Agricultural Statistics Service (NASS), for example, provides geographic data collected for its Census of Agriculture surveys as dot-density maps. The dots do not represent actual locations of the respondents but instead are randomly placed within land use polygons according to a customized statistical algorithms (USDA, 2012a; USDA, 2012b). Another approach is to assign every record in the dataset the coordinates of the centroid of the administrative unit in which it is located (Allshouse et al., 2010). However, this means that the range of potential offset—the maximum distance between the true location of the individual and the location assigned to the record in the dataset—varies depending on the size of the administrative unit. The size of an administrative unit tends to vary with its population density, which, in turn, may be determined by environmental factors (e.g., presence of deserts or rivers) and could confound the results of a geospatial analysis.

The aggregation approach protects confidentiality at the expense of spatial resolution and may mask health outcomes with a focal or clustered distribution, particularly in disease patterns that cross geopolitical boundaries (Allshouse et al., 2010; Armstrong et al., 1999). Examples of surveys using aggregation are summarized in Table 1.

1.1.2 Geographic Masking

Another approach to minimizing disclosure risk is geographic masking, or geomasking. This approach alters a record's geographic location in an unpredictable way that is sufficient for preserving the spatial distribution of the variables while minimizing the possibility of identification of individuals (Allshouse et al., 2010). There are three main methods of geomasking: swapping, truncating, and displacing coordinates.

The systematic swapping of locations is used by the U.S. Department of Agriculture (USDA) on a small proportion of the plot coordinates sampled in its Forest Inventory Analysis (FIA) program (McRoberts et al., 2005). Information from one plot location is exchanged with the information from another plot location. Another method of geomasking is truncating or rounding the coordinates to a specified number of decimal places or significant digits. When mapped, the points appear at the vertices of a grid or graticule. The range of potential geographical error, which depends on the number of significant digits, is quantifiable for all records.

A third geomasking method is the random, deterministic relocation of respondents' location identifiers to within a given distance of their true location—a process called "displacing" coordinates (although elsewhere, the same or similar processes have been called "spatial skewing," "fuzzing," "perturbing," "geo-scrambling," or "geographical off-setting") (Allshouse et al., 2010; Armstrong and Ruggles, 2005; Brownstein et al., 2006; Curtis et al., 2006; Hampton et al., 2010; VanWey et al., 2005). Displacement is the process of systematically introducing error to GPS coordinates data by "shifting" the coordinates under set parameters. In this method, each record is assigned the coordinates of a randomly selected point that falls within a circular buffer around the original point; the radius of the buffer corresponds to a specified maximum displacement distance. Each displaced coordinate can be thought of as having a circular "error" buffer around it within which the data user can be certain that the true location falls. The maximum displacement distance (the buffer radius) is specified according to the needs of the particular survey, for example, the level of disclosure risk that may be tolerated given the sensitivity of the information collected. As displacement distance increases, the likelihood of an individual respondent

being identified decreases; at the same time, however, the amount of spatial error introduced increases. This pattern occurs because points displaced over greater distances are, on average, less likely to be similar to their original points (with respect to spatially determined attributes) than points displaced over a shorter distances (McRoberts et al., 2005).

The U.S. Forest Service routinely displaces GPS coordinate data on the location of plots sampled in its FIA program, a process whereby the coordinates of plots are relocated within one mile of the original location. This action is in addition to the swapping of coordinate information mentioned earlier (McRoberts et al., 2005; U.S. Forest Service, 2011). The process is restricted so that the displaced location falls within the same US county as the true location and points are not displaced into large bodies of water (McRoberts et al., 2005). A study that looked at the effects of this method of displacement (for purposes of geomasking) found that when the estimated values for spatial attributes obtained from the displaced points were compared with the true points, the differences were negligible—always less than 1.0% and usually less than 0.5% (McRoberts et al., 2005).

1.2 Demographic and Health Survey Georeferenced Data

The Demographic and Health Surveys (DHS) project has earned a worldwide reputation for collecting and disseminating accurate, nationally representative data on fertility, family planning, maternal and child health, gender, HIV/AIDS, malaria, and nutrition. Data from the DHS household surveys are widely used to advance global understanding of health and population trends in developing countries as well as for planning and monitoring of development programs. Since the beginning of the project in 1984, it has provided technical assistance to more than 300 surveys in over 90 countries and it is committed to making this data openly available. All individual, household and cluster identifiers are removed from the datasets of all household surveys prior to their release. These standard respondent confidentiality measures are carried out on all three "DHS household surveys," the classic Demographic and Health Survey (DHS), the AIDS Indicator Survey (AIS), and the Malaria Indicator Survey (MIS).

The DHS household surveys primarily use a two-stage cluster sampling design within sample domains. DHS household survey samples are designed to give indicator estimates that are nationally representative, as well as representative at the lower level of DHS regions and urban/rural residence. DHS "regions" are sub-national units defined for purposes of the survey that usually correspond to existing administrative units or groupings of these units. Increasingly, DHS household surveys are representative at lower levels of administrative units. The urban/rural "residence" of clusters, as defined

by the country's census bureau, is usually also part of the sampling domain. Clusters are preexisting, geographic groupings within the population of interest that are the primary sampling units. In the majority of DHS household surveys, census enumeration areas (EAs), as defined by the country's census bureau, become the survey clusters. An EA can be a city block or apartment building in urban areas, while in rural areas it is typically a village or group of villages. The population and size of sampled clusters vary between and within countries; typically, clusters contain 100 to 300 households, of which 20 to 30 households are randomly selected for survey participation.

The DHS project started georeferencing coordinate data of cluster locations in the late 1980s and began making georeferenced GPS datasets available to the public in 2003. The georeferenced datasets can be linked to individual records in DHS household surveys through unique identifiers; however, the georeferenced datasets are kept separate from the main household data files and are available only by special permission. Through an online application process, researchers requesting access to the georeferenced data must submit an abstract describing how the GPS coordinate locations will be used in their project; additionally, they must agree to the conditions of use specified by the DHS project. Since the GPS data became available, numerous peer-reviewed articles have been published based on geospatial analyses of DHS data. In 2012 alone, the DHS project approved 731 requests for access to GPS coordinate datasets.

A particular focus of this report is to describe the DHS project's georeferenced data release policy and, specifically, the GPS coordinate displacement process. Three case studies are presented to illustrate the impact of the GPS coordinate displacement process on three spatial attributes: 1) distribution of displaced coordinates, 2) displacement restriction, and 3) enumeration area disclosure.

2 DHS Georeferenced Data-release Policy

Over a period of years, the DHS project has developed a standard georeferenced data-release policy that guides the manner in which georeferenced data from household surveys are released to the public. The policy, which focuses on cluster data, seeks to significantly reduce the disclosure risk associated with the use of spatial data, while preserving the usefulness of this information for reference mapping and GIS analysis. The DHS georeferenced data-release policy has two separate components. First, the cluster is assigned the coordinates of the center of the sampled EA—a type of aggregation. Second, the data are geomasked using a GPS coordinate displacement process.

Initially, the GPS coordinate displacement process was applied only to survey datasets that had an HIV-testing component; other datasets were released with undisplaced coordinates. In 2008, all DHS household surveys started using displacement procedures because, in addition to HIV status, other types of sensitive data were being collected. Georeferenced data collected from 2003-2008 was retroactively displaced and, subsequently, only displaced data have been publically released. Currently, in surveys with HIV testing the original undisplaced georeferenced data, sample frame, questionnaires, raw data files, and scramble-link file are destroyed before the survey dataset is publically released. For surveys without HIV testing, these files are achieved, not destroyed. As of August 2013, the DHS georeferenced data-release policy had been applied to 113 publically released georeferenced DHS household survey datasets; of these, 38 included HIV testing. Some early datasets that were georeferenced through the West Africa Spatial Analysis Project (WASAP) were 100% gazetted and never displaced (Hill, 1998).

It should be noted, that the GPS coordinate displacement process is distinct from “cluster number scrambling,” done only in surveys with HIV testing. Scrambling, which randomly reassigns DHS survey ID numbers to clusters, prevents the possibility of tracing backward from the final data files to the sample frame or individual questionnaires.

2.1 GPS Coordinate Data Collection, Aggregation, and Validation

In most DHS household surveys, field teams routinely use GPS receivers with a positional accuracy of 15 meters or less to georeference the location of the center of the populated areas of the sampled clusters (ICF Macro, 2011). To date, the DHS project has georeferenced more than 58,000 cluster survey locations. For clusters without GPS readings, other means are used to determine the coordinates. Coordinates may be extracted from paper maps, gazetteers of settlement names, or preexisting census data files provided by the country's census agency or statistics authority. When coordinates cannot be georeferenced, the cluster’s location is marked as “missing.” The DHS project never releases the GPS coordinates of individual households. In some surveys, household-level GPS data is collected for survey logistic purposes only and that data is used to calculate the coordinates of the centroid of the surveyed households.

Many countries have increased their geospatial infrastructure and created electronic files of the geographic boundaries of census enumeration areas. When a survey country is willing to share their data with the DHS project the country's spatial data can be used for georeferencing if it is of good quality, thus avoiding the need for additional GPS data collection. In such cases, the geographic center of the EA boundary is calculated and used as the cluster coordinate location. However, the preferred method is still for the DHS survey to collect the GPS coordinate data and to use the EA boundaries as part of the data validation process; this procedure gives an estimate of the populated center of the cluster, not just the geographic center of the cluster.

Prior to release of the GPS coordinate dataset for a survey, the cluster coordinates are verified by the DHS project geographic data specialists. This process includes three main steps:

1. The datum and projection of the data are converted to WGS84, the standard that is used across all datasets.
2. Data are mapped in a GIS and checked to ensure that each coordinate falls in its correct administrative unit.
3. The naming conventions in the data scheme are standardized to match the final survey datasets.

In Step 2, verification of the GPS coordinate locations is done to the lowest administrative unit that exists in both the sample frame and the most accurate geographic boundary GIS file. The geographic boundary files are usually either provided by the country or obtained from publicly available sources such as the United Nation's Second Administrative Level Boundaries (SALB) dataset and the Global Administrative Areas (GADM) database (UNGIWG, 2013; GADM, 2012). There are often large variations in the precision and accuracy of administrative boundary datasets (for example along coastlines or rivers or due to generalization), which makes borders less precise. This means that the ability to properly verify cluster locations depends on the accuracy of the boundaries and at times judgment calls need to be made when a coordinate lies just over a boundary. Gazetteers of village names, where available, may be used to validate data. Rural coordinates that fall less than 10 kilometers from the border of their correct administrative unit (according to the sample file) are accepted and are manually displaced into the proper administrative unit. Urban coordinates that fall less than two kilometers from the border of their correct administrative unit are also manually displaced into the proper administrative unit.

In 2012, the DHS project carried out a complete data audit of all georeferenced data files. For non-HIV- testing survey, this process included re-verifying that all the original georeferenced data coordinates were located within the correct DHS survey region and verifying that the displaced data coordinates were located within the correct DHS survey region. In addition, for all datasets an updated file structure was created with extensive metadata. The updated data schema now used in all datasets and the attribute definitions are shown in Appendix A.

2.2 GPS Coordinate Displacement Process

In DHS household surveys, a GPS coordinate displacement process is carried out as follows:

- Urban clusters are displaced a distance up to two kilometers.
- Rural clusters are displaced a distance up to five kilometers, with a further, randomly selected 1% of the rural clusters displaced a distance up to ten kilometers.

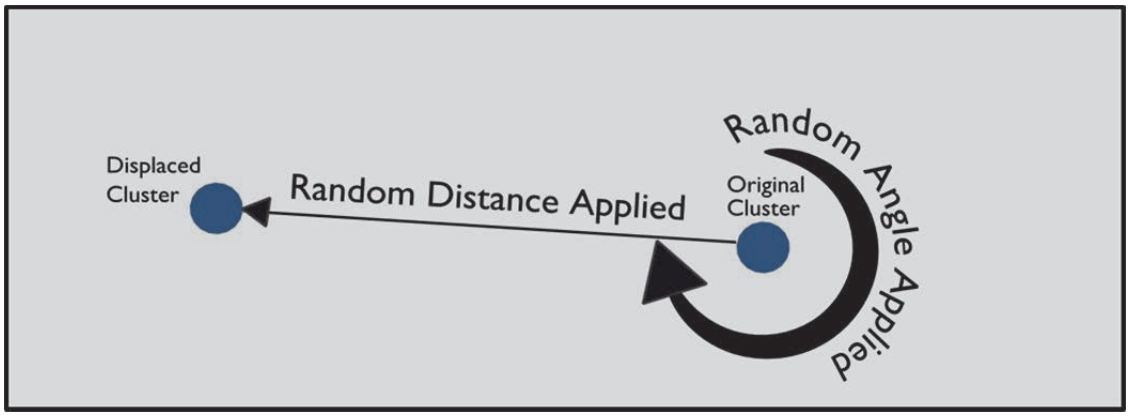
The reason for treating urban and rural clusters differently, according to VanWey et al., is that clusters in sparsely populated, rural areas need larger displacement distances to obtain the same level of (reduced) disclosure risk as clusters in densely populated, urban areas.

GPS coordinates are displaced according to the "random direction, random distance" method. Since March 2011, the displacement process has been automated through the use of a custom-built Python tool in ArcGIS for Desktop (ESRI, 2012; Collins, 2011). Code for the displacement process is presented in Appendix B. For datasets released from 2003 to 2010, displacement was carried out using a Microsoft Excel table that used a random generation formula but did not restrict the data. Both tools use the following basic steps for each coordinate (see Figure 1):

1) Select a random direction (angle) between 0 and 360 degrees.
2) Select a random distance according to the urban and rural parameters.
3) Combine the results of steps 1 and 2; assign the new coordinate to the cluster.

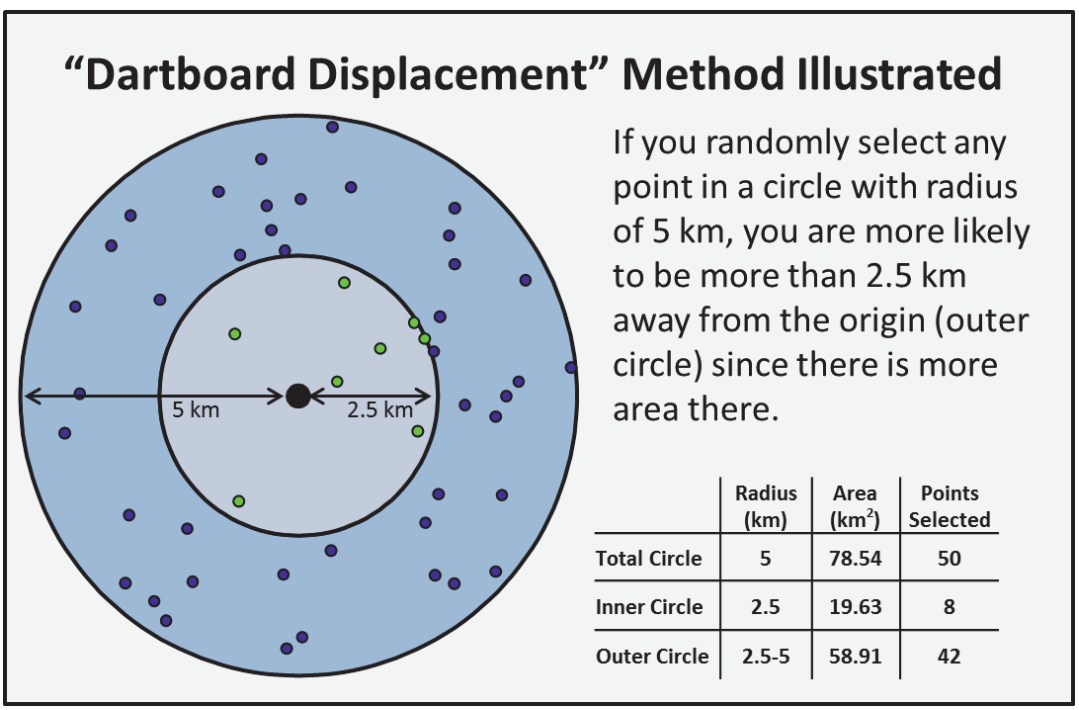
Using the Python script, the new coordinate is checked to make sure it falls within the designated administrative unit and has not been moved across a national or internal administrative boundary. If it has, the process repeats and the coordinate is re-displaced until it falls in a location that does not violate the restriction. The displacement is restricted so that the coordinates stay within the country and within the DHS survey region. In surveys conducted after 2008 the displacement is further restricted for some surveys to one level below the DHS survey region (usually the second administrative unit). Prior to the availability of Python script, this check was done manually for the DHS survey region unit only.

Figure 1: DHS household survey displacement process



The "random direction, random distance" method contrasts with an alternative but similar method, "dartboard displacement" (see Figure 3). In this method, a random coordinate is selected within a circular radius around the original coordinate. Coordinates randomly selected in this method stand a greater chance of being a further from the original coordinate because the area of the outer circle is larger than the area of the inner circle. Figure 2 illustrates this approach using a circle with radius of five kilometers.

Figure 2: Dartboard Displacement Method Illustrated



	Radius (km)	Area (km^2)	Points Selected
Total Circle	5	78.54	50
Inner Circle	2.5	19.63	8
Outer Circle	2.5-5	58.91	42

3 Case Studies

Three case studies were conducted to examine the impact the GPS coordinate displacement procedure has on spatial attributes. Case Study 1 examines the distribution of displaced coordinates both with simulated and real datasets. Case Study 2 examines the impact of administrative unit displacement restriction on the average displacement distance. Finally, Case Study 3 examines household numbers in displaced potential enumeration areas and the population density of displacement buffers.

3.1 Case Study 1: Distribution of Displaced Coordinates

When applied across all clusters in a survey country with no administrative unit restrictions, the GPS coordinate displacement process produces a near uniform distribution, with an average displacement of 1.0 kilometers for urban areas and 2.5 kilometers for rural areas. To illustrate this result, the authors used, a simulated dataset with 10,000 records, all located at latitude 0 and longitude 0. The coordinates in the simulated dataset were then displaced according to the "random direction, random distance" method described above, with no restriction for administrative units. The coordinates were displaced once using the urban parameters and again using the rural parameters. Figure 3 shows a histogram of the distribution of displacement distances for urban clusters and displays the location of the 10,000 coordinates for the original and displaced coordinates. As expected, there is largely uniform distribution of distances for urban clusters. Figure 4 shows a histogram of the distribution of displacement distances for rural clusters based on the simulated dataset of 10,000 coordinates and the spatial location of these coordinates around the original rural location. The distribution of distances for rural clusters is fairly uniform within the 5-kilometer buffer, and only a small number of coordinates are scattered in the 10-kilometer buffer.

In practice, because only 1% of rural coordinates are displaced up to 10 kilometers, the number of coordinates displaced 5 to 10 kilometers accounts for a very small percentage of clusters. In the simulated dataset of 10,000 coordinates, only 52 (approximately 0.52%) had displacement distances greater than 5 kilometers. On average, surveys with georeferenced data have about 300 rural clusters, which means that approximately three coordinates could be displaced up to 10 kilometers, with a 50% chance of these coordinates falling less than or equal to 5 kilometers from the original location. Therefore, doubling the maximum displacement distance for 1% of rural coordinates does not affect the expected average rural displacement distance for any given country.

Table 2 shows the average displacement distance for 40 recent DHS household survey datasets. The surveys were selected are a mixture regarding countries, survey types, inclusion of HIV-testing, and displacement restrictions. For rural clusters, the average displaced distance was 2.45 kilometers (varying from 2.21 to 2.68 kilometers), while the average displaced distance for urban clusters was 0.96 kilometers (varying from 0.80 to 1.13 kilometers). It is possible to have very little displacement distance for a given coordinate, which is why the range in the Table 2 shows 0.00 in some cases. For some countries, no rural points were displaced more than 5 kilometers.

Figure 3: Urban Displaced Distance Distribution Simulation

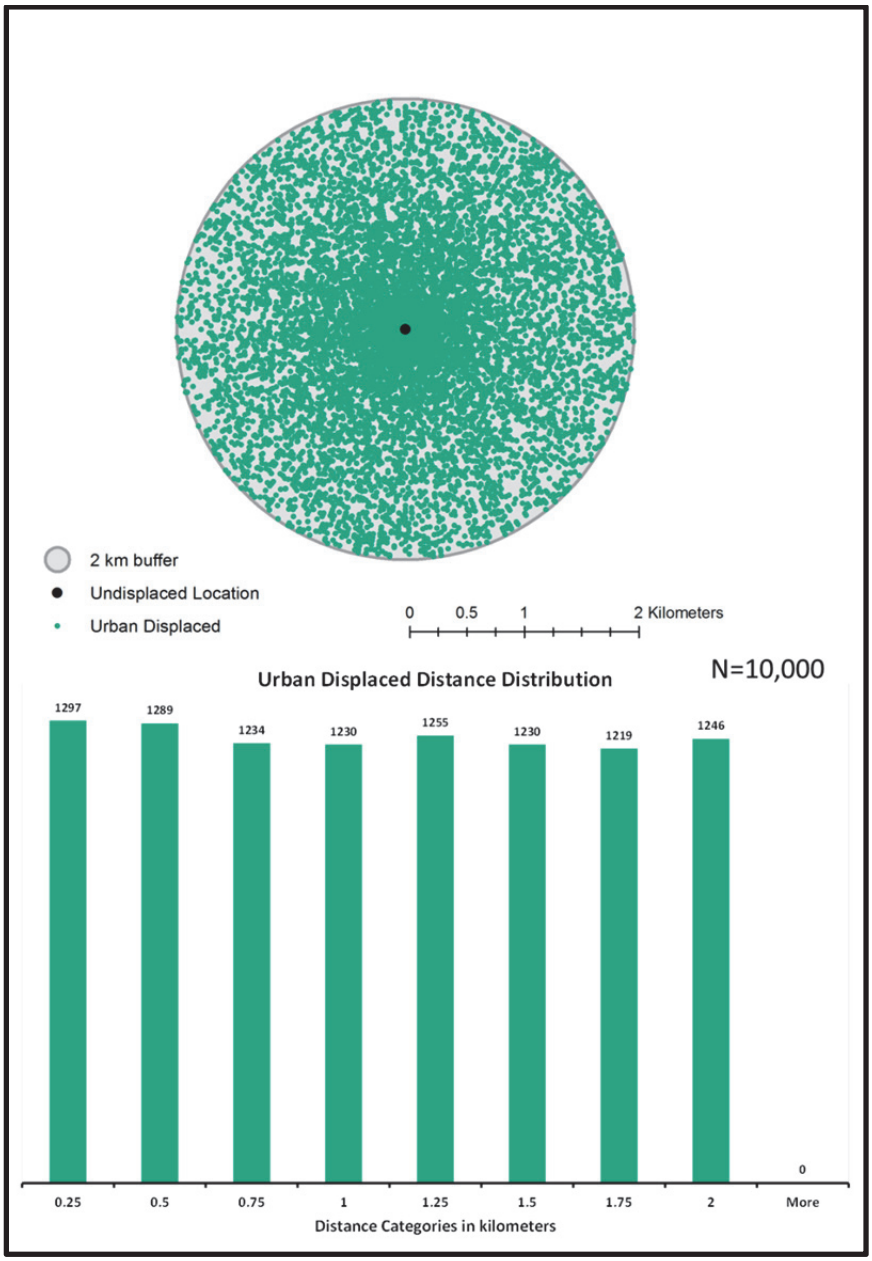


Figure 4: Rural Displacement Distance Distribution

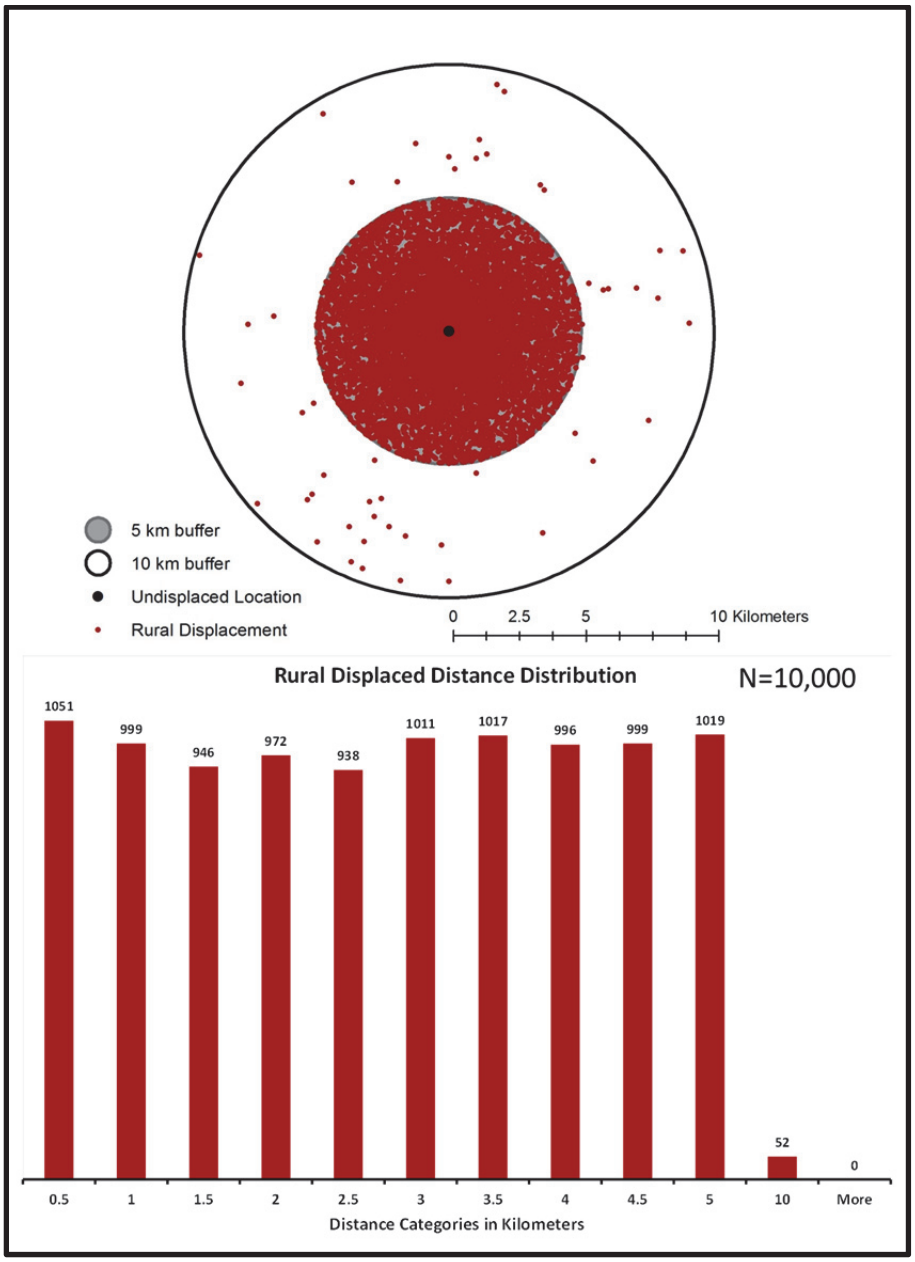


Table 2: DHS household survey displacement distances

	HIV testing	Total Clusters	Missing	Urban			Rural			Restriction
				Clusters	Mean	Range	Clusters	Mean	Range	
Albania DHS 2008	No	450	0	245	1.04	0.00-1.95	205	2.53	0.03-8.98	DHS region
Bangladesh DHS 2011	Yes	600	0	207	0.91	0.00-1.84	393	2.21	0.00-7.63	Admin2
Benin DHS 2001	No	247	0	117	1.05	0.00-1.97	130	2.54	0.02-9.01	DHS region
Bolivia DHS 2008	No	1000	2	593	1.04	0.01-2.00	405	2.68	0.02-9.92	DHS region
Burkina Faso DHS 2010	Yes	573	32	171	0.90	0.00-1.95	370	2.26	0.00-8.63	Admin2
Burundi DHS 2010	Yes	376	0	75	0.80	0.01-1.97	301	2.27	0.01-6.80	Admin2
Cambodia DHS 2010	No	611	4	188	0.85	0.01-1.94	419	2.26	0.01-7.36	Admin2
Cameroon DHS 2011	Yes	580	2	295	1.05	0.00-2.00	283	2.52	0.06-7.26	Admin2
Côte d'Ivoire DHS 1998	No	140	0	100	1.01	0.00-2.00	40	2.46	0.18-4.94	DHS region
Egypt DHS 2008	Yes	1267	20	586	0.98	0.00-2.00	661	2.47	0.01-9.90	DHS region
Ethiopia DHS 2011	Yes	596	25	163	0.97	0.01-1.95	408	2.42	0.02-9.14	Admin2
Ghana DHS 2008	No	411	7	179	0.98	0.00-2.00	225	2.53	0.00-5.01	DHS region
Guinea DHS 1999	No	293	0	115	0.88	0.00-2.00	178	2.55	0.02-5.02	DHS region
Guyana DHS 2009	No	325	13	87	1.00	0.01-1.97	225	2.47	0.08-8.95	Admin2
Haiti DHS 2005	Yes	339	7	158	0.98	0.00-1.99	174	2.49	0.01-4.94	DHS region
Honduras DHS 2011	Yes	1148	20	494	0.97	0.00-1.99	634	2.38	0.00-9.06	DHS region
Indonesia DHS 2002	No	1392	73	557	1.02	0.00-2.00	762	2.59	0.03-9.68	DHS region
Jordan DHS 2007	No	930	4	633	1.03	0.00-1.99	293	2.61	0.01-8.13	DHS region
Kenya DHS 2003	Yes	400	1	129	0.97	0.00-2.00	270	2.46	0.01-9.28	DHS region
Lesotho DHS 2009	Yes	400	5	94	0.89	0.02-1.99	301	2.48	0.01-9.47	DHS region
Madagascar DHS 2008	No	594	9	146	1.05	0.01-2.00	439	2.49	0.01-8.13	Admin2
Malawi DHS 2010	Yes	849	22	151	1.00	0.00-1.95	676	2.33	0.01-9.03	Admin2
Mali DHS 2001	Yes	402	3	129	1.01	0.03-1.98	270	2.50	0.01-7.83	DHS region
Moldova DHS 2005	No	400	1	233	0.96	0.01-1.99	166	2.55	0.00-7.49	DHS region

	HIV testing	Total Clusters	Missing	Urban			Rural			Restriction
				Clusters	Mean	Range	Clusters	Mean	Range	
Mozambique DHS 2011	Yes	610	1	255	0.88	0.00-1.85	354	2.36	0.02-7.31	Admin2
Namibia DHS 2006	No	500	9	208	1.01	0.01-1.99	283	2.39	0.00-9.49	DHS region
Nepal DHS 2011	No	289	0	95	0.81	0.01-1.75	194	2.24	0.03-6.43	Admin2
Niger DHS 1998	No	268	0	90	0.99	0.02-2.00	178	2.57	0.05-5.76	DHS region
Nigeria DHS 2008	No	886	0	279	1.03	0.00-2.00	607	2.58	0.00-9.24	DHS region
Pakistan DHS 2006	No	972	15	385	0.88	0.00-1.78	572	2.22	0.00-8.44	DHS region
Peru DHS 2000	No	1414	6	842	1.03	0.00-2.00	566	2.52	0.00-9.45	DHS Region
Philippines DHS 2003	No	819	3	441	1.00	0.00-2.00	375	2.63	0.00-5.47	DHS Region
Rwanda DHS 2010	Yes	492	0	79	0.96	0.01-1.98	413	2.38	0.00-8.73	Admin2
Senegal DHS 2010	Yes	391	6	147	0.92	0.02-1.91	238	2.36	0.00-9.54	Admin2
Sierra Leone DHS 2008	Yes	353	3	143	0.89	0.01-1.96	207	2.58	0.05-7.85	DHS Region
Tanzania AIS 2011	Yes	583	10	133	0.98	0.02-1.97	440	2.43	0.01-8.03	Admin2
Timor-Leste DHS 2009	No	455	1	116	0.97	0.00-2.01	338	2.47	0.01-9.14	DHS Region
Togo DHS 1998	No	288	1	134	0.91	0.00-1.99	153	2.36	0.02-4.99	DHS Region
Uganda DHS 2011	Yes	404	4	119	1.13	0.02-1.99	281	2.54	0.01-7.85	DHS Region
Zimbabwe DHS 2010	Yes	406	13	165	0.83	0.01-1.90	228	2.47	0.08-8.24	Admin2

3.2 Case Study 2: Administrative Unit Displacement Restriction

For most countries, the restriction of the displacement process to lower-level administrative units is not a problem because it decreases only slightly the average displacement. However, in countries where the administrative units are already geographically very small, implementing this restriction can substantially shorten the average displacement distance; the result is failure to adequately reduce respondent disclosure risk. The usual displacement restriction is one level below the DHS survey region level. In countries where the DHS survey region corresponds to the administrative two units, e.g., the 2010 Malawi DHS, the restriction stayed at that level. Table 3 shows the difference in average displacement distances, according to level of restriction, for three surveys: the 2008 Nigeria DHS, the 2011 Nepal DHS, and the 2011 Bangladesh DHS. These surveys were chosen because of their varying sizes and the availability of administrative unit geographic files. The original undisplaced coordinates were used for this case study because the surveys did not include HIV testing. Unlike the simulated data used in the earlier illustration, only one displacement process was carried out for each level in the three surveys. If the standard restriction had been implemented, the results would be slightly different for each application of the displacement procedure on the same dataset for the same restriction.

The results of examining the data from the three surveys show that the restriction has the greatest impact in rural areas where the displacement distance is larger and, therefore, coordinates are more likely to cross a border. That said, urban administrative units tend to be smaller to begin with. In general, as the number of sub-national units used for the restriction increases, the average displacement distance in rural and urban areas decreases (although the ranges remain similar). Table 2 includes recent DHS household survey datasets that were restricted below the DHS survey region. Of the 40 datasets, 15 were restricted to the administrative two units. The average displacement for rural clusters in the restricted surveys was below the expected average of 2.5 kilometers, going as low as 2.21 kilometers in the Bangladesh survey and as high as 2.52 kilometers in the Cameroon survey. Among the 25 datasets only restricted to DHS region, the rural displacement ranged from a minimum of 2.22 kilometers in Pakistan to a high of 2.68 in Bolivia.

Table 3: Displacement Restriction Case Study

				Area in km^2			Urban Clusters			Rural Clusters		
	Admin area	Admin Level	Number of areas	Minimum	Maximum	Average	Number of Clusters	Average displacement	Range	Number of Clusters	Average displacement	Range
Bangladesh DHS 2011	World	None	-	-	-	-	207	1.01	0.02 - 1.97	391	2.47	0.01 - 5.00
	National	0	1	167,663	167,663	167,663		0.97	0.00 - 1.83		2.24	0.01 - 6.93
	DHS Region/Division	1	7	11,752	37,587	23,954		0.91	0.00 - 1.82		2.22	0.02 - 5.57
	District	2	64	843	6,812	2,625		0.90	0.01 - 1.83		2.21	0.01 - 6.52
	Upazila	3	527	2	1,929	319		0.86	0.01 - 1.82		1.98	0.02 - 8.46
Nepal DHS 2011	World	None	-	-	-	-	95	0.96	0.01 - 1.98	194	2.41	0.04 - 5.01
	National	0	1	191,470	191,470	191,470		0.92	0.02 - 1.78		2.23	0.01 - 7.27
	DHS Region/Region	1	5	26,165	55,925	38,294		0.90	0.05 - 1.76		2.18	0.01 - 5.37
	Zones	2	14	9,146	28,370	13,676		0.89	0.01 - 1.73		2.12	0.08 - 4.38
	District	3	75	136	10,512	2,553		0.85	0.03 - 1.75		2.15	0.02 - 4.35
Nigeria DHS 2008	World	None	-	-	-	-	279	0.99	0.00 - 2.00	605	2.51	0.014 - 9.04
	National	0	1	942,145	942,145	942,145		1.00	0.02 - 1.97		2.57	0.02 - 9.34
	DHS Region	DHS region	6	29,176	292,486	157,024		1.00	0.00 - 1.97		2.46	0.00 - 7.53
	State	1	37	3,300	76,286	25,463		0.97	0.01 - 1.98		2.47	0.02 - 8.98
	Local Government Area	2	779	5	11,613	1,210		0.93	0.00 - 1.97		2.43	0.003 - 6.99

3.3 Case Study 3: Enumeration Area Disclosure

Prior to introducing the DHS georeferenced data release policy in 2003, the DHS project carried out exploratory studies on geographic data from the 2000 Malawi DHS and the 2000 Cambodia DHS. The georeferenced coordinates of the sampled clusters where overlaid with GIS data: EA boundary polygons for Malawi and village locations for Cambodia. Buffers were generated around the original cluster coordinates—2-kilometers for urban coordinates and 5-kilometers for rural coordinates. The purpose of the studies was to calculate the average number of EAs and households that fell within the buffers in order to quantify the reduction in disclosure risk resulting from the displacement policy. Figure 5 shows an example of a 2-kilometer and a 5-kilometer buffer around an original cluster located in EA 920. The 2-kilometer buffer overlays four EAs—the original EA, 920, as well as 919, 921, and 922. The 5-kilometer buffer overlays eight EAs—the four from the 2-kilometer buffer and four additional EAs: 918, 924, 938, and 940.

Figure 5: Cluster and Enumeration Area Illustration

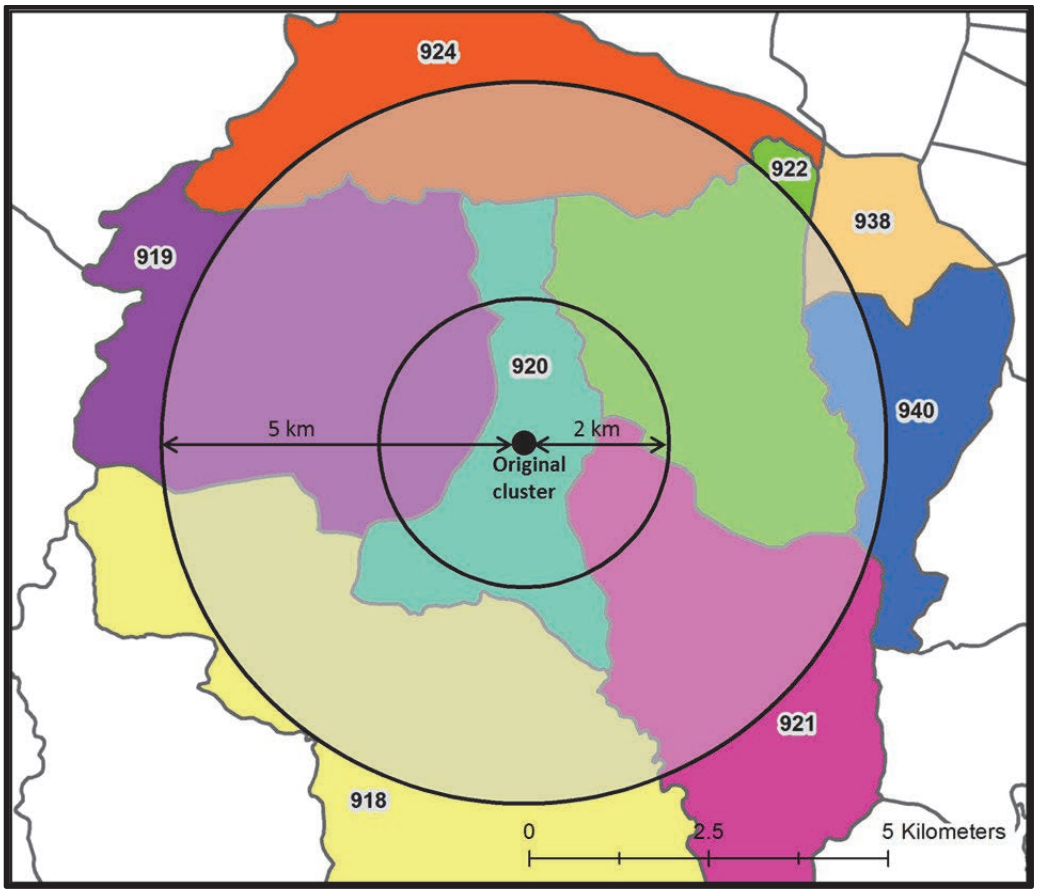


Table 4 summarizes the data for both these initial studies done in 2003 for Malawi and Cambodia. The probability of any given household being included in the sample from the original cluster was compared with the probability of that household being included among all the households that fell within the displacement buffer. In each case, the addition of the displacement process significantly reduced the chances of identifying a particular household compared with using the true geographic location of the cluster. In Malawi, for example, the number of households eligible for the sample increased in rural areas from 214 (in the original EA) to 2,568 households (in all possible EAs); in urban areas, the increase was from 260 to 2,340 households.

The 2003 analysis has been repeated using data from the 2009 Timor-Leste DHS, which had highly detailed EA information (see Table 4). This analysis added in a new component to the original study—estimating the population of the potential displacement area buffer—using the AsiaPop population density raster (Tatem et al., 2013), which has a 100-meter-by-100-meter resolution. Similar to the 2003 studies, the displacement increases the number of potential households more than twofold in urban areas and 18-fold in rural areas. Examining the estimated population from the AsiaPop data, we see a fivefold increase in average population density between the EA areas only and the displacement buffer, for both urban and rural areas.

Table 4: Enumeration Area Case Study

Country and survey year	Malawi DHS 2000		Cambodia DHS 2000		Timor-Leste DHS 2009	
Cluster type	Urban	Rural	Urban	Rural	Urban	Rural
Number of clusters	111	449	94	377	116	338
Buffer size (km)	2	5	2	5	2	5
Average number EA in buffer	8	11	5	16	20	15
Estimated households in original EA	260	214	360	176	861	408
Average estimated households in possible EAs	2,340	2,568	1,800	2,816	1,752	7,676
Average population of original EA from population density	-	-	-	-	1,398	1,216
Average population of buffer from population density	-	-	-	-	7,092	6,402
Notes	Uses 1998 EA boundaries provided by Malawi National Statistical Office and JICA		Villages used not Enumeration Areas		Population density data from AsiaPop 10-year adjusted estimations	

4 Discussion and Conclusion

The DHS household survey georeferenced data release policy aims to balance the need to protect respondent confidentiality with making available to the public analytically useful data. The DHS project's approach to displacement has recently been adopted by the World Bank's Living Standards Measurement Study (LSMS) in several of its surveys (CSA/Ethiopia and World Bank, 2013; NIS/Niger, 2013; NSO/Malawi, 2012). The policy incorporates two levels of protection: first by aggregating the enumeration areas to a single coordinate and then by geomasking the coordinate through the GPS coordinate displacement process. Analysis of both simulated and real DHS household survey data show that the GPS coordinate displacement process produces data with displaced distances that are uniformly distributed. Furthermore, the addition of 1% of the rural coordinates displaced up to 10 kilometers impacts very few coordinates and does not change the overall average distribution of rural coordinates. At the same time, it reduces disclosure risk in rural areas.

Analysis of the level of restriction of the data shows that in most cases the restriction does not significantly change the average displacement distance, as long as the units used for restrictions are not too small. In most countries, the displacement should remain at the administrative two units. In a few countries that have very small administrative two units or the GIS files are not considered accurate, a judgment call may be needed to determine at which level to make the restriction, in order to maintain the desired level of confidentiality. Ultimately, average displacement and the impact of restriction is very country dependent. The addition of restrictions to the displacement process allows for areal unit assignment below the DHS survey region, which is an important feature of the data for many researchers. However, confusion and erroneous analysis can still occur when using current administrative boundary data for past surveys, particularly in cases where the name of the administrative unit stays the same but the borders change (Burgert et al., 2012). Guidance on changes in DHS survey regions over time is available in the boundaries section of the DHS Spatial Data Repository website http://spatialdata.measuredhs.com.

Although limitations to the use of DHS household survey georeferenced data do exist, the geospatial survey data provide many opportunities for analysis that can contribute to improved health and development outcomes and programs. Consideration must be given, however, to the DHS policy regarding the release of georeferenced data for household surveys. Along with the spatial scale of the proposed analysis and the accuracy and temporality of the ancillary data being used, researchers need

to maintain the highest possible levels of respondent confidentiality. The next report in this series, DHS Spatial Analysis Report Number 8, outlines the potential bias in results that can occur with direct distance measurement, continuous and categorical raster point extraction, and areal unit assignment (Perez-Heyrich et al., 2013). The report also presents some potential tools for counter-acting possible misspecification that arises due to the displacement. Inaccuracies in ancillary data can also introduce errors in the analysis, in addition to those introduced by the DHS household survey GPS coordinates. In another recent study, displacement was shown to have relatively little impact on the four studied methods for linking DHS household data with facility-based surveys (Skiles et al., 2013).

The methods and displacement distance selected by researchers are grounded in real data and balance the use of DHS household survey georeferenced datasets with respondent confidentiality. As the DHS survey countries transition from having very little spatial data infrastructure to having reliable and accurate EA shapefiles and accurate population density layers, other approaches could be considered that allow for a reduction in displacement distance. For example, the increased accuracy of population density layers may allow for variable size displacement distance based on maintaining the same population density for all clusters. This and other questions related to the protection of respondent confidentiality will be examined by the DHS project in the coming years; ultimately, the welfare of survey respondents supersedes all other aspects of data use.

Appendix

Appendix A

DHS GPS data files Data Schema

DHSID = The 14 character DHS identification code - DHSCC & DHSYEAR & DHSCLUST (with 8 digits) from survey documentation.

DHSCC = The 2 letter DHS country code (http://www.measuredhs.com/data/File-Types-and-Names.cfm).

DHSYEAR = The 4 digit year of data collection from the survey documentation.

DHSCLUST = The integer cluster identification number. This variable will match v001 in the DHS recode file.

CCFIPS = Federal Information Processing Standards (FIPS) 2 letter country code (http://www.itl.nist.gov/fipspubs/fip10-4.htm).

ADM1FIPS = Federal Information Processing Standards (FIPS) 2 letter country code plus 2 letter/digit first sub-national administrative division code (http://www.itl.nist.gov/fipspubs/fip10-4.htm).

*NOTE: If this information is not available, this field will be "NULL".

ADM1FIPSNA = Federal Information Processing Standards (FIPS) first sub-national administrative division name (http://www.itl.nist.gov/fipspubs/fip10-4.htm).

*NOTE: If this information is not available, this field will be "NULL".

ADM1SALBCO = Second Administrative Level Boundaries (SALB) first sub-national administrative division code (http://www.unsalb.org).

*NOTE: The website requires free registration for downloads.

*NOTE: If this information is not available, this field will be "NULL".

ADM1SALBNA = Second Administrative Level Boundaries (SALB) first sub-national administrative division name (http://www.unsalb.org).

*NOTE: The website requires free registration for downloads.

*NOTE: If this information is not available, this field will be "NULL".

ADM1DHS = First sub-national administrative division code when the DHS sample is representative at the admin 1 level. This variable will usually match v024 in the DHS recode file.

*NOTE: If survey is not representative at the admin 1 level, this field will be "9999".

ADM1NAME = First sub-national administrative division name when the DHS sample is representative at the admin 1 level. This variable will usually match v024 in the DHS recode file.

*NOTE: If survey is not representative at the admin 1 level, this field will be "NULL".

DHSREGCO = The integer region code associated with the DHS region created for sampling. This variable will match either v024 or the country specific region variable in the DHS recode file.

*NOTE: In older templates, REPAR1DHS was used. This field has been renamed DHSREGCO. The REPAR1DHS field is no longer used.

DHSREGNA = The name associated with the DHS region created for sampling. This variable will match either v024 or the country specific region variable in the DHS recode file.

*NOTE: In older templates, REPAR1NAME was used. This field has been renamed DHSREGNA. The REPAR1NAME field is no longer used.

SOURCE = The source of data used to determine the latitude and longitude coordinates:

“GPS” for data collected by the survey team with a global positioning system receiver;

"CEN" for preexisting data provided by the census agency/ministry;

“GAZ” for data extracted from a gazetteer of village/place names;

“MAP” for data extracted from a paper map;

"MIS" for clusters in which data could not be fully verified. Clusters marked as "MIS" will have coordinates 0, 0.

URBAN_RURA = The cluster's Urban (U) and Rural (R) DHS sample classification.

LATNUM = The cluster's latitude coordinate in decimal degrees.

*NOTE: Clusters marked as "MIS" will have coordinates of 0, 0.

LONGNUM = The cluster's longitude coordinate in decimal degrees.

*NOTE: Clusters marked as "MIS" will have coordinates of 0, 0.

ALT_GPS = The cluster's elevation/altitude (in meters) recorded from the GPS receiver.

*NOTE: If this information is not available, this field will be "9999".

ALT_DEM = The cluster's elevation/altitude (in meters) from the SRTM (Shuttle Radar Topography Mission) DEM (Digital Elevation Model) for the specified coordinate location.

*NOTE: Elevations are regularly spaced at 30-arc seconds or approximately 1 kilometer (http://dds.cr.usgs.gov/srtm/version1/SRTM30).

*NOTE: If coordinates are missing, this field will be "9999".

DATUM = The coordinate reference system and geographic datum. It is always "WGS84" for the World Geodetic System (WGS) 1984.

Appendix B

GPS displacement process Python Script

```
@author: bcollins, Blue Raster, LLC
'''
from __future__ import division

import os
import sys
import random
import math
import arcpy
import traceback

class Displacer(object):
    '''
    Displaces a point location while preserving its location
    inside a given polygon
    '''

    def __init__(self):
        pass

    def displacePoint(self, x, y, maxDistance=5000):
        '''
        calculates new point up to a given distance away
        from original point.  All values should be provided
        in meters

        point = (x,y)
        '''
            #The number pi
        PI = 3.14159267

            #Generate a random angle between 0 and 360
        angle_degree = random.randint(0, 360)

            #Convert the random angle from degrees to radians
        angle_radian = (angle_degree) * (PI/180)

            #Generate a random distance by multiplying the max distance by a random number between 0
and 1
        distance = random.random() * maxDistance

            #Generate the offset by applying trig formulas (law of cosines) using the distance as the
hypotenuse solving for the other sides
        xOffset = math.sin(angle_radian) * distance
        yOffset = math.cos(angle_radian) * distance
```

```
        # if(angle_degree > 90 and angle_degree <= 270): xOffset *= -1
        # if(angle_degree > 180): yOffset *= -1

            #Add the offset to the orginal coordinate (in meters)
        new_x = x + xOffset
        new_y = y + yOffset

        return (new_x, new_y)

class GeometryHelpers(object):
    import arcpy
    import math

    def __init__(self):
        pass

    def getCoordinateUnits(self, feature_class):
        sr = arcpy.Describe(feature_class).spatialReference
        units = [sr.type, sr.name, sr.linearUnitName, sr.angularUnitName]
        return units

    def XYToPointGeometry(self, x, y, spatialReference):
        point = arcpy.Point(x, y)
        ptGeometry = arcpy.PointGeometry(point, spatialReference)
        return ptGeometry

            #Convert decimal degrees to meters
    def degreesToMeters(self, xLong, yLat):
            #A fixed conversion factor from degrees to radians
        DEG_TO_RAD = 0.017453292519943296
            #The number pi
        PI = 3.14159267
            #The earth's radius in meters
        EARTH_RADIUS = 6378137

    #This function will provide wrapping around the world, but only to half way back around.
    #This assertions protect against wacky coordinates
        assert (xLong < 360 and xLong > -360), 'longitude outside of wrapping bounds'
        assert (yLat < 180 and yLat > -180), 'latitude outside of wrapping bounds'

    #Wrap around values if necessary
        if(yLat <= -90): yLat = yLat % 90
        if(yLat >= 90): yLat = (yLat % 90) - 90
        if(xLong <= -180): xLong = xLong % 180
        if(xLong >= 180): xLong = (xLong % 180) - 180
```

```
        #The y formula uses yLat as a scalor to correct for differences in the number of meters in a
degree of latitude across the earth
    y = EARTH_RADIUS * math.log(math.tan(((yLat * DEG_TO_RAD) + (PI / 2))/2))
    x = EARTH_RADIUS * (xLong * DEG_TO_RAD)

    return (x, y);

        #Convert meters to decimal degrees
  def metersToDegrees(self, xLong, yLat):
        #A fixed conversion factor from radians to degrees
    RAD_TO_DEG = 57.295779513082322
        #The number pi
    PI = 3.14159267
        #The earth's radius in meters
    EARTH_RADIUS = 6378137

  #Convert meters to decimal degrees
    lat = RAD_TO_DEG * ((2 * math.atan(math.exp(yLat / EARTH_RADIUS))) - (PI/2));
    lon = RAD_TO_DEG * (xLong / EARTH_RADIUS);

  #This function will provide wrapping around the world, but only to half way back around.
  #This assertions protect against wacky coordinates
    assert (lon < 360 and lon > -360), 'longitude outside of wrapping bounds'
    assert (lat < 180 and lat > -180), 'latitude outside of wrapping bounds'

  #Wrap around values if necessary
    if(lat<=-90): lat = lat % 90
    if(lat>=90): lat = (lat % 90) - 90
    if(lon<=-180): lon = lon % 180
    if(lon>=180): lon = (lon % 180) - 180

    return (lon, lat)

  def isGeographicProjection(self, feature_class):
    feature_class_description = arcpy.Describe(feature_class)
    proj_type = feature_class_description.spatialReference.type
    return (proj_type == 'Geographic')

  def validateGeometries(self, point_feature_class, polygon_feature_class):
    point_description = arcpy.Describe(point_feature_class)
    polygon_description = arcpy.Describe(polygon_feature_class)

    point_sr = point_description.spatialReference.name
    polygon_sr = polygon_description.spatialReference.name

    assert (point_sr == polygon_sr), 'Point and Polygon Spatial Reference Mismatch'

  def relatePointsToPolygons(self, point_feature_class, polygon_feature_class):
```

```
        '''
        Input a point and polygon feature class and a receive a dictionary of which points are in which
polygons
        '''
        returnDict = { }
        point_rows = arcpy.SearchCursor(point_feature_class)

        for p in point_rows:
            ppoint = p.getValue('Shape')
            polygon_rows =  arcpy.SearchCursor(polygon_feature_class)
            for q in polygon_rows:
                poly = q.getValue('Shape')
                if(ppoint.within(poly)):
                    returnDict[p] = poly
            del polygon_rows
        del point_rows

        return returnDict

    def createTimestamp(self):
        '''creates a timestamp which can be used to create a unique name.'''
        from time import localtime, strftime
        l = localtime()
        return strftime("%Y-%m-%d_%H_%M", l)

if __name__ == '__main__':
    try:
        #Helper Classes
==========================================================================================
        oDisplacer = Displacer()
        oGeometryHelpers = GeometryHelpers()

        #Input Parameters
==========================================================================================
        POINTS_PATH = arcpy.GetParameterAsText(0)
        POLYGON_PATH = arcpy.GetParameterAsText(1)
        MAX_DISTANCE = int(arcpy.GetParameterAsText(2))
        UPDATE_LAT_LON_MODE = arcpy.GetParameterAsText(3)
        LAT_FIELD = arcpy.GetParameterAsText(4)
        LON_FIELD = arcpy.GetParameterAsText(5)
        URBAN_RURAL_MODE = arcpy.GetParameterAsText(6)
        URBAN_RURAL_FIELD = arcpy.GetParameterAsText(7)
        URBAN_VALUE = arcpy.GetParameterAsText(8)
        RURAL_VALUE = arcpy.GetParameterAsText(9)
        OUTPUT_DATASET = arcpy.GetParameterAsText(10)
        REPORT_LOCATION = arcpy.GetParameterAsText(11)
```

```
    #Setup Basic Report
==========================================================================================
    REPORT_NAME = '_'.join(['Point_Displacement_Report', oGeometryHelpers.createTimestamp() +
'.txt'])
    REPORT_FULL_PATH = os.path.join(REPORT_LOCATION, REPORT_NAME)
    report = open(REPORT_FULL_PATH, 'w')
    report.write(REPORT_NAME + '\n\n')
    report.write('Input Parameters: \n\n')
    report.write('POINTS INPUT PATH: %s \n' % POINTS_PATH)
    report.write('POLYGON INPUT PATH: %s \n' % POLYGON_PATH)
    report.write('MAX_DISTANCE: %i \n' % MAX_DISTANCE)
    report.write('URBAN_RURAL_MODE: %s \n' % URBAN_RURAL_MODE)
    report.write('URBAN_RURAL_FIELD: %s \n' % URBAN_RURAL_FIELD)
    report.write('URBAN_VALUE: %s \n' % URBAN_VALUE)
    report.write('RURAL_VALUE: %s \n' % RURAL_VALUE)
    report.write('OUTPUT_DATASET: %s \n\n' % OUTPUT_DATASET)

    report.write('TOOL MESSAGES: \n\n')

    arcpy.AddMessage('Report Location: %s' % REPORT_FULL_PATH)

    #GET INFORMATION ON POINTS LAYER
=========================================================================
    point_description = arcpy.Describe(POINTS_PATH)
    point_shapefield = point_description.shapeFieldName
    point_fields = arcpy.ListFields(POINTS_PATH)
    point_field_dict = dict([(f.name, f.type) for f in point_fields])
    point_count = arcpy.GetCount_management(POINTS_PATH).getOutput(0)
    point_sr = point_description.spatialReference

    #GET INFORMATION ON POLYGON LAYER
========================================================================
    polygon_description = arcpy.Describe(POLYGON_PATH)
    polygon_shapefield = polygon_description.shapeFieldName
    polygon_fields = arcpy.ListFields(POLYGON_PATH)
    polygon_count = arcpy.GetCount_management(POINTS_PATH).getOutput(0)
    polygon_sr = polygon_description.spatialReference

    WORKSPACE = os.path.split(POINTS_PATH)[0]
    OUTPUT_WORKSPACE = os.path.split(OUTPUT_DATASET)[0]

    #Assert Statement to validate inputs
=======================================================================
    assert arcpy.Exists(POINTS_PATH), 'Point Feature Class does not appear exist'
    assert arcpy.Exists(POLYGON_PATH), 'Polygon Feature Class does not appear exist'
    assert point_description.shapeType == 'Point', 'Point layer does not appear to be a Point layer'
    assert polygon_description.shapeType == 'Polygon', 'Polygon layer does not appear to be a Polygon
layer'
```

```
    assert point_count > 0, 'Dude...there are not any points in this file'
    assert os.path.exists(OUTPUT_WORKSPACE), 'Output Workspace does not appear to exist'
    assert oGeometryHelpers.isGeographicProjection(POINTS_PATH), 'Points file not in geographic
projection'
    assert oGeometryHelpers.isGeographicProjection(POLYGON_PATH), 'Polygon file not in geographic
projection'

    arcpy.env.workspace = WORKSPACE
    arcpy.env.overwriteOutput = True

    '''
    Many of the point datasets imported from excel contain Lat/Long fields
    which are static attributes.  We want these fields to be updated to the new
    lat/long after the point is displaced.  The script writes to these fields
    later on in the 'while' loop, but the code immediately below is used to make
    sure the fields supplied in the tool dialog actually exist.
    '''
    if(UPDATE_LAT_LON_MODE.lower() == 'true' or UPDATE_LAT_LON_MODE == '1'):
      UPDATE_LAT_LON_MODE = True
      assert (LAT_FIELD in point_field_dict.keys()), LAT_FIELD + ' not in point feature class fields'
      assert (LON_FIELD in point_field_dict.keys()), LON_FIELD + ' not in point feature class fields'
      assert (point_field_dict[LAT_FIELD] != 'String'), LAT_FIELD + ' appears to be a String field but
should be a Double'
      assert (point_field_dict[LON_FIELD] != 'String'), LON_FIELD + ' appears to be a String field but
should be a Double'

    else:
      UPDATE_LAT_LON_MODE = False

    #Figure out which Tool Mode we are in (URBAN_RURAL_MODE true/false)
=====================================
    if(URBAN_RURAL_MODE.lower() == 'true' or URBAN_RURAL_MODE == '1'):
      mode_name = ' Urban / Rural Mode'
      URBAN_RURAL_MODE = True
    else:
      mode_name = ' Maximum Displacement = ' + str(MAX_DISTANCE) + 'm'
      URBAN_RURAL_MODE = False

    if(URBAN_RURAL_MODE):
      assert (URBAN_RURAL_FIELD in point_field_dict.keys()), URBAN_RURAL_FIELD + ' not in point
feature class fields'
      point_search = arcpy.SearchCursor(POINTS_PATH)
      for p in point_search:
        locality = p.getValue(URBAN_RURAL_FIELD)
        assert (locality == URBAN_VALUE or locality == RURAL_VALUE), str(locality) + ' does not match
urban/rural values provided. Please check your attribute table and make sure all urban/rural values
match what was entered in the tool dialog.'
      del point_search
```

```
    #Copy Feature Class as not to mess with the original data
==============================================
    arcpy.CopyFeatures_management(POINTS_PATH, OUTPUT_DATASET)

    #MAIN BUSINESS LOGIC
=========================================================================================
    point_rows = arcpy.UpdateCursor(OUTPUT_DATASET)
    arcpy.SetProgressor("step", "Displacing Points...", 0, int(point_count), 1)

    displaced_points = 0
    rural_displaced_points = 0

    #Loop through each of the points
    for p in point_rows:
        point_geom = p.getValue(point_shapefield)

        #Get Max Distance based on tool mode and current point attributes (urban vs. rural)
======================
        if(URBAN_RURAL_MODE):
            locality = p.getValue(URBAN_RURAL_FIELD)

            #Urban Point Displacement Logic
            if(locality == URBAN_VALUE):
                max_displace_distance = 2000

            #Rural Point Displacement Logic
            elif(locality == RURAL_VALUE):
                rural_displaced_points += 1
                max_displace_distance = (rural_displaced_points % 100 == 0) and 10000 or 5000

        #Use distance supplied by user if not in URBAN/RURAL MODE
        else:
            max_displace_distance = MAX_DISTANCE

        #Loop through each of the polygons
        polygon_rows = arcpy.SearchCursor(POLYGON_PATH)
        point_within_study_area = False
        for q in polygon_rows:
            poly_geom = q.getValue(polygon_shapefield)
            if(point_geom.within(poly_geom)):
                point_within_study_area = True

                new_point_within = False
                while(not new_point_within):
                    ppoint = point_geom.firstPoint
```

```
                #Convert from degrees to meters
                meter_x, meter_y = oGeometryHelpers.degreesToMeters(ppoint.X, ppoint.Y)

                #Run displacement function
                displaced_x, displaced_y = oDisplacer.displacePoint(meter_x, meter_y,
max_displace_distance)

                #Convert output back to degrees
                new_x, new_y = oGeometryHelpers.metersToDegrees(displaced_x, displaced_y)
                new_point = arcpy.Point(new_x, new_y)
                new_geometry = arcpy.PointGeometry(new_point, point_sr)

                #Check if point still remains inside the original polygon, if so
                #then this loop will end, if not the process begins again
                new_point_within = new_geometry.within(poly_geom)


            #Update Shape Field
            p.setValue(point_shapefield, new_geometry)

            #Update Static Lat/Lon fields
            if(UPDATE_LAT_LON_MODE):
                p.setValue(LAT_FIELD, new_y)
                p.setValue(LON_FIELD, new_x)

            point_rows.updateRow(p)
            displaced_points += 1
            arcpy.SetProgressorLabel('Displaced ' + str(displaced_points) + ' of ' + str(point_count) +
mode_name)
            arcpy.SetProgressorPosition()

        #Handle if point is not found within any of the study area polygons
==================================
        if(not point_within_study_area):
            arcpy.AddWarning('Point Detected outside of study area.  Please check report for more
details.')
            report.write('POINT OUTSIDE STUDY AREA: \n')
            for f in point_field_dict.keys():
                if(f != point_shapefield):
                    v = p.getValue(f)
                    report.write('    %s: %s \n' % (f, str(v)))

            report.write('\n\n')
        del polygon_rows
        del point_rows

    except AssertionError, e:
        report.write('AssertionError: %s' % e)
```

```
    raise e

except Exception, e:
    tb = sys.exc_info()[2]
    msg = "An error occurred on line %i" % tb.tb_lineno
    report.write('Exception: %s' % e + '\n')
    report.write(msg + '\n')
    arcpy.AddMessage(msg)
    arcpy.AddMessage(arcpy.GetMessages(2))

finally:
    report.close()
```

References

Allshouse, W.B., M.K. Fitch, K.H. Hampton, D.C. Gesink, I.A. Doherty, P.A. Leone, M.L. Serre, and W.C. Miller. 2010. Geomasking sensitive health data and privacy protection: An evaluation using an E911 database. *Geocarto International* 25(6): 443-452. doi: 10.1080/10106049.2010.496496.

Armstrong, M.P., G. Rushton, and D.L. Zimmerman. 1999. Geographically masking health data to preserve confidentiality. *Statistics in Medicine* 18(5): 497-525.

Armstrong, M.P., and A.J. Ruggles. 2005. Geographic information technologies and personal privacy. *Cartographica: The International Journal for Geographic Information and Geovisualization* 40(4): 63-73. doi: 10.3138/RU65-81R3-0W75-8V21.

Collins, Brendan. Boundary Respecting Point Displacement.Python Script. Blue Raster, LLC. 2011.

Brownstein, J.S., C.A. Cassa, I.S. Kohane, and K.D. Mandl. 2006. An unsupervised classification method for inferring original case locations from low-resolution disease maps. *International Journal of Health Geographics* 5: 56. doi: 10.1186/1476-072x-5-56.

Brownstein, J.S., C.A. Cassa, and K.D. Mandl. 2006. No place to hide—reverse identification of patients from published maps. *New England Journal of Medicine* 355(16): 1741-1742. doi: 10.1056/NEJMc061891.

Burgert, C., B. Zachary, and A. Way. 2012. Response to "Problems of spatial linkage of a geo-referenced Demographic and Health Survey (DHS) dataset to a population census: A case study of Egypt." *Computers, Environment and Urban Systems* 36(6): 626-627.

Centers for Disease Control and Prevention (CDC). 2013. *Research Data Center - Geocodes*. Atlanta, Georgia, USA: Centers for Disease Control and Prevention; Hyattsville, Maryland, USA: National Center for Health Statistics. [Accessed August 7, 2013]. Available at: http://www.cdc.gov/rdc/B1DataType/Dt123Geocod.htm.

Central Statistical Agency (CSA) [Ethiopia], and World Bank. 2013. *Living Standards Measurement Study-Integrated Surveys on Agriculture (LSMS-ISA): Ethiopia Rural Socioeconomic Survey (ERSS): Basic Information Documents*. Addis Ababa, Ethiopia: Central Statistical Agency and World Bank. Available at: http://www.csa.gov.et/

Curtis, A.J., J.W. Mills, and M. Leitner. 2006. Spatial confidentiality and GIS: Re-engineering mortality locations from published maps about hurricane Katrina. *International Journal of Health Geographics* 5: 44. doi: 10.1186/1476-072X-5-44.

Dupriez, O., and E. Boyko. 2010. *Dissemination of microdata files: Principles, procedures and practices*. International Household Survey Network (IHSN), IHSN Working Paper No. 005. Available at: http://www.ihsn.org.

ESRI. 2012. *ArcGIS for Desktop*. (Version 10.1). Redlands, CA: ESRI.

GADM. 2013. *GADM Database of Global Administrative Areas.* (Version 2.0, January 2012) [Accessed August 20, 2013]. Available at: http://www.gadm.org/home.

Hampton, K.H., M.K. Fitch, W.B. Allshouse, I.A. Doherty, D.C. Gesink, P.A. Leone, M.L. Serre, and W.C. Miller. 2010. Mapping health data: Improved privacy protection with donut method geomasking. *American Journal of Epidemiology* 172(9): 1062-9. doi: 10.1093/aje/kwq248.

Hill, Nicolas. 1998. West Africa Spatial Analysis Prototype Exploratory Analysis: Creating Social Borders from WASAP datasets. DHS Spatial Analysis Reports No 1. Calverton, Maryland, USA: Macro International.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Naylor, E.S. Nordholt, G. Seri, and P.-P. De Wolf. 2010. *Handbook on statistical disclosure control*. (Version 1.2) The Hague, Netherlands: ESSNet SDC, A Network of Excellence in the European Statistical System in the field of Statistical Disclosure Control.

ICF Macro. 2011. *Incorporating geographic information into Demographic and Health Surveys: A field guide to GPS data collection*. DHS Toolkit. Calverton, Maryland, USA: ICF Macro.

McRoberts, R.E., G.R. Holden, M.D. Nelson, G.C. Liknes, W.K. Moser, A.J. Lister, S.L. King, E.B. LaPoint, J.W. Coulston, W.B. Smith, and G.A. Reams. 2005. Estimating and circumventing the effects of perturbing and swapping inventory plot locations. *Journal of Forestry* 103(6): 275-279.

MICS Team. 2013. Email exchange with Josh Colston, July 29.

National Academies, National Research Council, Division of Behavioral and Social Sciences and Education. 2005. Workshop on Confidentiality Issues in Linking Geographically Explicit and Self-Identifying Data, December 9-10, Washington D.C.

National Institute of Statistics (NIS) [Niger]. 2013. *2011 National Survey on Household Living Conditions and Agriculture* (*ECVM/A-2011*): *Basic information document*. Niamey, Niger: National Institute of Statistics. Available at: http://www.agrodep.org/fr/dataset/niger-national-survey-living-conditions-and-agriculture-2011.

National Statistical Office (NSO) [Malawi]. 2012. *Malawi Third Integrated Household Survey (IHS3) 2010-2011*: *Basic information document*. Zomba, Malawi: National Statistical Office. Available at: http://siteresources.worldbank.org/INTLSMS/Resources/3358986-1233781970982/5800988-1271185595871/IHS3.BID.FINAL.pdf.

Perez-Heydrich, C., J.L. Warren, C.R. Burgert, and M.E. Emch. 2013. *Guidelines on the use of DHS GPS data*. DHS Spatial Analysis Reports No 8. Calverton, Maryland, USA: ICF International.

Skiles, M.P., C.R. Burgert, S.L. Curtis, and J. Spencer. 2013. Geographically linking population and facility surveys: Methodological considerations. *Population Health Metrics* 11: 14.

Tatem, A., C. Linard, and A. Gaughan. 2013. Timor Leste. AsiaPop (Edited by AsiaPop). http://www.clas.ufl.edu/users/atatem/index_files/Timor.htm: AsiaPop. Original edition, Alpha.

U.S. Census Bureau. 2013. *2010 Census data products: United States - At a glance* (Version 2.7) [Accessed July 23, 2013]. Available at: http://www.census.gov/population/www/cen2010/glance/.

United Nations Geographic Information Working Group (UNGIWG). 2013. *Second Administrative Level Boundaries (SALB)*. New York, USA: United Nations. [Accessed August 20, 2013]. Available at: http://www.unsalb.org/.

U.S. Department of Agriculture (USDA). 2012a. *2007 Census Ag Atlas Maps*. USDA Census of Agriculture. Washington, D.C.: National Agricultural Statistics Service (NASS). [Accessed July 23, 2013]. Available at: http://www.agcensus.usda.gov/Publications/2007/Online_Highlights/Ag_Atlas_Maps/.

U.S. Department of Agriculture (USDA). 2012b. *2007 Census of Agriculture: Ag Atlas Maps*. (Geographic Fact Sheets). USDA Census of Agriculture. Washington, D.C.: National Agricultural Statistics Service (NASS). http://www.agcensus.usda.gov/Publications/2007/Online_Highlights/Fact_Sheets/Geographic/ag_atlas.pdf

U.S. Forest Service. 2013. *Forest Inventory and Analysis National Program: Privacy policy and authority.* Last modified: February 22, 2011. [Accessed July 1, 2013]. Available at: http://www.fia.fs.fed.us/tools-data/spatial/Policy/default.asp.

VanWey, L.K., R.R. Rindfuss, M.P. Gutmann, B. Entwisle, and D.L. Balk. 2005. Confidentiality and spatially explicit data: Concerns and challenges. *Proceedings of the National Academy of Sciences of the United States of America* 102(43): 15337-15342. doi: 10.1073/pnas.0507804102.