Spatial Anonymization

Guidance Note prepared for the Inter-Secretariat Working Group on Household Surveys

January 2021

Acknowledgements

This document summarizes key issues related to the dissemination of spatial data with household surveys. The content is for review and comment and is subject to change. The document was produced by members of the DHS program of ICF and the Development Data Group of the World Bank. The core team included Trevor Croft, Trinadh Dontamsetti and Benjamin Mayala from ICF, and Brian Blankespoor and Siobhan Murray from the World Bank. We thank senior management for enabling this collaboration. We are grateful for early insights and ideas provided by World Bank colleagues Jon Kastelic, Talip Kilic and Dunstan Matekenya. We have also benefited from very constructive comments received from reviewers Thijs Benschop (FAO), Mark Montgomery (Stony Brook University) and Dana Thomson (University of Southampton). We thank them for their efforts.

Contents

I. Motivation

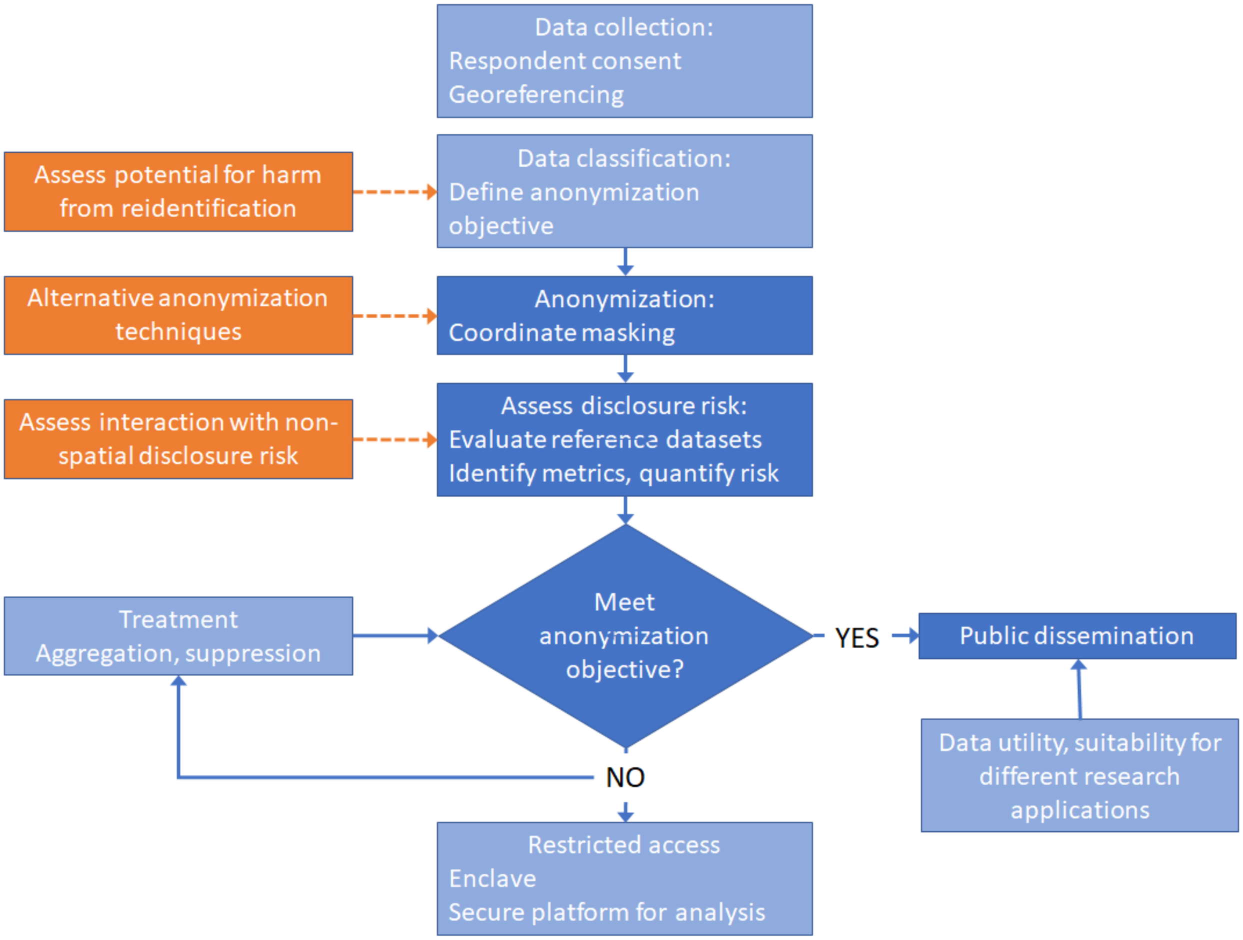
International development organizations, national statistics agencies and research institutions invest significant resources each year in the collection of household, health facility, farm and firm surveys. The collection of geographical locations of dwellings, agricultural parcels and plots, facilities and other locations directly associated with surveyed communities and respondents is now common practice. Exact coordinates could be used to identify individual respondents and are therefore typically removed from public-use datasets. Recognizing the added value of GPS[1] data, efforts have been made to disseminate anonymized spatial data that meets a spectrum of research needs while maintaining confidentiality: these include dissemination of anonymized coordinates, production and dissemination of relevant spatial covariates and production of spatial variables by request, among others.

While the dissemination strategies listed above have thus far provided a satisfactory solution, several issues are driving a renewed interest in the topic of spatial anonymization, fueled by the potentially huge analytical gains of greater access. First, the concept of data privacy and obligation of data providers to safeguard personal data has come into greater focus. The EU General Data Protection Regulation (GDPR) includes location data in its definition of "personal data", holding collectors of such data to high standards in data protection and security. At the same time, there are numerous examples of unintentional but consequential data exposure. Such examples are made increasingly likely because of the advancement in technologies, expansion of publicly available data and satellite imagery, as well as open source tools and cloud computing that facilitate integration of data from many different sources. Lastly, this rich data landscape and analytical applications in turn spur greater demand for access to more precise location information for valuable cutting-edge research applications.

In response to this changing context a review of existing protocols is warranted. Drawing on more than a decade of experience by the teams of the USAID funded Demographic and Health Surveys Program (DHS Program) and the Living Standards Measurement Study - Integrated Surveys on Agriculture (LSMS-ISA) program of the World Bank, this report describes current spatial anonymization protocols and alternatives under consideration, presents some methods for assessing spatial disclosure risk and offers a set of recommended best practices. Figure 1 below summarizes issues covered in the report; blue boxes contain topics covered in detail, light blue indicates minimal coverage and orange boxes highlight important areas for future work.

[1] Global Positioning System (GPS) is used here as the more familiar term for generic Global Navigation Satellite System (GNSS)

Figure 1 Spatial anonymization topics and coverage in the report


Data collection:
Respondent consent
Georeferencing
Assess potential for harm from reidentification
Data classification:
Define anonymization objective
Alternative anonymization techniques
Anonymization:
Coordinate masking
Assess interaction with non-spatial disclosure risk
Assess disclosure risk:
Evaluate reference datasets
Identify metrics, quantify risk
Treatment
Aggregation, suppression
Meet anonymization objective?
YES
Public dissemination
Data utility, suitability for different research applications
NO
Restricted access
Enclave
Secure platform for analysis

II. Review of literature

The proliferation of free and open source high spatial resolution data has led to a rapid increase in detailed publicly available datasets in the past few decades (Wegmann, Leutner & Dech 2016). At the same time, interest in local data for local analyses and decision-making has heightened demand on society to provide detailed geographic and even individual level data (Samarati and Sweeney 1998; Pickle et al. 2006; Mennis and Guo 2009, United Nations General Assembly 2015). The benefits of integrating high spatial resolution data from remote sensing or mobile data with survey data have been demonstrated by many researchers; gaining new insights into poverty (e.g. Jean et al. 2016, Steele et al. 2017), health outcomes from environmental changes (Brown et al. 2014), HIV elimination strategies (Coburn et al. 2017) and health (Richardson et al. 2013, Buckee et al. 2013).

Although the opportunities of using spatially referenced microdata are well demonstrated, assessment of the magnitude of spatial risk of disclosure, or positive identification of place, is challenging (Duncan and Lambert 1989). Richardson et al. (2015) discuss several legal and ethical challenges, as well as weaknesses in standards and practices. While standards exist for non-spatial microdata (Dupriez and Boyko 2010), the guidance does not extend to the spatial domain. In this context data providers have used a variety of methods to anonymize location information, to partially address these challenges and access the untapped potential of spatially referenced microdata.

Anonymization techniques include aggregation, obfuscation, and record swapping. Aggregating a sufficiently large enough number of individual records within a geographic area can address individual privacy concerns (Armstrong et al. 1999; Wieland et al. 2008). This can be applied using hierarchical locality variables, or directly to GPS coordinates. The DHS Program pioneered a geographic displacement method based on cluster-level aggregation and fixed range offset for public use datasets (Burgert et al. 2013, Perez-Heydrich et al. 2013). This method uses different ranges for urban or rural strata, but it does not account for other local characteristics, such as population density. Variants include the donut method, which ensures both a minimum and maximum distance from the location (Hampton et al. 2010). The Verified neighbor method by Richter (2017) "guarantees the specified level of privacy even where population density is uneven while minimizing spatial distortion and changes to the values of environmental variables assigned to subjects". The location swapping method selects a location to replace the original from the universe of masked locations within a specified proximity that have similar spatial characteristics (Zhang et al. 2015). The Adaptive Areal Elimination method considers both the population density to ensure a minimum k-anonymity or an aggregation to the median centers of the areas (Kounadi and Leitner 2016). Cassa et al. (2006) developed a method based on a population-density-based Gaussian spatial blurring. Horey et al. (2012) develop a Negative Quad Tree algorithm to reconstruct geographic density as an extension of the concept of the negative survey. Lastly, Wieland et al. (2008) developed a linear programming model to address the risk of identification.

However, methods to mask the identification may introduce too much distortion, resulting in greatly diminished data utility for certain applications. A seminal work by Samarati and Sweeney (1998) defined k-anonymity when an individual cannot be identified from at least k-1 individuals in the individual level data. This method is a popular approach to introduce minimum generalization, which attempts to provide maximum utility while achieving a k-anonymity of a given threshold (Samarati and Sweeney 1998; Narayanan and Shmatikov 2006). The concept of using a target k-anonymity is compelling due to the release of high spatial resolution gridded population datasets in recent years, making an approach focusing on location parameters technically feasible on a wider scale. However, it is also important to understand the sources of gridded population data (Leyk et al. 2019) and examine uncertainty in the underlying population models (e.g. Hillson et al. 2014; Gaughan et al. 2019). Some constraints to standardizing this approach are that the ideal minimum size of k is still undefined, and it is a use-case-specific decision. In addition, there are conditions under which the measure of k-anonymity may not be sufficiently private (Machanavajjhala et al. 2007) as the approach does not necessarily incorporate other relevant parameters besides location.

Multiple studies have recently demonstrated that high dimensionality of data that is coupled with greater and quicker access to data presents a concern to previous practice of anonymization of the data (e.g. Hartter et al. 2013). For example, when the user can take advantage of the combination of variables such as gender, birth date, etc. it can be possible to reidentify individuals such as a list of voters (e.g. Sweeney 1997 in Samarati and Sweeney 1998). In mobile data, de Montjoye et al. (2013), found that four spatial-temporal points were sufficient to uniquely identify individuals and concluded that even coarse spatial resolution data provide a risk. Brownstein et al. (2006) also documented how easily one can reverse the identification of patients from low resolution disease maps.

In the following sections we grapple with this challenge, present new methods tested on multiple household survey datasets, and propose some tests for validation of spatial disclosure risk. In the final section we address other issues in the area of data provision.

III. Geomasking Methods

Spatial anonymization has dual objectives: to provide a geographic reference that enables users to integrate additional information from spatial datasets into a household survey and at the same time preserve confidentiality of place, preventing positive identification of the location of survey respondents. Geomasking, by altering or “blurring” coordinates, serves to conceal the actual location and, when mask parameters are revealed, also enables users to incorporate uncertainty into spatial variables derived using the anonymized locations. In this section we present a range of options for generating masked coordinates, starting with a method developed by the DHS Program and used in the dissemination of survey datasets since the early 2010s. Subsequently, the World Bank LSMS program applied this method as part of the dissemination of household survey data with masked coordinates. Referring to this as the “current” method we assess variations of and alternatives to the current method that may impart superior protections to respondents while preserving the data’s spatial integrity.

The section below is based on an in-depth review of methods that has been undertaken by the DHS program. The review seeks to strike a balance between respondent privacy and data availability, with the intent to produce a revised protocol for dissemination. It is informed in part by consultations with DHS users (e.g. researchers, academic institutions), and other groups in the household survey sphere, which revealed the consensus that the existing geomasking method overly displaces urban sample points[2], compromising analyses of DHS data. Additionally, these groups suggest that rural clusters may be inadequately displaced to properly protect survey respondents. The DHS Program explored several population-based alternatives to the current method. While these approaches hold the greatest promise for striking the correct balance between respondent confidentiality and preservation of spatial integrity, additional work will be necessary to identify population datasets that most accurately reflect the survey clusters visited. This, in turn, will allow for the best performance of the population displacement tools.

A. Urban/Rural displacement

Under the current method, clusters are displaced based upon their urban-rural status. Circular buffers are drawn surrounding each cluster, with a 2km radius for urban clusters and a 5km radius for rural clusters. Using an automated tool, clusters are displaced using a random angle and random distance within these buffers. An additional 1% of rural clusters are displaced up to 10km to further minimize disclosure risk (Burgert et al. 2013). This approach is also coded to ensure that these displacement buffers do not allow the point to be displaced out of the administrative level 2 area within which it is located. This method has been in use since the early 2010s and has become the standard for geomasking of household survey data.

The current displacement method has been adopted widely since its implementation and extensively studied and accounted for by analysts using DHS GPS data. The main advantage of this method is the simplicity with which the displacement occurs (built on a buffer-based approach). However, it has also

[2] Sample sites are interchangeably referred to in this document as sites, enumeration areas (EAs) or clusters

been surmised that the current displacement method overly displaces urban points – compromising the spatial relationships of these clusters. Conversely, rural points may be considered inadequately displaced from a confidentiality perspective. As it stands, the displacement of rural clusters does not necessarily make it more difficult to identify the original, undisplaced rural cluster, whereas a greater displacement distance may do so. This does mean, however, that the spatial relationships affecting rural clusters will be further compromised for analytic purposes.

To address some of the limitations of the current method, two parameter modifications have been evaluated, both based on buffer ranges. In the first modification, the urban displacement is reduced from 2 km to 1 km, while rural clusters are displaced up to 10 km and 1% of rural clusters being displaced up to 30 km. A second, more extreme modification is aimed directly at addressing the excessive displacement in urban and inadequate displacement in rural areas. In this implementation, the urban range is reduced to 0.5 km, while rural clusters are displaced up to 20 km and 1% of rural clusters displaced up to 30 km. While it has not been tested, it is worth noting that the parameter of percentage of rural clusters subjected to the maximum displacement range might also be modified. For instance, an increase from 1% to 10% could serve to encourage users to recognize the upper bound, without significantly increasing the average offset.

The main advantages of the parameter modifications are consistency with a well-documented and familiar method and use of existing tools. The modifications seek to address the two main concerns with the existing displacement: that urban points are overly displaced, compromising analysis, and rural clusters are inadequately displaced, potentially compromising respondent confidentiality. However, the negative impact of increased displacement on analysis of rural clusters remains unclear.

B. Population equation

An alternative to the existing displacement method, the population equation method is predicated upon the inverse relationship between the population at a given cluster and the distance the cluster must be displaced to adequately obfuscate it among its neighbors. As the population increases, the distance a cluster must be moved to reach a target population within which to hide decreases. A tool for implementation has been developed that measures the population at the undisplaced cluster, compares it to the threshold population value set by the user, and plots this value in the population equation to identify the maximum distance the cluster must move to achieve the threshold population. This maximum distance is used as the radius for a circular buffer drawn around the cluster, and the point is displaced at a random angle and random distance within the buffer.

An advantage of the population equation method is that it is dynamic due to the use of underlying population raster data. This means that each cluster can be displaced uniquely based on the population at and around the undisplaced cluster, ultimately doing away with the need for static “urban” and “rural” classifications for each cluster. A dynamic method helps reduce the risk of users reverse-engineering the displacement logic to identify undisplaced points. The method is also easy to modify by fitting custom equations that best capture the distribution of population between urban and rural clusters across surveys, and this information can in turn be used to “train” the model to increase its compatibility and applicability across surveys. However, this method will perform poorly if using an unoptimized equation to displace the GPS data. Further, it is reliant on the quality of the population raster dataset used in conjunction with the equation tool, another potential source of error, where the user must decide the best-fit.[3]

C. Adaptive masking by population buffer

The population buffer method also uses a threshold population value chosen by the user as the basis for the displacement of DHS cluster data. A 0.5 km buffer is drawn surrounding each cluster, and the population within the buffer is summed. If this value equals or exceeds the threshold, the cluster is displaced at a random angle and distance within the buffer. If the threshold is not equaled, the buffer is redrawn with a radius of 1 km, the population is re-estimated, and the process is repeated. This will continue with buffers of increasing radii until the threshold population is met. The sequential buffer generation may alternatively be implemented by a process of cumulative sum when adjacent pixels are sorted by distance to the cluster centerpoint, or seed location (adapted from R package gridsample). Once the cells required to meet the population threshold are identified, the maximum distance or final radius is the buffer within which the cluster is displaced using a random angle and distance.

As with the population equation, this approach is population-based and dynamic, meaning each cluster is displaced uniquely based on the cluster's population. However, the buffer tool also takes into account characteristics of the surrounding environment. An advantage of this is that it is built on the current method's basic framework of buffers being built around the undisplaced cluster, allowing for easier interpretation by users. Urban/rural classifications are discarded as population count is used instead: the measured population at each undisplaced cluster is compared to a threshold population value (k). For these analyses, the threshold *k* was selected following consultation with sampling and data quality experts at The DHS Program to identify a rough estimate of *k*. Transparency of mask parameters is up to the data provider. Withholding the exact value of "k" in public metadata will increase the difficulty of reverse engineering the method. However, without explicit knowledge of the displacement range the end user cannot properly account for uncertainty in derived spatial variables. The method is also computationally more intense and must be optimized to ensure it can be used with larger datasets. Much like the population equation, the buffer tool is reliant on – and limited by – the quality of the population raster dataset used in tandem with the buffer tool itself.

D. Summary of method results

These methods have been preliminarily assessed by the DHS Program for their ability to adequately displace clusters a minimum distance while ensuring respondent confidentiality. Currently, the population buffer tool was ranked highest performing against the current method (the existing urban/rural displacement method). This tool, on average, decreased the distance urban clusters were displaced and ensured each cluster, urban or rural, was displaced within an area encompassing the threshold k (target population), conferring adequate obfuscation of the respondent cluster.

The population equation tool performed worse than the buffer method, ranking last among the tested methods. The primary drawback of the tool was the unpredictable displacement that resulted from the tool's equation. Urban clusters would often be displaced further than they are under the existing displacement parameters, and rural clusters often displaced far less. Although uncommon, there are also instances where an extreme outlying data point can strongly skew the displacement of clusters

[3] see Leyk et al. 2019 for a review of global and continental population gridded datasets

within a dataset. However, this is likely due to the tool being trained with insufficient data, resulting in a sub-optimal equation that models the relationship between population and displacement distances.

The other displacement methods tested, including the two modifications to the existing displacement methodology, demonstrated mixed results. Both modifications successfully reduced the distance urban clusters are displaced while increasing the displacement of rural clusters. However, the selected parameters resulted in excessive displacement of rural clusters, to the extent that the population (k) value within which rural clusters were hidden were far above the threshold required. A summary of the methods tested, their parameters, and the successes and limitations of each is found in **Table 1**, below.

Table 1: Summary of DHS Displacement Method Testing

Displacement Method	Parameters	Pros	Cons
Urban/Rural (Current)	U: ≤ 2 km R: ≤ 5 km 1% of R: ≤ 10 km	• Operationalized by DHS and other groups for 11 years • Well documented & studied	• Excessive urban cluster displacement • Insufficient displacement of some rural clusters
Modification 1	U: ≤ 1 km R: ≤ 10 km 1% of R: ≤ 30 km	• Successful reduction of urban displacement • Successful increase of rural displacement	• Excessive rural cluster displacement • Excessive *k* values for rural clusters • Insufficient *k* values for urban clusters
Modification 2	U: ≤ 0.5 km R: ≤ 20 km 1% of R: ≤ 30 km	• Successful reduction of urban displacement • Successful increase of rural displacement	• Insufficient *k* values for urban clusters • Excessive *k* values for rural clusters • Excessive rural cluster displacement
Population Equation	1.593 + 824.209 * 1/x	• Achieves *k* anonymization for both urban and rural clusters	• Extremely excessive displacement for both urban and rural clusters • Greatly exceeds target *k* values for urban and rural clusters
Population Buffer	*k* = 5,000	• Achieves moderate reduction of displacement distance for urban clusters • Achieves moderate increase of displacement distance for rural clusters	• Displacement for rural clusters in one test was reduced, rather than increased • Computationally intensive • Could not successfully process large datasets

E. Other Considerations

It is important to consider the effect of location variables when planning for dissemination of any household survey. The displacement tools employed by the DHS Program, including the newly developed population-based methods, are constrained to the administrative 2 level unit of the country. That is, regardless of the displacement distance estimated by the tools, the cluster will not be displaced outside of the administrative 2 level unit from which it originates, which helps maintain the accuracy of the dataset. At the same time the actual size of the known zone, or target population, is reduced in cases where the zone overlaps an administrative boundary, by the process of elimination. The second administrative level was chosen for DHS surveys as it is, generally, the lowest level of administrative division across all DHS survey countries that can be provided to the user – in conjunction with geomasked GPS data – without compromising the privacy of the survey respondents. There is increasing demand from users to provide the most accurate geospatial data possible, including a standard data release at the administrative 3 level, but further investigation is required.

Exclusion in geomasking is a technique wherein an exclusion zone, or donut, is drawn around the undisplaced clusters within which the tool will *not* displace the cluster. Rather, the cluster is displaced in the area outside of the exclusion zone, thereby providing a minimum safe zone of obfuscation surrounding the respondents. While this method does theoretically confer protection to the respondents by way of this exclusion zone, it also reduces the overall area within which a cluster can be displaced within the administrative unit wherein the respondents are located. Ultimately, this reduces the effectiveness of the method in conferring protections to the survey respondents. Although a popular method explored in literature, this technique has not been incorporated into the population-based approaches tested by the DHS and World Bank teams.

IV. Quantifying risk of disclosure

Most household survey datasets include location variables, region or district or other place name, that define a base level of spatial disclosure risk. The provision of masked coordinates allows for spatial refinement, or reduction, of these areas, adding to this risk. This could act as a deterrent to the provision of masked coordinates or additional spatial attributes. However, by clearly defining anonymization objectives and making deliberate efforts to meet these objectives, the risk can be measured and therefore managed. In the following sections we illustrate how the concept of k-anonymity can be applied to the spatial dimension, how violations can be identified using different reference data sources and how the choice of reference data may impact results. We employ LSMS-ISA datasets anonymized using the Urban/Rural (current) and the Population Buffers method, as described in the previous section, to demonstrate these approaches. We then explore the uncertainty associated with different inputs and additional risk deriving from the depth of attribute information available in the spatial domain. We close with a brief discussion of data sensitivity and disclosure tolerance.

A. Spatial k-anonymity

We start with a naïve interpretation of spatial k-anonymity as a function of the characteristics of the zone of uncertainty, or anonymizing spatial region (ASR), within which the survey site is contained. For this analysis we make use of mask parameters, or displacement ranges, for both the fixed range and variable range adaptive method. We generate the cluster-level ASR using the fixed radius (Urban/Rural

Displacement) and two implementations of the Population Buffer method, with a high (10,000) and low (5,000) target population threshold using the WorldPop gridded population dataset.

Availability of spatial data to accurately assess the population count for an ASR is an undeniable challenge. Ideally this would be based on resident information for every building, but this level of information is rarely if ever available. Although aggregated, census Enumeration Area (EA) counts are accurate at the time of census. Survey listing results may be more current but would not fully cover the ASR. Some modeled population datasets are gridded at a finer spatial resolution, however pixel-level accuracy is affected by the administrative level of input population data, as well as modeling approach and other spatial inputs. For the purpose of this analysis we use enumeration areas as the authoritative source to measure k-Anonymity.

Acknowledging that digital EA boundaries are not always available; we make use of datasets that are becoming increasingly commonplace in the public domain to measure potential for reidentification. We compare results using three gridded population models: High Resolution Settlement Layer[4], WorldPop High Resolution Population[5], Global Human Settlement Population[6]. We also assess the suitability of feature datasets including building footprints[7], village locations[8], populated places[9] and small settlement areas[10]. In each case we benchmark the zonal statistics against the authoritative source, digital EAs, in terms of identifying violations of a k-anonymity threshold.

1. Household re-identification risk

To estimate the probability of re-identification of a household within the explicit zone of uncertainty, we use the spatial intersection to derive attributes of population and building count. The total population within this zone, divided by average household size, represents the minimum risk as it does not account for other known household attributes. For reference datasets we choose thresholds based on averages representative of the area of interest. In this case a threshold of 5,000 population represents approximately 1,250 households. For building footprints, we assume that 60% of building stock is residential[11], so a threshold of 2,000 total buildings represents a ratio of almost one building per household. These thresholds are somewhat arbitrary and would require thorough investigation for actual implementation.

For demonstration we use an EA count of less than 5 as the definition of a violation of the anonymization objective, where the zone of uncertainty provides insufficient protection. Results presented in Table 2 show that violations are more common in urban areas, across all methods. This finding is counter to expectations and may be an indication that use of a lower EA count would be acceptable in urban areas, given the high density of population. Also notable is that a large portion of

[4] High Resolution Settlement Layer. Facebook Connectivity Lab and Center for International Earth Science Information Network - CIESIN - Columbia University. 2016

[5] WorldPop spatial distribution of population in 2015 (DOI: 10.5258/SOTON/WP00645), November 2018

[6] GHSL Global Population Grids, R2015A. JRC and CIESIN. October 2015

[7] Building footprints from DigitizeAfrica, a Bill and Melinda Gates Foundation and Government of Canada funded effort implemented by Maxar and Ecopia. September 2017

[8] Village locations downloaded from ICRAF landscapeportal.org, incomplete metadata

[9] GRID3 Settlement Extents Version 01, Alpha accessed from data.humdata.org 12/22/2020

[10] NGA GEOnet Names Server (GNS). Populated places accessed from https://geonames.nga.mil 1/3/2021

[11] Residential building stock estimation based on number of features in the building footprint dataset and number of housing units reported by Habitat for Humanity (https://www.habitat.org/)

the dataset produced using low-threshold Population Buffer method does not meet k-5 Anonymization criteria. However, this could be mitigated by withholding cluster-level mask parameters. The other methods perform similarly to each other, with a slight advantage for Urban/Rural in urban settings and for high-threshold Population Buffer in rural.

Table 2 Violations of k5-Anonymity and Potential for Household Re-identification

		Violations	Percent of violations identified by proxy datasets					
	Strata	EA count < 5	HRSL < 5000	WorldPop < 5000	GHSPOP < 5000	Buildings < 2000	Combined (any)	Concordance (2 or more)
Urban/Rural	rural	13%	50%	60%	54%	27%	62%	54%
	urban	25%	57%	75%	68%	39%	79%	71%
Population Buffer (5,000)	rural	61%	41%	64%	44%	22%	77%	54%
	urban	85%	44%	52%	38%	69%	86%	56%
Population Buffer (10,000)	rural	9%	7%	7%	7%	11%	17%	7%
	urban	37%	17%	24%	15%	32%	46%	20%

We further explore how many violations of k5-anonymity identified by EA count are also identified by proxy datasets of population and building footprints using the thresholds described above. Findings are that proxy datasets are least successful overall with the high-threshold Population Buffer method and that for the other methods WorldPop performs best across datasets. Combining all proxy datasets results in some gains, however the gains come with many false positives (not shown) that would undermine the effectiveness of this approach in the absence of EA counts. If datasets are given equal weight, then a more reliable approach might be to look at agreement of at least 2 proxy datasets (presented in the last column), where a maximum of 54% of the violations are identified in rural areas and 71% in urban areas.

The data-driven anonymization methods (Population Buffer) should have an advantage with respect to population summary statistics in that the threshold is encoded in the process when an explicit population target is specified. Nevertheless, there are occurrences in Table 2 of zones of uncertainty that do *not* meet the target thresholds. This is an indication that while the target population is met around the actual location, this property does not extend to a zone of uncertainty derived around the anonymized location.

2. Community re-identification risk

Re-identification may also occur at the neighborhood, community or town level. To illustrate, we again use violations of k5-Anonymity defined by EA count, and assess how well these critical cases are captured in proxy datasets representing community features. Results presented in Table 3 show that in most cases proxy datasets appear to perform better in urban areas. However, this is a misleading conclusion, stemming from the fact that the absence of features is the determinant of success (although we do exclude zero counts). In fact, it is an indicator of incompleteness of the village and populated place datasets in the urban context: urban neighborhoods are not well captured.

.

Table 3 Violations of k5-Anonymity and Potential for Community Re-identification

		Violations	Percent of violations identified by proxy datasets				
	Strata	EA count < 5	Village count < 5	Populated places < 5	Small settlement areas < 5	Combined (any)	Concordance (2 or more)
Urban/Rural	rural	13%	17%	35%	17%	53%	12%
	urban	25%	43%	64%	NA	79%	57%
Population Buffer (5,000)	rural	61%	15%	46%	29%	63%	23%
	urban	85%	62%	17%	NA	73%	11%
Population Buffer (10,000)	rural	9%	4%	48%	28%	59%	19%
	urban	37%	46%	29%	NA	61%	15%

Furthermore, the large difference between combination and concordance measures across the datasets indicates a lack of agreement which does not promote confidence. As such, we can only conclude that the datasets used in this analysis would not support accurate estimation of community-level re-identification risk, particularly in the urban context, and there is a need for additional data exploration.

3. Actual k-Anonymity

Two additional factors affecting risk of disclosure are the use of constraining boundaries in the displacement process and the degree of overlap between zones of uncertainty. Note that in previous sections we have consistently used data that is constrained by known administrative units, and that should be standard procedure. However, to gain some understanding of the effect of constraining by administrative unit we present a comparison of summary statistics in Table 4.

Table 4 Comparison of Methods and Effect of Masking by Administrative Unit

			Unmasked		Masked by known administrative unit	
	Strata	Mean displacement (km)	Mean EA count	Mean pop WorldPop	Mean EA count	Mean pop WorldPop
Urban/Rural	rural	2.4	16	19,344	12	13,451
	urban	1.0	26	48,220	11	20,471
Population Buffer (5,000)	rural	1.5	7	8,238	4	5,050
	urban	0.4	4	6,239	3	5,033
Population Buffer (10,000)	rural	2.1	13	15,829	9	10,085
	urban	0.5	9	16,066	6	8,838

The largest mean displacement in high density areas (Urban/Rural method for urban) results in significantly higher reduction in anonymization metrics: mean zonal population is reduced by nearly 60 percent, from 48,220 to 20,471. This category also exhibits potentially excessive displacement with respect to other methods, with an average displacement twice that of the high Buffer method.

Finally, the issue of overlap between zones of uncertainty affects actual k-anonymity. We find this to be a significant factor, as evidenced by the summary shown in Table 5. Not unexpectedly, the overlap is closely related to zone size within stratum and is most acute within large urban zones.

Table 5 Extent of overlap between zones of uncertainty

	Strata	Mean displacement (km)	Percent overlap
Urban/Rural	rural	2.4	25%
	urban	1.0	46%
Population Buffer (5,000)	rural	1.5	12%
	urban	0.4	5%
Population Buffer (10,000)	rural	2.1	25%
	urban	0.5	17%

B. Uncertainty in modeled population data

Accurate and complete spatial datasets are necessary for understanding spatial risk of disclosure. Whether they are used to define or describe ASRs, the reliability of input datasets determines the level of confidence in the product of anonymization. Uncertainty itself may confer some protection in the short-term. However, given the current trend of increasing data availability and tools for integration, this protection is likely to wane over time with improvements in data and modeling.

In an effort to understand the differences between population datasets, we use ASRs generated using the fixed radius by stratum method to derive zonal population sums for each dataset in two country contexts. Table 6 shows general agreement of basic summary statistics, with the exception of GHSPOP in country B.

Table 6 Cluster level summary statistics from gridded population datasets

		HRSL		WorldPop		GHSPOP	
	Strata	mean	stdev	mean	stdev	mean	stdev
Country A	rural	14,637	8,687	13,451	7,904	14,667	8,740
	urban	22,571	19,600	20,471	18,349	22,515	19,739
Country B	rural	24,636	26,958	22,590	31,955	37,280	66,233
	urban	66,197	75,651	65,644	70,245	112,991	90,108

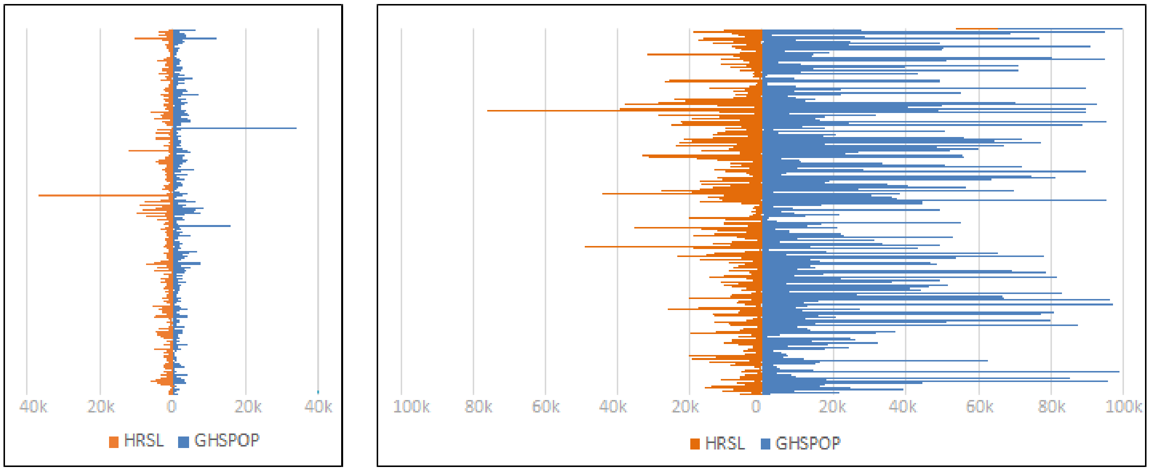
Table 7 provides further evidence that GHSPOP is potentially problematic in the rural context of country B. The number of missed clusters, defined as zonal total population less than 100 people, is high, indicating that small rural communities are not well captured in the dataset for country B.

Table 7 Number of missed clusters (population < 100)

	Strata	HRSL	WorldPop	GHSPOP
country A	rural	4	4	2
	urban	0	0	0
country B	rural	4	1	128
	urban	0	0	3

Looking more closely at cluster level population sums across the datasets we find striking differences between the two countries. Figure 2 shows deviation of cluster level population from HRSL and GHSPOP, from WorldPop (represented by zero vertical), using the same scale for both countries. While there are outliers in both graphs, the magnitude of differences is far greater in country B, shown on the right.

Figure 2 Cluster level difference from WorldPop for country A (left) and country B (right)



We hypothesize that one factor driving the differences is the administrative level of input population data. The number of input administrative units is more than an order of magnitude greater for country A than country B[12]. Differences are also more extreme for small extraction zones, when grids are sampled over very limited areas. Zonal statistics at the cluster level serve to highlight uncertainties in the modeled output, with implications for the suitability of the datasets in different contexts for applications such as adaptive masking and assessment of risk of disclosure.

C. Second order risk

Place names provide linkages with a profusion of data points in the public domain (Facebook advertising data and public posts, news sources, Twitter, and Google) or privately held repositories. Natural language processing algorithms and AI are powerful tools for data mining, allowing users to extract relevant and potentially identifying information from massive amounts of text data. Even unstructured location data provides linkages, as evidenced by the pooling of data by location known as geofencing. Deeper exploration of the implication for reidentification deriving from these linkages is beyond the

[12] Center for International Earth Science Information Network - CIESIN - Columbia University. 2016. Documentation for the Gridded Population of the World, Version 4 (GPWv4). Palisades NY: NASA Socioeconomic Data and Applications Center (SEDAC). http://dx.doi.org/10.7927/H4D50JX4.

scope of this guideline. However, the methods described above provide some idea of the potential narrowing of the field, through auxiliary datasets representing towns, localities and other place descriptors.

Additionally, there may be pseudo-spatial attributes in microdata that can be geocoded with minimal effort. These include both record identifiers that include official administrative codes and responses that make direct reference to spatial features (distance to main/tarred road, distance to health clinic or secondary school, etc), among others. Although often imprecise or generalized, these responses are highly valuable in the absence of masked coordinates. However, they can significantly increase disclosure risk when combined with an explicitly spatial zone of uncertainty around a masked location. An initial screening of microdata for such pseudo-spatial information, and treatment prior to dissemination will help manage this risk.

D. Tolerance

Any discussion of risk of disclosure is incomplete without considering the consequences of re-identification associated with a particular survey. The potential consequences are wide ranging and highly contextual. For example, in longitudinal surveys, panel interference is a concern in the case of reidentification at EA level. Alternatively, at the individual level respondents may be targeted for advertising, or more directly harmful actions based on specific characteristics. The control of personal information may be lost.

This guideline stops short of making explicit recommendations regarding risk thresholds. This is a judgement specific to each survey, although there are some guardrails common to all: respondent notification and consent, internal governance structures. Many providers of data have an internal advisory board which would dictate protocols. However, in the absence of such an authority, data providers may look to other areas of applied research for guidance. The ethical matrix originally presented as a conceptual tool designed to help in making decisions about the ethical acceptability of different technologies applied in the fields of food and agriculture (Mepham et al. 2006) has lately been promoted as a framework for assessing the pros and cons of powerful algorithms that affect the daily lives of individuals (O'Neil, 2016). This framework could provide a useful tool for assessing potential for harm associated with reidentification from household survey in greater depth.

Several initiatives have been launched more recently, specifically aimed at promoting ethical use of geographic data. This is in part due to the increased use of location data for tracking and monitoring individuals, as well as research, during the global covid-19 pandemic. One example is an effort by the Dutch government foundation Geonovum to develop an ethical framework to guide the collection and use of location data. Similarly, the EthicalGEO Initiative is supporting adoption of an international charter, the Locus Charter, on the ethical use of geodata. These resources may be helpful in design of dissemination strategy for household survey geodata.

V. Data utility trade-off

The use of anonymized locations comes at some analytical cost. The magnitude of this cost is currently not known across a wide range of data applications but can be estimated as a function of the spatial resolution and smoothness of geospatial variables researchers would be seeking to integrate with the survey data, as well as the contextual scale of relevant spatial information. This is an area of ongoing research for both DHS and LSMS-ISA programs. We discuss work that serves to inform both users, on data limitations, as well as providers, in the preparation of data for dissemination.

In the guidance document for users of DHS GPS data three common location-based research applications, each illustrated using case studies, are investigated to demonstrate the effect of anonymization on analysis. In the first study, the effect of anonymization on distance-based measures and closest facility allocation is shown to result in some bias and is related to spatial density of target features. The second study assesses the impact of anonymization on integration of spatial variables from raster sources and finds that spatial autocorrelation of source data is a factor in effect bias. The final study evaluates the results of attribute information derived using areal overlays, presents a method for quantifying the probability of misspecification and proposes the use of weighted covariates to reduce the effect of errors. It should be noted that these studies assume transparency of geomasking parameters, or explicit range of uncertainty.

As noted above, the effect of displacement on data for research can be captured in part by comparing key spatial statistics derived using raw and anonymized locations. However recent research on the utility of displaced geographic data for health research (Broen et al. 2021) reinforces the notion that there is a clear trade-off. In an assessment of multiple anonymization techniques, methods that produced the least impact on spatial metrics were also the most vulnerable to de-identification. Although the displacement ranges used in this work are smaller than is currently used in dissemination of LSMS-ISA and DHS survey data, their findings with regard to effectiveness of spatial anonymization are cautionary. Assessment of risk is necessary and iterative approaches may be helpful in mitigating the risk.

Forthcoming research from the World Bank examines the relationship between anonymization techniques and the quality of machine learning (ML) approaches to mapping economic well-being. The experiment is grounded in consumption-expenditure data surveyed in Ethiopia and Malawi, georeferenced using actual EA locations, and three anonymization methods. A convolutional neural network (CNN) is built to generate estimates of wealth at each location using satellite imagery from Sentinel-2 and VIIRS. As part of our work, we additionally document whether the impact of using confidential versus spatially anonymized varies by (i) the spatial resolution of the satellite imagery – specifically using Sentinel-2 (10-m resolution) versus Landsat (30-m resolution), and (ii) the size of tiles (i.e. the geographic extent) that the satellite imagery is parsed into for a convolutional neural network to isolate spatial features that are used as poverty predictors in the model. Preliminary findings are that the effect of anonymization can be largely mitigated relative to the non-anonymized EA locations by employing a larger geographic extent as input to the model as well as by leveraging inputs that capture critical information but at a coarser spatial resolution, such as VIIRS. The implication is that reducing displacement distance would not necessarily improve the predictive power of comparable models.

Other forthcoming research (van der Weide et al. mimeo) compares the performance of an approach that estimates poverty through a geographically weighted regression model that uses: (a) public use coordinates (Burgert et al. 2013) versus (b) precise coordinates provided by the Malawi National Statistical Office that are not publicly available. The authors find that the use of public coordinates reduces the in-sample goodness-of-fit of the model, as is to be expected, although the reduction in adjusted R-square is reasonably modest. The regression models based on public versus private coordinates provide qualitatively similar results; all regression coefficients remain significant and are of the same order of magnitude. Furthermore, the estimates of these two models of poverty are highly correlated with a gold-standard benchmark estimate of poverty from the method developed by Elbers, Lanjouw and Lanjouw (2003).

There is a clear need for further research to gain a deeper understanding of the impact of spatial anonymization on data utility in a variety of applications. Nevertheless, there are some general lessons data providers may glean from existing evidence: reduced displacement is not a guarantee of improved data utility, mask parameters have been a valuable reference for users in the assessment of potential error due to anonymization. Furthermore, understanding the trade-off between reduced accuracy and the impact on analysis can inform decisions on dissemination. Incurring additional risk for little gain in analytical value could be avoided, as well producing data that are not fit for purpose due to excessive displacement. Lastly it is worth noting that at least in select examples of poverty estimation described here, use of anonymized coordinates does not unduly diminish model performance.

VI. More options for supporting priority/cutting-edge research

The provision of explicit spatial references, for example anonymized coordinates, is the most direct method for enabling users to integrate additional spatial variables of interest. However, there are a range of other options for enabling research through dissemination, services, remote access, on-site access.

A. Dissemination of spatial variables

The dissemination of record-level spatial variables with survey data can be a means of providing more spatial detail, when unmodified coordinates are used, or simply promoting ease of access and standardization, when generalized coordinates, ranges or polygons are used.

IPUMs and the DHS program have collaborated on a set of spatial variables described as “contextual variables”, which characterize the general landscape surrounding survey clusters. The DHS Program also provides geospatial covariate datasets, linking survey cluster locations to ancillary data - known as covariates – that contain data on topics including population, climate, and environmental factors. In a similar vein, all waves of LSMS-ISA data are accompanied by a set of spatial variables, mostly related to agriculture and household welfare.

These datasets are typically generated using anonymized location data and do not add to the existing risk of disclosure. However, they may not accurately capture characteristics of sample sites, particularly with high spatial resolution or surfaces with high local variation. Some variable-level tests would improve utility and reliability of the datasets. The DHS Program is planning to investigate the effects of the existing displacement method on the accuracy of the geospatial covariate data produced by DHS. Similar efforts should be done to measure the impact of the population-based displacement methods on geospatial covariate accuracy.

While the approaches described above can add great value to survey datasets, they do not likely meet research needs for detailed spatial information. Spatial variables that are generated using more precise location data than anonymized location (cluster centerpoint or household location) will add to the risk of disclosure. Untreated variables produced in this way using publicly available reference datasets create a spatial signature. Depending on the number of variables and characteristics of the reference datasets, the spatial signature can dramatically refine the zone of uncertainty for sample points. Some approaches to managing risk of disclosure include reducing dimensionality (number of spatial variables) by producing a core set of the most relevant variables for analysis (i.e. accessibility, rainfall, vegetation indices, terrain for agricultural surveys). In addition, inputs datasets with high value depth (number of values) or high spatial resolution should be managed by rounding or ranging results. Each of these measures necessarily reduces the specificity of variables, and there is strong evidence that it is not possible to produce a dataset suitable for public dissemination without undermining the primary purpose. Although the terms of use likely preclude re-engineering of location, it is advisable to restrict access to any datasets that present an increased risk of disclosure.

B. Interpolated surfaces / small area estimation

Since 2016, The DHS Program has been producing interpolated modeled surfaces for a suite of health and demographic indicators, providing decision-makers with high spatial resolution and small area estimates for relevant issues that are central to meeting the sustainable development goals (SDGs). These surfaces are freely available on the Spatial Data Repository. These methods are cutting-edge solutions to the questions of inadequate funding to increase survey sample size, allowing decision makers to get the high-quality health and demographic information needed to successfully address the pressing issues their countries face in resource-constrained environments.

The DHS program has been collecting various geospatial covariate layers that are used for spatial analysis. These layers are obtained from a myriad of sources, and hence therefore have different spatial reference, projections, extents and dimensions. Thin plate smoothing spline and bilinear interpolation methods are used to downscale and resample the geospatial covariates to the same spatial resolution used in the modeling process. It should be noted that interpolation is not suitable for all types of variables. A summary of key characteristics to be considered is outlined in the DHS Spatial Analysis Report 9.

Another approach to downscaling is the use of small estimation techniques. Statistical models leverage the spatial relationships between survey locations and spatial covariates in order to produce reliable estimates of key variables at a level below representative unit (e.g. Mayala et al. 2019).

C. Restricted access

There are limits to the effectiveness of spatial anonymization by design. Even after treatment, location data may not be sufficiently anonymized, or required modifications render a data product unsuitable for its intended purpose. In this case access to confidential data may be the only option to support some research applications. Access can be provided through a secure physical or virtual enclave, where approved users interact with the data but cannot make copies or extract source data, or by way of analytical platforms.

Some National Statistical Agencies and organizations do provide this type of access. The US Census Bureau, for example, enables access to confidential data through a system of Federal Statistical Research Data Centers. Likewise, the UK Data Service provides controlled access to potentially disclosive data through both physical and virtual enclaves. The South African DataFirst initiative provides access to restricted data through a secure physical enclave at the University of Cape Town.

These examples all benefit from an economy of scale and, as national institutions, have a common frame of reference with respect to national statistical laws and security standards. Nevertheless, considerable resources are required to build and sustain the effort. They must support the maintenance of secure physical or digital infrastructure, with proper safeguards. Dedicated staffing is required to engage with data depositors in defining level of access and standards for disclosure control, to review applications and to evaluate results derived within the enclave environment. Furthermore, the hosting agency is responsible for ensuring compliance with confidentiality agreements.

Unfortunately, many National Statistical Agencies, particularly in low-income countries, lack the resources to maintain these services. If legal and regulatory frameworks allow, one option to consider is external hosting of the restricted data. While such arrangements are not common, one example is the platform being developed for the Millennium Challenge Corporation (MCC) by the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan. The goal is to expand access to microdata, including restricted data through a virtual data enclave, collected in multiple countries where MCC works and conducts evaluation projects.

In lieu of providing direct access to restricted data, there is the idea of bringing research to the data through online services. The Open Algorithms (OPAL) Project maintains data on-site but allows certified algorithms to be applied to the data resource. Lastly, data providers may choose to provide analytical services, wherein confidential location data is used to run analyses in response to specific requests that have a demonstrated need for higher locational accuracy. Though only the results are shared, as with spatial variables, results must be assessed for disclosure risk and treatment applied as needed.

VII. Recommendations for discussion

This section pulls together evidence from preceding sections. Although not prescriptive we attempt to identify best practices and highlight areas for further research.

A. Current best practices

1. Tolerance for disclosure risk should be based on sensitivity of microdata, as well as respondent consent and organizational practices. Given the availability of tools for automating the displacement procedures it is feasible to test different approaches and assess their utility with respect to a particular dataset and anonymization objective.

2. With increasing focus on privacy and data breaches, it is important to implement due diligence to verify that anonymization objectives are achieved. Checks should be implemented to measure risk of disclosure, aimed at identifying both excessive and inadequate displacement.

3. EA boundaries or centerpoints are optimal reference datasets for assessing risk of disclosure. Alternative datasets are not always suitable substitutes. Some level of data curation, for completeness and accuracy, should be applied to all reference datasets.

4. Flexibility in treatment of outliers may be required. If a chosen method of anonymization is inadequate in a small subset of sparsely populated regions, use of aggregation or administrative level reference may be an option.

5. Assessment of spatial risk of disclosure should be cumulative, with respect to all spatial and pseudo-spatial attributes in a dataset, as well as derived outputs. In other words, the total risk of disclosure should be based on the spatial intersection of all relevant information.

6. Efforts to ensure spatial anonymity will be strengthened by disclosure control (reducing sample uniqueness) applied to linked non-spatial microdata. In some cases, data providers must also consider suppression of sensitive microdata.

7. Standardization of procedures promotes transparency, ease of use, tracking and compliance:
   - standardized data download page and request protocol similar to that currently employed by DHS
   - standard level of disclosure, similar to that currently employed by programs such as DHS
   - standardized "template" for displacement (methodological options) allowing data providers to modify as needed

B. Future work

1. Transparency in mask parameters (what information to provide to the end-user) is a critical issue. Withholding mask parameters is a powerful way for data providers to ensure anonymity. However, withholding this information undermines the ability of users to incorporate discrete measures of spatial uncertainty. Further investigation into the impacts of withholding this information is required.

2. Replication would build confidence in important findings and help hone messaging. Future work should include the training of these tools using the full suite of DHS and LSMS-ISA GPS datasets, to ensure the methods work consistently over time within the same survey regions, and across regions.

3. There is a trend toward increased volume and specificity of individual-level data in the public domain (Facebook Advertising, etc.) as well as democratization of data analytics: open source tools and lower-cost computing resources (Google Earth Engine, cloud services). More research is needed to understand how these developments will affect disclosure risk.

4. The interaction of non-spatial microdata attributes and spatial information needs to be investigated.

5. A better understanding of the data utility trade-off associated with spatial anonymization will help guide data providers in defining dissemination protocols.

6. Current findings with respect to adaptive approaches are driven by inaccuracies in the input gridded population datasets. Improvements in quality of inputs would potentially make these approaches more feasible. They should be reassessed with advancements in modeling methods and increased availability of data.

References

Alegana, V. A., P. M. Atkinson, C. Pezzulo, A. Sorichetta, D. Weiss, T. Bird, E. Erbach-Schoenberg, and A. J. Tatem. 2015. “Fine Resolution Mapping of Population Age-structures for Health and Development Applications.” *Journal of the Royal Society Interface* 12 (105): 20150073. https://doi.org/10.1098/rsif.201

Boulos, M. N. K., Curtis, A. J., & AbdelMalik, P. (2009). Musings on privacy issues in health research involving disaggregate geographic data about individuals. *International Journal of Health Geographics*, *8*, 46.

Brown, M. E., Grace, K., Shively, G., Johnson, K. B., & Carroll, M. (2014). Using satellite remote sensing and household survey data to assess human health and nutrition response to environmental change. *Population and environment*, *36*(1), 48-72.

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

Broen, K., Trangucci, R., Zelner, J. (2021). Measuring the impact of spatial perturbations on the relationship between data privacy and validity of descriptive statistics. *Int J Health Geographics*, 20:3

Buckee, C. O., Wesolowski, A., Eagle, N. N., Hansena, E. & Snow, R. W. Mobile phones and malaria: Modeling human and parasite travel. Travel. Med. Infect. Dis 11, 15–22 (2013)

Burgert, C. R., Colston, J., Roy, T., & Zachary, B. (2013). Geographic displacement procedure and georeferenced data release policy for the Demographic and Health Surveys.

Burgert-Brucker, C. R., Dontamsetti, T., & Gething, P. W. (2018). The DHS program's modeled surfaces spatial datasets. *Studies in family planning*, *49*(1), 87-92.

de Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, *3*, 1376.

DHS Spatial Interpolation Working Group. 2014. Spatial Interpolation with Demographic and Health Survey Data: Key Considerations. DHS Spatial Analysis Reports No. 9. Rockville, Maryland, USA: ICF International.

Duncan, G., & Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business & Economic Statistics*, *7*(2), 207-217.

Duncan, G. T., Elliot, M., & Salazar-González, J. J. (2011). Why Statistical Confidentiality?. In *Statistical Confidentiality* (pp. 1-26). Springer, New York, NY.

Dupriez, O., & Boyko, E. (2010). *Dissemination of Microdata Files; Principles, Procedures and Practices.* International Household Survey Network (IHSN).

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. Econometrica, 71(1), 355-364.

El Emam, K., Jonker, E., Arbuckle, L., & Malin, B. (2011). A systematic review of re-identification attacks on health data. *PloS one*, *6*(12), e28071.

Fronterrè, C., Giorgi, E., & Diggle, P. (2018). Geostatistical inference in the presence of geomasking: a composite-likelihood approach. *Spatial Statistics*, *28*, 319-330.

Gaughan, A. E., Oda, T., Sorichetta, A., Stevens, F. R., Bondarenko, M., Bun, R., Krauser, L., Yetman, G., & Nghiem, S. V. (2019). Evaluating nighttime lights and population distribution as proxies for mapping anthropogenic CO2 emission in Vietnam, Cambodia and Laos. Environmental Research Communications, 1(9), 091006.

Grace, K., Nagle, N. N., Burgert-Brucker, C. R., Rutzick, S., Van Riper, D. C., Dontamsetti, T., & Croft, T. (2019). Integrating Environmental Context into DHS Analysis While Protecting Participant Confidentiality: A New Remote Sensing Method. *Population and development review*, *45*(1), 197.

Gutmann, M. P., Witkowski, K., Colyer, C., O'Rourke, J. M., & McNally, J. (2008). Providing spatial data for secondary analysis: Issues and current practices relating to confidentiality. *Population research and policy review*, *27*(6), 639-665.

Han, Y. S., de Wolf, P. P., & de Jonge, E. (2019). Comparing methods of safely plotting variables on a map.

Hartter, J., Ryan, S. J., MacKenzie, C. A., Parker, J. N., & Strasser, C. A. (2013). Spatially explicit data: stewardship and ethical challenges in science. *PLoS Biology*, *11*(9), e1001634.

Hermes, K., & Poulsen, M. (2012). A review of current methods to generate synthetic spatial microdata using reweighting and future directions. *Computers, Environment and Urban Systems*, *36*(4), 281-290.

Hillson, R., Alejandre, J. D., Jacobsen, K. H., Ansumana, R., Bockarie, A. S., Bangura, U., ... & Stenger, D. A. (2014). Methods for determining the uncertainty of population estimates derived from satellite imagery and limited survey data: a case study of Bo City, Sierra Leone. *PloS one*, *9*(11), e112241.

Horey, J., Forrest, S., & Groat, M. (2012, April). Reconstructing spatial distributions from anonymized locations. In *2012 IEEE 28th International Conference on Data Engineering Workshops* (pp. 243-250). IEEE.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790-794.

Leyk, S., Gaughan, A. E., Adamo, S. B., de Sherbinin, A., Balk, D., Freire, S., ... & Comenetz, J. (2019). The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use. *Earth System Science Data*, *11*(3).

Loonis V., Bellefon M.-P.(dir.) (2018). Manuel d'analyse spatiale. Théorie et mise en œuvre pratique avec R, Insee Méthodes n° 131, Insee, Eurostat, 392 p.

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2007, March). l-diversity: Privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data 1, 1, Article 3, 52 pages. DOI=10.1145/1217299.1217302 http://doi.acm.org/10.1145/1217299.1217302

Machanavajjhala A., D. Martin, D. Kifer, J. Gehrke, and J. Halpern. Worst case background knowledge. In *ICDE*, 2007.

Mayala, B.K., T. Dontamsetti, T.D. Fish, and T.N. Croft. 2019*. Interpolation of DHS Survey Data at Subnational Administrative Level 2*. DHS Spatial Analysis Reports 17. Rockville, MD, USA: ICF International.

https://dhsprogram.com/publications/publication-sar17-spatial-analysis-reports.cfm

MEASURE GIS Working Group 2008.

Mennis, J., & Guo, D. (2009). Spatial data mining and geographic knowledge discovery—An introduction. *Computers, Environment and Urban Systems*, *33*(6), 403-408.

Mepham, B. & Kaiser, Matthias & Thorstensen, Erik & Tomkins, S. & Millar, K.. (2006). Ethical Matrix Manual. Agricultural and Forest Meteorology - AGR FOREST METEOROL.

National Research Council 2007. *Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data*. Washington, DC: The National Academies Press. https://doi.org/10.17226/11865.

Narayanan, A., & Shmatikov, V. (2006). How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105*.

Øhrn, A., & Ohno-Machado, L. (1999). Using Boolean reasoning to anonymize databases. *Artificial Intelligence in Medicine*, *15*(3), 235-254.

O'Neil, Cathy. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. First edition. New York: Crown, 2016.

Primault, V., Mokhtar, S. B., Lauradoux, C., & Brunie, L. (2015, August). Time distortion anonymization for the publication of mobility data with high utility. In *2015 IEEE Trustcom/BigDataSE/ISPA* (Vol. 1, pp. 539-546). IEEE.

Perez-Heydrich, C., Warren, J. L., Burgert, C. R., & Emch, M. (2013). *Guidelines on the use of DHS GPS data*. ICF International.

Richardson, D. B., Kwan, M. P., Alter, G., & McKendry, J. E. (2015). Replication of scientific research: addressing geoprivacy, confidentiality, and data sharing challenges in geospatial research. *Annals of GIS*, *21*(2), 101-110.

Runfola, D., Marty, R., Goodman, S., Lefew, M., & BenYishay, A. (2017). geoSIMEX: A generalized approach to modeling spatial imprecision. *http://aiddata. org/aiddata-working-paper-series*.

Samarati, P., & Sweeney, L. (1998). *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression* (p. 19). technical report, SRI International.

Seidl, D. E., Jankowski, P., & Tsou, M. H. (2016). Privacy and spatial pattern preservation in masked GPS trajectory data. *International Journal of Geographical Information Science*, *30*(4), 785-800.

Sweeney, L. (1997). Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics*, *25*(2-3), 98-110.

United Nations General Assembly 2015. Transforming our World: The 2030 Agenda for Sustainable Development. New York, New York, USA: United Nations Population Fund. https://www.unfpa.org/resources/transforming-our-world-2030-agenda-sustainable-development.

van der Weide, R., B. Blankespoor, C. Elbers, and P. Lanjouw. Can a Poverty Map Based on Remote Sensing Data Replicate One Based on Census Data? An Assessment for Malawi. Mimeo.

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

Wartenberg, D., & Thompson, W. D. (2010). Privacy versus public health: the impact of current confidentiality rules. *American journal of public health*, *100*(3), 407-412.

Wegmann, M., Leutner, B. & Dech, S. 2016. Remote Sensing and GIS for Ecologists: Using Open Source Software. Pelagic Publishing. Exeter, UK.

Zandbergen, P. A. (2014). Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. *Advances in medicine*, *2014*.