Population and Environment (2021) 42:445–476
https://doi.org/10.1007/s11111-020-00372-4

ORIGINAL PAPER

Working toward effective anonymization for surveillance data: innovation at South Africa's Agincourt Health and Socio-Demographic Surveillance Site

Lori M. Hunter[1,2] · Catherine Talbot[1,2] · Wayne Twine[3] · Joe McGlinchy[5] · Chodziwadziwa W. Kabudula[4] · Daniel Ohene-Kwofie[4]

Accepted: 29 December 2020 / Published online: 24 March 2021

Abstract

Linking people and places is essential for population-health-environment research. Yet, this data integration requires geographic coding such that information reflecting individuals or households can appropriately be connected with characteristics of their proximate environments. However, offering access to such geocoding greatly increases the risk of respondent identification and, therefore, holds the potential to breach confidentiality. In response, a variety of "geographic masking" techniques have been developed to introduce error into geographic coding and thereby reduce the likelihood of identification. We report findings from analyses of the error introduced by several masking techniques applied to data from the Agincourt Health and Socio-Demographic Surveillance System in rural South Africa. Using a vegetation index (Normalized Difference Vegetation Index (NDVI)) at the household scale, comparisons are made between the "true" NDVI values and those calculated after masking. We also examine the tradeoffs between accuracy and protecting respondent privacy. The exploration suggests that in this study setting and for NDVI, geomasking approaches that use buffers and account for population density produce the most accurate results. However, the exploration also clearly demonstrates the tradeoff between accuracy and privacy, with more accuracy resulting in a higher level of potential respondent identification. It is important to note that these analyses illustrate a process that should characterize spatially informed research but within which particular decisions must be shaped by the research setting and objectives. In the long run, we aim to provide insight into masking's potential and perils to facilitate population-environment-health research.

Keywords Anonymity · Confidentiality · Geomasking · Geographic Masking · Jittering · Spatial Error · Agincourt · South Africa

✉ Lori M. Hunter
Lori.Hunter@colorado.edu

Extended author information available on the last page of the article

Introduction

Vulnerability to environmental change and its implications for human and ecological well-being remain critical challenges within global development. Research on the myriad dimensions of vulnerability has grown rapidly over the past decade, and while vulnerability as a concept has been usefully theorized, understanding the patterns and implications of differential vulnerability also requires accessible data. However, a critical challenge in vulnerability scholarship is that research linking people and place requires knowledge of individuals' or households' geographic location. Such knowledge can compromise confidentiality.

This paper begins a methodological exploration to facilitate the availability of detailed socio-ecological data to advance understandings of population-environment-health connections. Our goal is ultimately to fuel research to inform policy designed to safeguard human and ecological well-being. The methodological exploration presented here involves innovative processing and analyses of data from a low-income setting in rural South Africa, a region from which social surveillance data are underutilized for the purposes of research on population-environment-health interactions. Several techniques of data anonymization are tested that are designed to mask true household locations in order to protect confidentiality. The primary contribution of this study is the identification of anonymization techniques—for this particular context and contextual measure—that yield more accurate estimations of environmental measures from anonymized household locations relative to the measures from the "true" household locations while sufficiently preserving confidentiality. Importantly, however, the process presented here is illustrative in that a wide variety of methodological decisions must be made that should be context-specific as well as driven by a project's particular research objectives.

Background

Understanding associations between human populations and their environments is becoming increasingly important as evidence of climate change continues to mount. Such understandings are particularly essential to inform programs and policies in settings of high climate vulnerability including many rural areas of the Global South (Byers et al., 2018). In many such regions, livelihoods remain intimately intertwined with proximate natural resources (e.g., Wisley et al., 2018), resources which may become increasingly scarce in areas anticipating shifts in rainfall and heat extremes (e.g., Olsson et al., 2014).

To increase the sustainability of rural livelihoods, research at the individual- or household-scale is of particular importance since livelihood decision-making is focused within these realms (e.g., Sumner et al., 2017). A wide variety of useful secondary social science datasets are available for such examinations. These often include confidential information from human subjects including age, gender, race, ethnicity, income, livelihood strategies, and, in some cases, particular health outcomes. Collection of such data typically comes with important assurances as to confidentiality and protection of individual right to privacy.

Yet, examination of the environmental dimensions of well-being requires linking this individual- or household-scale data with information reflecting local environments. As two examples, locational data allow linking households to information on proximate rainfall and temperature conditions, thereby facilitating environmental health research. Locational data also allow for disease mapping to identify spatial clusters of incidences or outbreaks for programmatic targeting. Linking the necessary social and environment data for these questions requires geocodes—the geographic locations of survey or census respondents. Yet, making available such geographically specific data would typically violate ethical and legal requirements regarding the confidentiality of microdata in that respondents' identities may be revealed. This tradeoff as related to protecting privacy while maintaining sufficient accuracy is the central issue addressed within this paper.

There are several approaches to address data confidentiality concerns with the most restrictive being complete non-disclosure such as the destruction of all location information after data collection has ceased. Partial disclosure can be implemented along a continuum, with any particular approach having its own drawbacks. There are no broadly representative comparative studies that examine the relative strengths of the many approaches nor that offer guidance on method choice. This paper's contribution is examination of the tradeoffs between anonymity and analytical precision within one such method—geographic masking—thereby offering important insight related to this commonly used approach. The analyses presented make use of an environmental variable measuring vegetation cover and are undertaken in a rural South African study site. The work is illustrative of the process, and the implications, of different geo-masking approaches and is intended as guidance for population-environment researchers interested in linking micro and contextual data while maximizing privacy.

Of the variety of methods to protect anonymity, restricted data enclaves offer access to the highly specific geographic information. An example is the US Census Bureau's network of Federal Statistical Research Data Centers (FSRDC) through which precise geographic information may be made available with highly structured confidentiality agreements and a requirement of travel to highly secured data centers for access. Such agreements also typically require high levels of oversight as to public presentation of results.

Offering more accessibility but substantially less spatial precision, geographic information is occasionally available within secondary data sets. In many cases, readily available spatial units are quite coarse (e.g., municipality in the Mexican Migration Survey). Yet, many organizations offer the possibility of more spatial precision through confidentiality agreements with individual researchers for specific projects. The Panel Study of Income Dynamics, for example, provides the option through a security agreement to access US Census linkages to tracts, block groups, or blocks which offer far more precision than positioning respondents in their larger-scale counties. Other methods to protect respondent identity include aggregation whereby data are transformed to characterize geographic units (e.g., county-scale age, gender, racial composition). Aggregation units can sometimes be quite small although the ecological fallacy remains a risk. This occurs when inferences are made about individuals or other subunits based on aggregated characteristics. Also, given the loss of microdata, aggregation yields less analytically useful data for the purposes of understanding micro-scale social and economic processes. Other approaches include spatial smoothing where data represent weighted averages of individual-level data averaged spatially by nearest neighbors (Zhou & Louis, 2010) and multiple imputation which simulates datasets that capture dependencies among variables in the original data (Wang & Reiter, 2012). Linear programming approaches add noise to individual locations based on mathematical probabilities to minimize risk of individual disclosures based on the desired level of privacy. Data swapping entails switching values between various records, while synthetic data entails the creation of a dataset which has similar properties to original data where individuals cannot be identified. The critical importance of the privacy/accuracy tradeoff is illustrated by the debate surrounding differential privacy methods proposed for US Census data. Broadly, differential privacy entails the addition of a precise amount of statistical noise—i.e., synthetic records—into the released dataset such that a user cannot identify individual data (Abowd & Schmutte, 1999). Ruggles et al. (2019) argue that proposed methods have the potential to reduce the utility of microdata and smaller-area estimates and may disproportionally impact racial/ethnic minorities and underrepresented individuals.

Another approach to respondent protection is illustrated by the Demographic Health Survey (DHS) which provides geocodes representing the center of a geographic cluster or small settlement. Each location is "geo-scrambled" to randomly add position error, the distance influenced the local population density. Within this approach, the sociodemographic and economic characteristics of the households, themselves, are unchanged but instead the location is altered. As such, the contextual variables generated for households will reflect the displaced, as opposed to original location. Several studies have cautioned researchers to carefully consider the spatial error introduced within analyses using the DHS (e.g., Elkies et al., 2015) and innovative alternatives to the DHS clusters has been proposed such as using characteristics of nearby communities as proxies for environmental conditions (Grace et al., 2019).

Related, and the focus of this paper, geographic masking—also known as "geomasking" or "jittering"—entails displacement of the individual or household location using predefined parameters typically related to direction and distance from the true location. Proposed as early as 1999 (Armstrong, Rushton, and Zimmerman), geomasking has received far more attention in public health and epidemiological research as compared to population science although potentially of substantial use to demographers.

Approaches to geographic masking

Masking techniques typically include some form of spatial dislocation to reduce the potential for identification of study households. A variety of techniques have been developed that structure displacement of the original locations through different

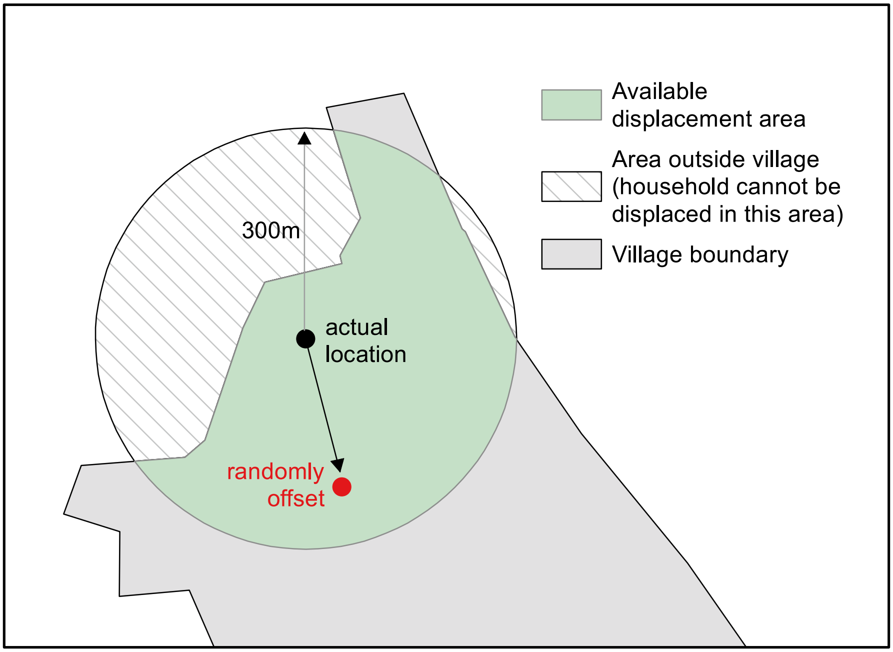

Available displacement area
Area outside village (household cannot be displaced in this area)
Village boundary
300m
actual location
randomly offset

Fig. 1 Example of a simple random offset for a households with available displacement area constrained by village boundary

approaches to randomization of distance and direction. After displacement, the original locations are removed from the dataset that is made publicly available or available through a data sharing agreement.

One of the most straightforward approaches is presented in Fig. 1 where a household is simply randomly offset within a buffer of predetermined size (in this case, 300 m). In the case of Fig. 1, the displacement is constrained by the village boundaries so that a household is not displaced outside the village.

The use of randomization has become more common in context-centered research (Armstrong et al., 1999; Cassa et al., 2008; Lu et al., 2012). However, there is little consensus as to the amount of displacement necessary to preserve confidentiality. One approach to quantitatively measuring privacy risk is the spatial *k*-anonymity factor, where *k* is represents the number of people (or households) needed within a buffer to preserve confidentiality (Sweeney, 2002). This is an extension of the concept of *k*-anonymity where data are released only if there is a minimum of *k*-1 individuals with the same combinations of characteristics (Zandbergen, 2014). In its spatial version, *k*-anonymity approaches consider the displacement distance necessary to protect privacy given a particular population density. Dense urban settings require less distance in displacements than sparsely populated rural areas (Cassa et al., 2008).

To better understand the implications of the various approaches to balancing research and confidentiality, we explore the differences that displacement brings for a particular contextual measure reflecting proximate vegetation, described below. We do so within a longstanding study site, the MRC/Wits-Agincourt Unit in rural South Africa.

Research setting

The Agincourt Health and Socio-Demographic Surveillance System (AHDSS)—situated in the far northeast of South Africa—is operated by the Medical Research Council (MRC) and University of the Witwatersrand (Wits) Rural Public Health and Health Transitions Research Unit (MRC/Wits-Agincourt Unit). The study area of 450 km^2 study includes 31 villages which are home to ~ 110,000 residents in ~ 22,700 households. Since 1992, the Agincourt Unit has conducted an annual census including the entire Agincourt HDSS population (Collinson, 2010).

A "homeland" area where black South Africans were forcibly resettled during the era of Apartheid, the study site is characterized by relatively high population densities (~ 170 persons per sq. km), high poverty, and a longstanding lack of development and access to state services (Collinson, 2010). The Agincourt study site's settlement pattern is fairly typical of rural communities across South Africa, and socioeconomically, it is characterized by a high reliance on remittances from the large proportion of adults who are migrant laborers on commercial farms and in towns and cities across the country. A substantial portion of households also depend heavily on the state pensions of elderly members (Collinson, 2010).

The region is generally dry (annual rainfall 550–700 mm), although an east–west rainfall gradient results in local variation in natural resource availability. Homestead plots are typically too small to fully support subsistence agriculture and some households farm assigned plots in the surrounding communal lands. Residents are highly dependent on the natural environment for a range of uses. These include grazing livestock and collecting fuelwood, wild foods, thatching grass, construction timber, and other domestic products both for household consumption and for generating income (Paumgarten & Shackleton, 2011).

The centrality of natural resources to livelihoods in rural South Africa is key to the illustrative research presented here. Case studies in two rural villages found that 70% of households made use of non-timber forest products, such as fuelwood, wild fruit, and edible herbs during times of shortage and crisis (Paumgarten & Shackleton, 2011). Even in rural South African villages with readily available electricity, over 90% of households use fuelwood as a primary energy source due to the cost of electricity and appliances (Matsika et al., 2013). This trend has been observed in and near the Agincourt study site where natural resources also act as buffers against household shocks such as a breadwinner's death (Hunter et al., 2007).

There is vast potential for linking population, environment and health data within Health and Demographic Surveillance Systems (HDSS). The INDEPTH network includes 48 such sites in low-income settings across sub-Saharan Africa, Asia, and Oceania. These study settings undertake continuous monitoring of all individuals within a defined study setting and, combined, the INDEPTH network data provides longitudinal health and demographic insight on nearly 4 million individuals in 18 countries, providing critically important opportunities for policy-relevant scholarship (INDEPTH, 2017).

The application of "geographic masking" to facilitate contextual research is, however, nascent within the HDSS community. A recent study with professionals possessing a working knowledge of surveillance systems found that most respondents (83.5%) were not aware of any written rule, policy, or regulation governing research with HDSS data although nearly 86% agreed that there was a need for such guidelines. The risk of personal or family data being compromised was of great concern, and 74% supported anonymizing data before release to researchers (Anane-Sarpong et al., 2016). Per special agreements with particular HDSS' (e.g., Africa Centre Demographic Information System), temporary access to locational information can sometimes be obtained by individual researchers (e.g., Tlou et al. 2017). But a key motivation of the present project is to explore the potential for geographic masking approaches to facilitate greater data sharing for the purposes of population-environment-health research especially as related to current and future climate vulnerability.

Data and methods

We anonymize data from the AHDSS using nine different approaches (described below) and evaluate the impact of the geomasking through comparison of measures of proximate natural resources. The AHDSS provided data on specific physical locations of households but without any additional individual or household information, and in total the locations represent 31 separate villages with 22,708 households (Fig. 2).

Given our focus on facilitating population-environment-health research, we examine the implications of geomasking for a vegetation measure reflective of proximate natural resources which are fundamental to most livelihoods in the Agincourt study site and predominantly collected from communal lands surrounding villages. Specifically, we use the normalized difference vegetation index (NDVI) which is well-correlated with vegetation amount and quality (Roerink et al., 2003; Wessels et al., 2004). It is important to acknowledge that NDVI is used here to illustrate the process involved in geomasking and to allow for exploration of the tradeoff between privacy and accuracy. This example demonstrates the important questions to be asked within the context of spatially-informed scholarship although the choice of specific data reflecting the proximate environment must be driven by research objectives.

NDVI values are derived from data from the Landsat 5, 7, and 8 missions and the index's calculation exploits vegetation's reflectance of near-infrared light and the absorption of red light (Tucker, 1979). The data used to calculate NDVI is consistent across the Landsat missions due to Collection 1 data processing completed by the

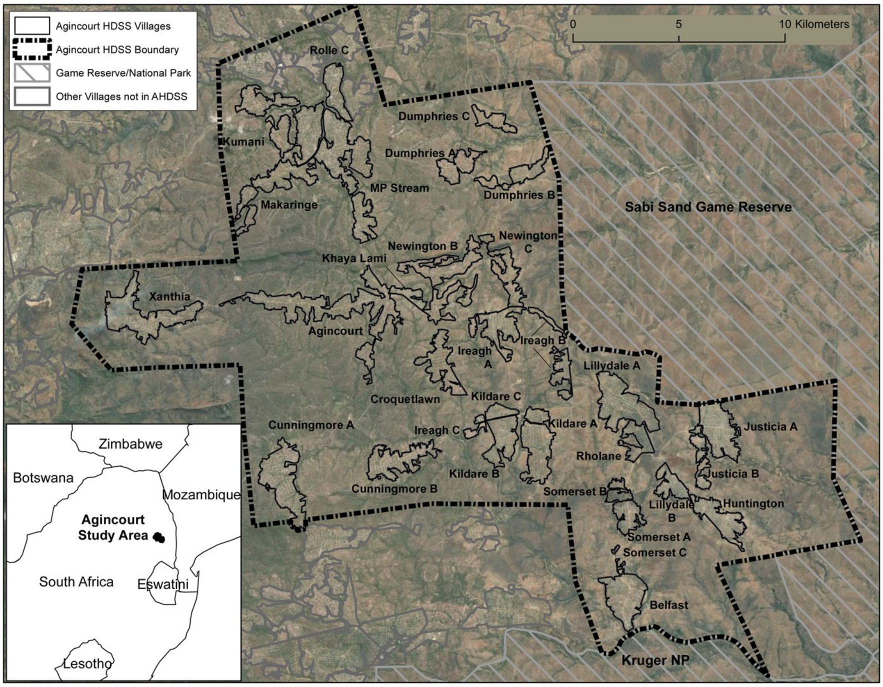


Fig. 2 Villages within the MRC/Wits-Agincourt Unit

USGS Earth Resources Observation and Science (EROS) Center. Values range from − 1 to + 1 with vegetation biomass and productivity positively correlated with NDVI (Foody et al., 2001; Mutanga & Skidmore, 2004; Wang & Rich, 2008). Low values (≤ 0.1) indicate barren land, rock, sand, or water, moderately positive values (0.2–0.3) may correspond to shrublands or grasslands, while high values (0.6–0.8) correspond to temperate or tropical rainforests (NASA, 2000).

The Landsat data includes NDVI estimates for a given location at least every 16 days, and every 8 days when considering Landsat 7 and Landsat 8 overpass, at a resolution of 30 m (~ 100 ft). Two corresponding files were obtained for each date; a NDVI image file containing processed NDVI values and the associated pixel QA file. The pixel QA file is a raster image with the same pixels as the remotely sensed image, but each pixel is given a number identifying its usability, reporting image quality issues for each associated pixel of the NDVI data.[1]

This project incorporates data from March 1997 to December 2017 to capture temporal changes in NDVI. Additional processing included filtering NDVI values based on typical quality control criteria. For instance, data reporting "Cloud

[1] The NDVI Values and Quality Assessment (QA) files were obtained from the Land Satellites Data System (LSDS) Science Research and Development (LSRD) repository provided by the US Geological Survey (USGS) Earth Resources Observation and Science (EROS) center (LSRD, 2018).

Shadow," "Cloud," or "Water" in each corresponding QA file were excluded from the analysis. Undertaking the data integration at the pixel scale allows for better coverage since areas with clouded coverage can be deleted from consideration without the need to discount the entire Landsat image from consideration. Areas within village boundaries were also eliminated from consideration since they do not represent the communal areas where resource collection takes place. We also excluded neighboring game reserves and parkland (see Fig. 2) since village residents typically do not have access to these spaces.

We use 2000 m (2 km) buffer zones within which the NDVI associated with a household point location is calculated—the choice was informed by research on typical travel distances for natural resource access (Giannecchini et al., 2007).[2] The determination of appropriate buffer size must be informed by cultural context and existing knowledge of population-environment linkages within that socio-ecological space. Median NDVI values within each individual buffer zone are estimated as a measure of the central tendency of NDVI values available to each household from March 1997 to December 2017. We also use a measure of household resource availability that is the sum of the NDVI values divided by the number of households in each individual household buffer zone. This metric serves as a proxy of relative resource availability accounting for both distance of access and the number of households that may share proximate resources. Households located further from village boundaries or in high-density areas have lower natural resource availability than households near village boundaries or in less-dense areas (e.g., see Leyk et al., 2012).

For this project, NDVI values are not aggregated over long periods of time (e.g., mean annual value) since such calculations necessarily diminish information on seasonality and other within-year changes. Instead, we incorporate 200 estimates of vegetation availability for each of the 22,708 households for an average of 10 months per year, 1997 to 2017 reflecting one measurement per month selected as close as possible to the middle of the month where available.

In the current study, we employ nine geomasking techniques; four of these use the "donut" approach such that no points are offset within a minimum radius. Approaches that use donut masking include (1) random displacement, (2) offsets that represent Gaussian distributions of displacement, and both (3) random displacement and (4) Gaussian displacement with a "distance/density factor" (explained below). For these four approaches, we consider a maximum radius of 300 m with a 150 m exclusion zone (i.e., residence locations displaced 150 to 300 m). This distance represents on average, the largest distance between any two households located within the same village. Larger distances were tested but could not be used due to smaller footprints of several villages. We also limit displacement to a household's village extent (see Fig. 1). The influence of the village boundary constraint varies by village spatial size; as would be anticipated, larger villages have greater displacement potential. Village shape matters, too. For instance, Ireagh B is

[2] Calculations were also made with 1 km buffers with no substantial differences in overarching conclusions.

long and narrow and has smaller potential areas of displacement than villages of a similar size (Dumphries B) for all methods of displacement.

As mentioned, the “distance/density factor” is also considered in several of our illustrative masking approaches. The factor is calculated by relating the local population density (e.g., village) to total population densities (e.g., study region) (Cassa et al., 2008). Basically, within such weighted displacement approaches, the baseline displacement distance within buffers (300 m) is adjusted to compensate for proximity of households to one another within a particular village. More dense villages require less displacement for privacy protection.

$$Total\ density\ multiplier = \frac{Average\ total\ household\ density}{Village\ household\ density}$$

We also examine a method that allows for adjustment of displacement distance that adds consideration of k-anonymity to the density adjustment (Allshouse et al., 2010). With this approach, the minimum (R_{ai}) and maximum (R_{bi}) displacement distance is defined by the density of households and user-defined levels of k-anonymity:

$$R_{ai} = \left(\left(A_i/\pi\right) \times \left(k_a/N_i\right)\right)^{1/2}$$

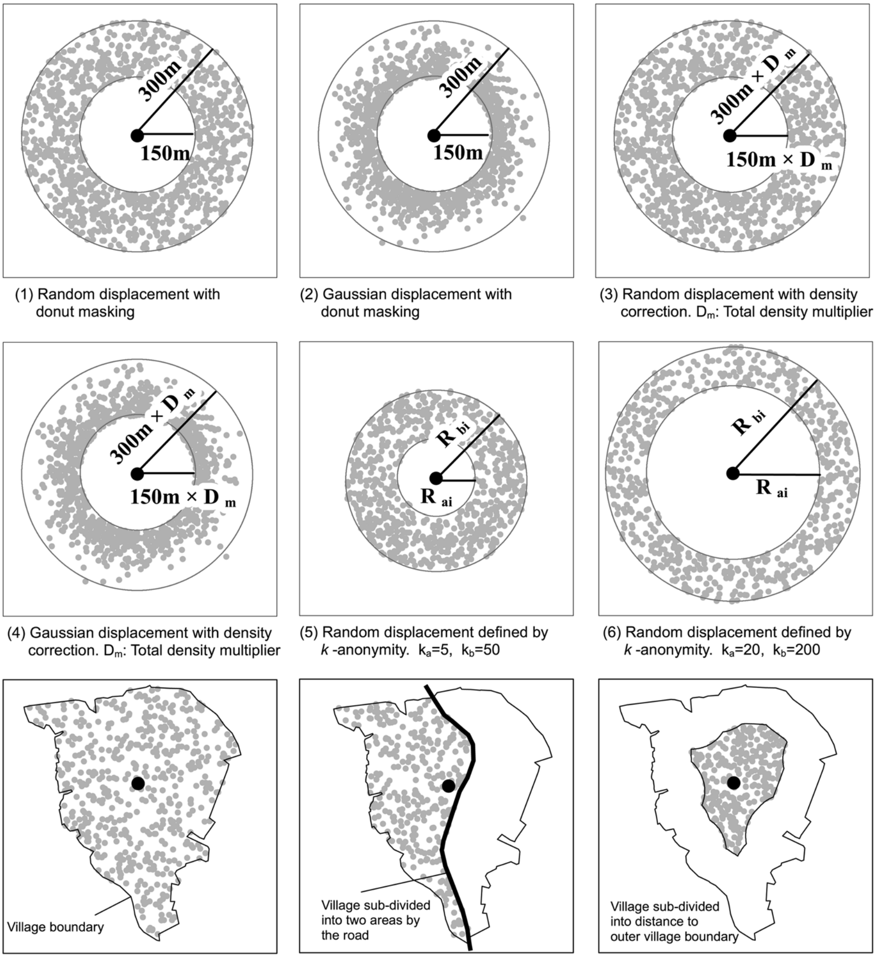
$$R_{bi} = \left(\left(A_i/\pi\right) \times \left(k_b/N_i\right)\right)^{1/2}$$

N_i represents the number of households in each village and A_i is the area. Here, k_a is the minimum displacement threshold and k_b equals the maximum displacement threshold. For example, specifying $k_a = 5$ assumes that at least five households will be in closer proximity to the true household location than the displaced household location. For this study, we specify $k_a = 5$ and 20 with $k_b = 10 \times \mathrm{k_a}$. The decision to use a multiplier of 10 follows work by Allshouse et al. (2010) and this particular measure is central to our exploration of the tradeoff between privacy vs. accuracy. These approaches are also displayed in Fig. 3 as (5) with $k_a = 5$ and $k_b = 50$, and (6) with $k_a = 20$ and $k_b = 200$.

Our final three illustrative approaches involve random displacement of households within a geographic area defined differently than a circular buffer. These three approaches focus on within-village geographic clusters pre-defined by spatial attributes that likely shape livelihood strategies such as proximity to a major road or distance to communal lands. These involve random assignment and are also presented in Fig. 3, numbered within (7) the entire village; (8) sub-village areas defined by physical boundaries such as roads, rivers, or railroads; and (9) sub-village areas defined by distance from edge of village boundary in 100 m buffer zones.

Village descriptive profiles

There is substantial variation in overall population size across the AHDSS villages (Table 1); the village of Lillydale A is home to over 1600 residents while one subsection of Somerset, Somerset B, has only 71 residents. This wide variation plays into


300m
150m
300m
150m
300m + D m
150m × D m
300m + D m
150m × D m
R bi
R ai
R bi
R ai
Village boundary
Village sub-divided into two areas by the road
Village sub-divided into distance to outer village boundary

(1) Random displacement with donut masking

(2) Gaussian displacement with donut masking

(3) Random displacement with density correction. D_m: Total density multiplier

(4) Gaussian displacement with density correction. D_m: Total density multiplier

(5) Random displacement defined by *k*-anonymity. k_a=5, k_b=50

(6) Random displacement defined by *k*-anonymity. k_a=20, k_b=200

(7) Random displacement within village

(8) Random displacement within sub-village by infrastructure (roads)

(9) Random displacement within sub-village by distance to village boundary

Fig. 3 The nine illustrative geographic masking techniques examined. Black circle represents the original location; gray dots represent simulated possible locations using each masking method with a radius of 300 m and exclusion zones of 150 m for masking approaches 1 and 2. For approaches 3 and 4, the radius and exclusion zones are adjusted from 300 and 150 m using a total density multiplier (D_m). D_m = average total household density/village household density. For approaches 5 and 6, possible locations are placed within a radius R_{bi} and exclusion zone R_{ai}. Approaches 7, 8, and 9 represent possible locations of displaced households within a village or sub-village boundaries

the differentials in density, with the highest density in Somerset C (1180 households/km^2) and lowest density in MP Stream (117 households/km^2), a tenfold distinction. These variations speak to the need for randomization methods that account for substantial differences in size and density in that offsetting households a large distance may not be practical in the village of Somerset C with its small geographic

Table 1 Summary statistics for true household locations, growing-season months, 2 km buffers, MRC/Wits-Agincourt Unit

Village Name	Number of households (household/km^2)	Area (km^2)	Density(household/km^2)	Distance Factor	Median NDVI					Sum NDVI/Households				
					Range	Average	Median	Min.	Max.	Range	Average	Median	Min.	Max.
Xanthia	941	5.2	182	1.43	0.795	0.447	0.460	0.013	0.807	65.19	8.30	7.43	0.14	65.33
Cunningmore A	1305	4.6	282	0.92	0.786	0.423	0.422	0.009	0.795	13.82	3.91	3.89	0.06	13.87
Cunningmore B	807	3.1	263	0.99	0.746	0.389	0.383	0.011	0.757	17.70	6.52	6.54	0.15	17.85
Agincourt	1406	7.8	180	1.44	0.776	0.388	0.374	0.010	0.786	68.16	7.62	5.37	0.06	68.22
Makaringe	582	1.2	472	1.64	0.776	0.383	0.369	-0.015	0.761	66.68	8.84	6.07	-0.17	66.50
Kildare B	908	2.6	350	0.74	0.737	0.377	0.370	0.012	0.749	6.72	1.72	1.60	0.03	6.75
Rolle C	399	1.9	216	1.21	0.710	0.374	0.365	0.008	0.718	16.66	4.24	3.25	0.04	16.69
Kildare C	256	0.7	347	0.75	0.717	0.371	0.365	0.012	0.729	5.61	2.31	2.30	0.06	5.67
MP Stream	1095	9.4	117	1.65	0.711	0.371	0.359	0.008	0.719	29.88	5.28	2.74	0.02	29.90
Kumani	1322	6.2	214	1.22	0.723	0.371	0.358	-0.022	0.701	19.34	3.04	2.02	-0.40	18.94
Ireagh C	298	0.3	872	0.30	0.707	0.371	0.363	0.012	0.719	7.45	3.10	3.13	0.08	7.54
Croquetlawn	719	2.7	270	0.96	0.707	0.368	0.358	0.011	0.718	12.91	5.65	5.45	0.09	13.01
Newington C	418	2.2	190	1.37	0.745	0.360	0.336	-0.006	0.740	24.02	7.07	6.29	0.08	24.10
Khaya Lami	813	0.8	1029	0.25	0.683	0.359	0.340	0.010	0.693	4.32	2.14	2.07	0.06	4.38
Justicia A	1240	3.5	354	0.74	0.758	0.358	0.333	0.011	0.770	8.01	2.07	2.01	0.05	8.06

Table 1 (continued)

Village Name	Number of households (household/km^2)	Area (km^2)	Density(household/km^2)	Distance Factor	Median NDVI					Sum NDVI/Households				
					Range	Average	Median	Min.	Max.	Range	Average	Median	Min.	Max.
Dumphries A	499	1.8	282	0.92	0.727	0.357	0.342	0.005	0.732	17.28	7.86	7.40	0.10	17.38
Newington B	656	4.6	144	1.81	0.725	0.356	0.339	0.008	0.733	23.43	3.95	3.34	0.06	23.49
Dumphries C	262	1.1	244	1.07	0.677	0.353	0.342	0.008	0.685	35.83	14.21	13.70	0.24	36.08
Ireagh B	506	1.8	287	0.91	0.712	0.350	0.347	-0.015	0.697	24.14	7.49	7.14	-0.40	23.74
Kildare A	996	3.9	255	1.02	0.739	0.350	0.337	0.011	0.750	15.83	2.45	2.00	0.04	15.87
Belfast	1316	3.7	360	0.72	0.700	0.348	0.331	0.014	0.714	6.84	2.59	2.55	0.08	6.92
Ireagh A	813	2.8	292	0.89	0.715	0.346	0.333	-0.011	0.704	16.29	3.95	3.85	-0.10	16.19
Somerset A	644	1.6	412	0.63	0.700	0.345	0.331	0.013	0.713	8.93	4.07	4.06	0.12	9.05
Huntington	787	2.6	308	0.85	0.707	0.343	0.320	0.013	0.721	12.96	4.52	4.24	0.11	13.08
Dumphries B	467	1.7	274	0.95	0.707	0.341	0.321	0.009	0.716	19.98	7.61	7.26	0.13	20.11
Lillydale B	515	1.5	343	0.76	0.701	0.340	0.319	0.013	0.714	8.26	3.76	3.75	0.11	8.37
Justicia B	320	1.3	242	1.08	0.720	0.340	0.314	0.012	0.731	5.71	2.27	2.23	0.06	5.77

Table 1 (continued)

Village Name	Number of households (household/km^2)	Area (km^2)	Density(household/km^2)	Distance Factor	Median NDVI					Sum NDVI/Households				
					Range	Average	Median	Min.	Max.	Range	Average	Median	Min.	Max.
Somerset B	233	0.8	275	0.95	0.692	0.337	0.323	0.013	0.705	11.40	4.30	4.23	0.13	11.53
Somerset C	71	0.1	1180	0.22	0.661	0.336	0.323	0.014	0.675	7.86	3.96	3.88	0.15	8.01
Lillydale A	1626	4.5	362	0.72	0.748	0.335	0.308	-0.016	0.731	5.65	1.68	1.60	-0.11	5.54
Rholane	488	1.5	317	0.82	0.709	0.326	0.297	0.012	0.721	6.57	2.22	2.05	0.05	6.62

Sorted by NDVI range = difference between highest and lowest household NDVI within village

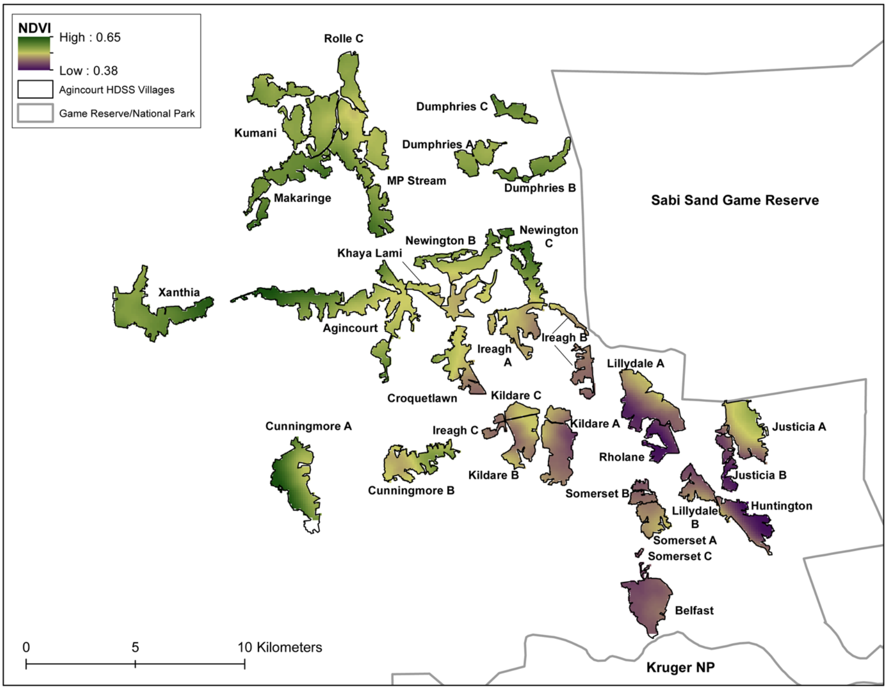


Fig. 4 NDVI values (Jan 2010) for true household locations, 2 km buffers, MRC/Wits-Agincourt Unit

footprint. However, for Somerset C, consideration must also be given to its low overall population which poses important challenges to privacy.

As to the environment, Fig. 4 presents estimated NDVI values for households during the summer of 2010 and reveals a substantial west–east greenness gradient. The study area's western side is indeed characterized by slightly higher elevations, greater variation in topography, and more precipitation. It is also clear that villages in the eastern portion confront substantial resource constrains given both low NDVI values and boundaries with fenced reserves.

Understanding implications of geomasking

Ultimately, this study's objective is to demonstrate a process whereby researchers might better understand the implications of geomasking. Here, we aim to examine the significance of the differences in NDVI values for each masking approach as compared to the NDVI calculated based on the true household location. As such, the focus is on differences between geomasking methods as opposed to differences between households. The reference category represents the NDVI estimates derived from true household locations.

To generate a quantitative understanding of the distinctions in NDVI values, we determined differences between the two NDVI estimates calculated for the true household locations relative to those displaced. To quantify distinctions, however, our evaluation considers only growing-season months (September to April); Table 1 also presents this village-scale descriptive information across the entire study period, sorted by the average of the monthly median NDVI. The results reveal wide variation in the range of median NDVI values between and within villages. For instance, median NDVI values in Xanthia varied by 0.79, a relatively large amount given that NDVI values generally range between 0.0 and 0.8 in the study area. The villages of Cunningmore A, Makaringe, and Agincourt also show a high degree of variation in median household NDVI values. This pattern is also observed in the range of NDVI values estimated as the sum of NDVI divided by the number of households; the highest levels of variation are in the villages of Xanthia, Makaringe, and Agincourt. Although there is not a particularly strong correlation between density and NDVI, the villages in the study area's western portion do tend to be larger in size and relatively less dense than villages with lower levels of variation of NDVI values which tend to be in the eastern portion of the study area (e.g., Somerset B, Lilydale A, and Rholane). Even so, the village with highest density, Xanthia, actually has especially high NDVI likely due to the area's rolling topography and proximity to surface water sources.

These descriptive glimpses into variation are intriguing, but we also aim for a more thorough sense of the significance of these differences. To this end, and given the study's longitudinal nature, we use multilevel models where repeated measures of NDVI (level 1) are nested within households (level 2) with both the intercept and slope for households varying across time. Since NDVI is a continuous measure, we use linear mixed regression models with fixed effects representing the different methods of randomization—the effect of the method used to estimate natural resource availability is reflected by the fixed effect's estimated coefficient. For example, a statistically significant coefficient for the fixed effect using randomization approach 2 (i.e., Gaussian displacement with donut masking) would suggest important differences between the NDVI values calculated via this method and those reflecting NDVI surrounding households' true locations while also accounting for temporal differences in NDVI. Analyses were conducted for each village to estimate variation of randomization methods at the village scale. This approach allows us to examine potential sources of deviation in given approaches that may be associated with regions of high or low NDVI values, or areas within the study site that have a high degree of variation in resource availability.

Results for median and sum NDVI are presented in Appendix Tables 3 and 4, respectively. Two particularly intriguing findings emerge. First, methods of displacement that account for housing density (random, Gaussian, and *k*-anonymity) provide results most similar to the median NDVI estimates of the true household locations as demonstrated by a consistent lack of statistically significant differences, especially for median NDVI estimates. While there are a greater number of statistically significant differences for the sum NDVI estimates, these differences are substantively small (less than 4%) for three specific methods (random density, Gaussian density and $k_a=5$ $k_b=50$). The greater number of significant differences is likely a result of the introduction of larger variation in NDVI estimates for masked household locations. For instance, sum NDVI estimates are more sensitive to the position of the household relative to the village boundary, where a house located in the interior portion of a village will have a smaller sum NDVI estimate than a house located near the boundary of a village and closer to the surrounding vegetation.[3]

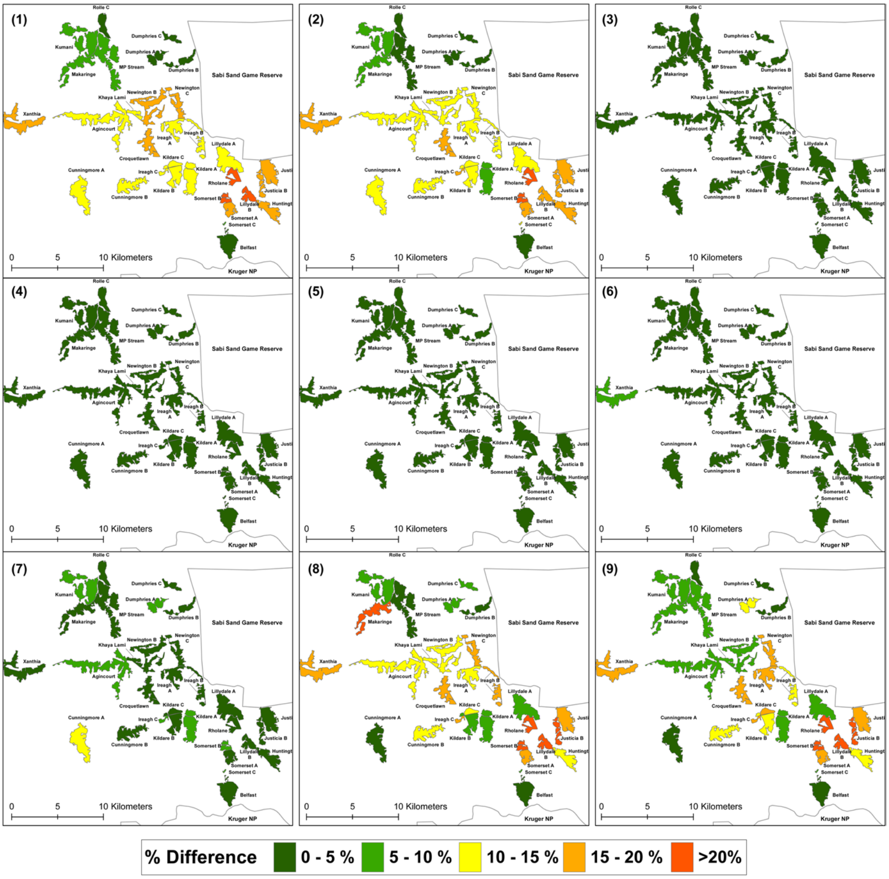


Fig. 5 Spatial distribution of difference in sum of NDVI values/number of households, true vs. geomasked locations, MRC/Wits-Agincourt Unit (1) Random, donut, (2) Gaussian, donut, (3) Random (density), donut, (4) Gaussian (density), donut (5) ka=5 kb=50, donut, (6) ka=20 kb=200, donut, (7) Random within village, (8) Random within sub-village by infrastructure (roads), (9) Random within sub-village by distance to boundary

[3] We also examined the impact of removing boundary constraints which increases the distance of displacement for households by an average of 14%. However, the displacement distances varied substantially across villages. For example, villages with lower household density did not see large gains in displacement as compared to when with masking methods that account for household density. Also, household displacement distances were on average, lower when using *k*-anonymity methods. In all, this suggests that the household density is more limiting in terms of constraining displacement distances than the village boundary.

Second, there are a few villages within which almost all of the geomasking approaches simply miss the mark, most notably Rholane, Somerset B, and Lilydale B. To spatially represent these patterns of error, Fig. 5 presents the differences between true locations for the sum NDVI estimates for each of the nine masking methods. As shown in Fig. 5, there is a high degree of variation in NDVI estimates depending on the method used. However, villages in the central and eastern portions of the study area consistently show a large difference between true and randomized NDVI values.

To explore potential explanations for the variation in accuracy, we examined the underlying distributions of NDVI, population, and density, as well as examining other potentially influential factors such as village proximity to other villages and the study region's edge. Regardless of the method of masking, villages that are smaller and more densely populated tend to have lower variation between true and displaced NDVI values (e.g., Somerset C). This is because smaller and more-dense villages tend to have relatively low displacement distances—especially when using density-dependent approaches—which then results in less difference between true and displaced NDVI values. In the case of both random and Gaussian approaches of geomasking (panel 5a and 5b), villages with less difference between true and randomized NDVI values are located in the northern part of the study area (Makaringe, MP Stream, Dumphries A, Dumphries B, Dumphries C, Rolle C, and Kumani). These villages generally exhibit higher levels of NDVI compared to the middle part of the study area where differences between true and randomized NDVI values are greater. Two southern villages—Belfast and Somerset C—also show better agreement. This is due to a combination of factors: the villages are relatively small and have small ranges of NDVI values. Belfast's location at the edge of the study site further reduces variability between true and displaced NDVI values. In all, the spatial variation in error is due to the influence of NDVI variability, population, and/or population density, as well as other factors such as household proximity to village edges, other households, as well as village proximity to other villages and protected areas. In a particular research project, it would be useful to consider how variation in these factors combine with your variable of interest to influence the accuracy of displaced values across the study region. Such understanding may ultimately influence the decision as to which geomasking method is most appropriate for a particular research project.

Tradeoffs between anonymity and accuracy

In addition to measurement error, it is important to understand potential implications for anonymity within each geomasking approach. In essence, this illustrates the balancing act that represents a key contribution of this paper but, as a reminder, these analyses illustrate a process which should be undertaken within spatially-explicit

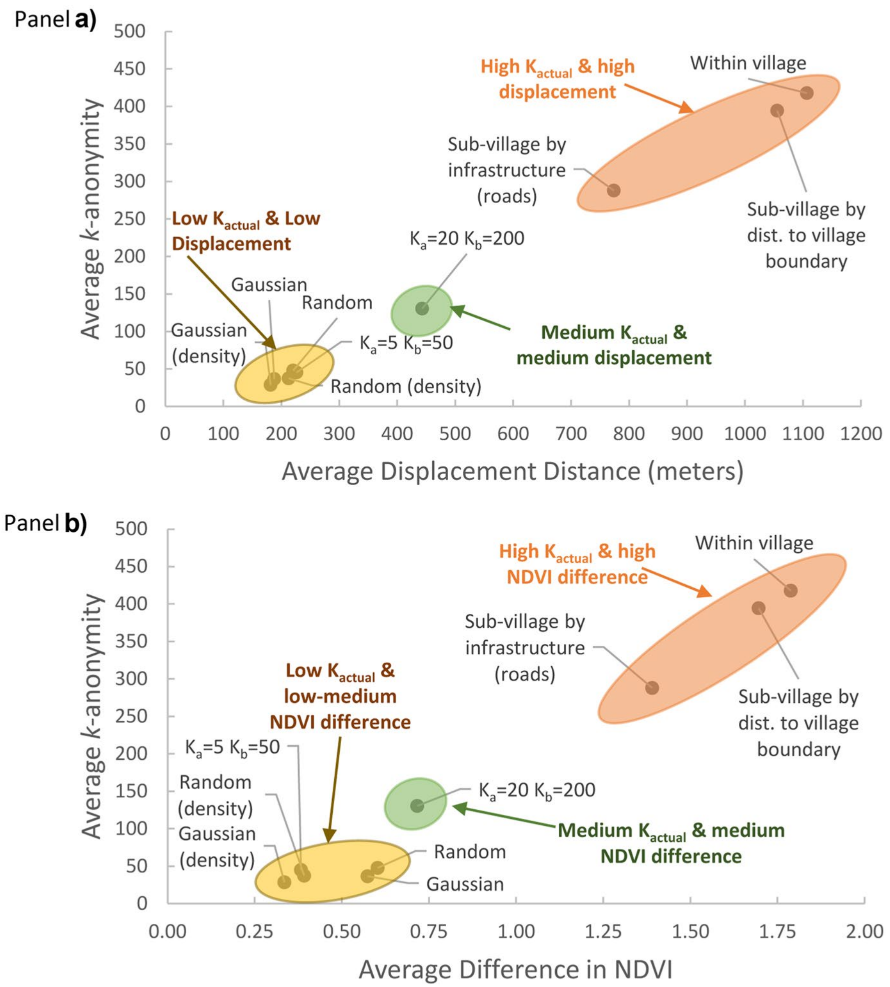


Fig. 6 Evaluation of accuracy and privacy for true vs. geomasked locations, MRC/Wits-Agincourt Unit. **a** Displacement distance versus *k*-anonymity. **b** *k*-Anonymity versus absolute difference in NDVI

scholarship where anonymity is a concern. Recall that there are research-specific choices to be made regarding relevant environmental variables, buffer sizes, required anonymity threshold, and the like.

To illustrate tradeoffs, we calculated average displacement between true household locations and masked locations and the estimated *k*-anonymity for each location averaged within the villages (defined as the number of households that are closer to the true location than masked). The village-specific results are presented in Appendix Table 5, while they are summarized within Fig. 6a. On average, and as would be expected, masking methods that tend to be closer to the true location tend to have lower *k*-anonymity. In particular, random displacement and Gaussian approaches adjusted for density (approaches 1–4) provide smaller displacement distances and lower levels of actual *k*-anonymity (yellow oval, Fig. 6a). As shown in Appendix Table 5, for instance, in the least populated village, Somerset C, density-informed, random geomasking result in an average of eight households being closer to the true location than the masked household. The density-informed Gaussian displacement lowered this to six.

Conversely, geomasking approaches that incorporate medium levels of both household density and *k*-anonymity (approaches 5 and 6) provide relatively higher levels of displacement and anonymity (green oval, Fig. 6a). On the other extreme, relatively high levels of anonymity and displacement characterize geomasking within village boundaries (approaches 7–9), as opposed to those based on buffers (orange oval, Fig. 6a). In particular, approach 9 that uses sub-village areas as defined by distance to village boundary reached as high as 833 in the densely populated village of Agincourt, the study site's namesake.

To evaluate implications of anonymity and accuracy, for each geomasking method we next plot the average *k*-anonymity versus the average difference in the sum of NDVI between true and masked household locations (Fig. 6b). We find that methods that have lower levels of *k*-anonymity (for example, random and Gaussian displacement) also tend to have smaller differences in NDVI measures (yellow oval, Fig. 6b). Conversely, methods with higher levels of *k*-anonymity tend to have larger differences in NDVI measures (for example, random village displacement). While methods that randomize locations within a village or sub-village boundary indicate higher levels of *k*-anonymity, these methods have relatively large differences in NDVI measures. Of particular interest, methods that incorporate both household density and *k*-anonymity as highlighted by the green oval in panel b ($k_a = 20$ $k_b = 200$) exhibit high levels of *k*-anonymity and relatively small differences in vegetation measures, suggesting perhaps the most effective approach for NDVI in this study setting.

These findings on tradeoffs are logical since larger displacements will yield greater anonymity although also being further from the true location and, therefore, more likely to yield a larger difference in true vs. displaced measures. Again, it is not our goal to firmly decide which of the nine approaches is "best," rather we aim to illustrate the process by which researchers might explore the implications of geomasking in a particular study context with particular research objectives.

The average level of anonymity provides insight into the utility of each method and the tradeoff between accuracy of NDVI measures and relative displacement. However, given that methods of anonymization assume homogeneous distribution of households, we extend the analyses further by establishing a minimum acceptable threshold of *k*-anonymity for this particular illustration of the process of geomasking. For example, a threshold of five is not met if there are less than five households located closer to the true location than the masked location. In this study, we find that at most, 1.2% of households do not meet the privacy standard of five households regardless of the method of anonymization. For a threshold of 10 households, only one method—$K_a = 20$ $K_b = 200$—provides a relatively high level of

Table 2 Percentage of Households That Do Not Meet Privacy Thresholds

Geographic masking methods	Number of Households $k_{act}=1$	% $k_{act}=1$	% $k_{act}<5$	% $k_{act}<10$
1. Random distribution	5	0.02%	0.3%	2.1%
2. Gaussian distribution	10	0.04%	0.5%	3.4%
3. Random density	10	0.04%	0.4%	2.7%
4. Gaussian density	11	0.05%	0.7%	4.6%
5. $K_a=5$ $Kb=50$	6	0.03%	0.4%	2.8%
6. $K_a=20$ $Kb=200$	1	0.00%	0.0%	0.2%
7. Within village	29	0.13%	0.4%	1.0%
8. Sub-village by infrastructure (roads)	89	0.39%	1.2%	2.7%
9. Sub-village by distance to village boundary	77	0.34%	1.0%	2.1%

K_{act} calculated as the number of households located closer to the true location than masked household location (averaged for each village)

anonymization with 99.8% of households meeting this privacy standard, while each method has at least one household that may be exposed to these privacy standards (Table 5). We recognize the disclosure of just one household is problematic as illustrated in Table 5, where all methods disclose at least one household. Yet, this analysis highlights that although masking methods may provide an average acceptable high level of protection (for example, approaches 7–9) it is important to consider how well a single household is protected. One response to this challenge is the removal of households that fail desired levels of anonymity from the dataset, of course with careful documentation. Another approach is to swap a specified number of households prior to displacement—or values within households—thereby adding additional uncertainty to lessen the likelihood of identification. While these approaches may not be acceptable in some cases, the study's investigators must balance the project's needs as related to the level of detail required—more detailed data requires a larger *k*-anonymity to minimize the likelihood of identification. (Table 2)

Discussion and conclusions

Substantial advancements in geographic information science have expanded the possibilities of scholarship linking the social and ecological worlds. Such scholarship is essential during this contemporary era of climate change in order to understand the processes shaping vulnerabilities.

The preliminary analysis presented here has been motivated by concern with facilitating socio-ecological research while preserving the confidentiality of social science survey respondents. Several studies have emerged over the past several years examining the potential error introduced by geomasking techniques, but most such research has examined error with regard to the creation of distance-based measures such as distance to health clinic (e.g., Warren et al., 2016). Results suggest that the creation of these distance-based covariates can yield consequential measurement error. But as opposed to distance-based measures, here we begin examination of the error introduced by geomasking for contextual measures reflecting proximate environmental conditions. We do so within the context of a health and demographic surveillance site, the Agincourt Health and Socio-Demographic Surveillance System in rural South Africa. The exploration suggests that in this study setting and for NDVI, geomasking approaches that use buffers and account for population density produce the most accurate results. However, the exploration also clearly demonstrates the tradeoff between accuracy and privacy, with more accuracy resulting in a higher level of potential respondent identification. In this way, the analysis demonstrates the critical tradeoff between geographic accuracy and confidentiality—a tradeoff that must be made carefully considered by study site administrators and research teams. It is important to remind readers that there are a variety of analytical choices that must be informed by specific study contexts, research questions, and the level of detail desired within the data.

While we focus on the Agincourt HDSS site, this is but one of nearly 50 such surveillance sites throughout sub-Saharan Africa, Asia, and Oceania. HDSS support research on the world's most pressing development questions, and they do so in regions where reliable and comprehensive data would typically not otherwise be available. Within these social surveillance systems, health, mortality, fertility, and migration are often focal topics. Socio-demographic characteristics such as age, gender, and employment status are also recorded. Combined across individuals, households and years, these data provide extraordinary opportunities to improve understanding of social and ecological processes and their connection with health including infant mortality, child health, and disease outbreaks. As such, the approaches tested here have the potential for far broader impact in understanding connections between human well-being and environmental change over longer time periods and in different settings.

In the long run, we aim to contribute to methodologies balancing research and privacy needs. As noted by Zandbergen (2014:9), "the lack of comparative analysis of masking techniques provides a clear indication for desirable future directions." Important next steps include examination of the error introduced within substantive analyses when using the NDVI estimates from anonymized household locations. Yet, the present manuscript offers an important first step in exploring anonymization prospects for health and demographic surveillance systems that are designed to facilitate essential population-health-environment in this contemporary era of global environmental change.

Appendix

Table 3 Estimated coefficients and significance of difference between true vs. displaced median NDVI estimates, nine illustrative geomasking approaches

Village (sorted by household size)	True household location	Geomasking within 300-m buffer zone		Geomasking within density-corrected 300-m buffer zone		Geomasking defined by *k*-anonymity		Random displacement within specific boundary		
		Random	Gaussian	Random (density)	Gaussian (density)	k_a=5 k_b=50	k_a=20 k_b=200	Within	Sub-village by infrastructure (roads)	Sub-village by distance to village boundary
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Constant	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.
Lillydale A	0.3508	*0.0000*	*– 0.0001*	*0.0000*	*0.0000*	*0.0000*	– 0.0004*	– 0.0006*	– 0.0004*	0.0002*
Agincourt	0.4003	– 0.0003*	*– 0.0001*	– 0.0004*	– 0.0003*	– 0.0005*	– 0.0005*	0.0015*	0.0006*	0.0011*
Kumani	0.3844	*0.0000*	*– 0.0001*	– 0.0001*	*– 0.0001*	*0.0000*	– 0.0004*	– 0.0007*	– 0.0003*	– 0.0005*
Belfast	0.3646	*0.0000*	*– 0.0001**	*– 0.0001*	*– 0.0001*	*0.0000*	– 0.0004*	– 0.0007*	– 0.0005*	– 0.0001*
Cunningmore A	0.4391	– 0.0003*	*– 0.0004**	– 0.0004*	*– 0.0001*	– 0.0004*	– 0.0013*	– 0.0016*	– 0.0007*	– 0.0004*
Justicia A	0.3769	0.0001*	*0.0000*	*0.0001*	*0.0001*	0.0002*	0.0002*	*0.0000*	*0.0005**	0.0004*
MP Stream	0.3853	*– 0.0001*	*0.0001*	*0.0000*	*0.0001*	*– 0.0001*	*0.0000*	0.0003*	0.0008*	*0.0000*
Kildare A	0.3603	– 0.0002*	*– 0.0002**	– 0.0004*	– 0.0002*	*– 0.0001*	– 0.0013*	– 0.0003*	0.0002*	– 0.0010*
Xanthia	0.4572	*– 0.0001*	*0.0000*	– 0.0003*	*– 0.0002*	*0.0000*	*– 0.0001*	0.0024*	0.0014*	0.0020*
Kildare B	0.3877	*0.0000*	*0.0000*	*0.0000*	*0.0001*	*0.0000*	*0.0000*	– 0.0008*	– 0.0002*	– 0.0003*
Ireagh A	0.3568	– 0.0001*	*– 0.0001*	*– 0.0001*	*– 0.0001*	*– 0.0001*	– 0.0002*	– 0.0001*	0.0001*	*– 0.0001*
Khaya Lami	0.3702	0.0001*	*0.0002**	*0.0000*	*0.0000*	*0.0000*	0.0002*	0.0023*	0.0001*	0.0025*
Cunningmore B	0.4006	– 0.0001	*0.0000*	*0.0000*	*0.0000*	*– 0.0001*	– 0.0001*	0.0003*	0.0003*	0.0001*
Huntington	0.3604	– 0.0001*	*– 0.0001*	*0.0000*	*– 0.0001*	– 0.0001*	– 0.0002*	0.0006*	0.0014*	*0.0000*
Croquetlawn	0.3791	– 0.0003*	*– 0.0001*	– 0.0002*	– 0.0003*	– 0.0003*	– 0.0011*	– 0.0015*	– 0.0015*	– 0.0019*
Newington B	0.3676	*0.0001*	*0.0000*	*0.0002*	*0.0001*	*0.0001*	*0.0001*	0.0008*	0.0007*	0.0005*
Somerset A	0.3608	*0.0000*	*0.0000*	0.0001*	*0.0000*	*0.0000*	0.0002*	0.0006*	0.0001	0.0002*
Makaringe	0.3959	*– 0.0001*	*0.0001*	*– 0.0001*	*0.0000*	*– 0.0001*	*– 0.0002*	0.0007*	– 0.0008*	*0.0000*
Lillydale B	0.3587	– 0.0001*	*0.0000*	– 0.0001*	*0.0000*	*0.0000*	0.0001*	– 0.0003*	– 0.0002*	*0.0000*

Table 3 (continued)

Village (sorted by household size)	True household location	Geomasking within 300-m buffer zone		Geomasking within density-corrected 300-m buffer zone		Geomasking defined by *k*-anonymity		Random displacement within specific boundary		
		Random	Gaussian	Random (density)	Gaussian (density)	k_a=5 k_b=50	k_a=20 k_b=200	Within	Sub-village by infrastructure (roads)	Sub-village by distance to village boundary
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Constant	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.	Coef.
Ireagh B	0.3618	_0.0000_	_0.0000_	_– 0.0001_	_0.0000_	_– 0.0001_	0.0001*	– 0.0001*	0.0000	– 0.0005*
Dumphries A	0.3685	0.0002*	_0.0001_	0.0002*	_– 0.0001_	0.0002*	_– 0.0001_	– 0.0005*	– 0.0003*	– 0.0003*
Rholane	0.3445	_0.0000_	_0.0002*_	_– 0.0001_	_– 0.0001_	– 0.0002*	– 0.0003*	– 0.0024*	– 0.0005*	– 0.0014*
Dumphries B	0.3516	0.0002*	_0.0000_	_0.0000_	_0.0001_	_0.0000_	_0.0000_	0.0023*	0.0006*	0.0026*
Newington C	0.3738	– 0.0002*	_– 0.0003*_	_– 0.0001_	_– 0.0001_	_– 0.0001_	– 0.0008*	– 0.0005*	– 0.0007*	– 0.0010*
Rolle C	0.3895	_– 0.0001_	_0.0001_	_0.0001_	– 0.0002*	_0.0000_	0.0004*	0.0005*	0.0005*	0.0003*
Justicia B	0.3590	_0.0000_	_0.0000_	_– 0.0001_	_0.0000_	_0.0001_	– 0.0002*	0.0003*	0.0000	_0.0000_
Ireagh C	0.3818	_– 0.0001_	_– 0.0002*_	_– 0.0001_	_0.0000_	_– 0.0001_	– 0.0002*	– 0.0007*	– 0.0005*	– 0.0009*
Dumphries C	0.3630	_0.0000_	_0.0000_	– 0.0002*	– 0.0002*	_– 0.0001_	– 0.0002*	0.0013*	0.0005*	– 0.0002*
Kildare C	0.3824	_0.0000_	_0.0000_	_0.0001_	_0.0000_	_0.0000_	_0.0000_	– 0.0004*	– 0.0003*	– 0.0004*
Somerset B	0.3513	_0.0000_	_0.0000_	_– 0.0001_	_0.0000_	_– 0.0001_	– 0.0003*	_– 0.0001_	_0.0003*_	– 0.0003*
Somerset C	0.3506	_– 0.0001_	_– 0.0001*_	_0.0000_	_0.0000_	_0.0000_	_0.0000_	_0.0000_	_– 0.0001*_	– 0.0001*

Cells highlighted in italics are non-significant difference between reference category (true household location)

*Results significant at $p < 0.05$

Table 4 Estimated coefficients and significance of difference between true and displaced sum NDVI estimates divided by the number of households, nine geomasking approaches

Village (sorted by household size)	True household location	Geomasking within 300-m buffer zone				Geomasking within density-corrected 300-m buffer zone				Geomasking defined by *k*-anonymity				Random displacement within specific boundary					
		Random		Gaussian		Random (density)		Gauss (density)		k$_a$=5 k_b=50		k_a=20 k_b=200		Within village		Sub-village by infrastructure (roads)		Sub-village by distance to village boundary	
		(1)		(2)		(3)		(4)		(5)		(6)		(7)		(8)		(9)	
	Constant	Coef.	% diff.	Coef.	% diff.	Coef.	% diff.	Coef.	% diff.	Coef.	% diff.	Coef.	% diff.	Coef.	% diff.	Coef.	% diff.	Coef.	% diff.
Lillydale A	1.672	− 0.2409*	14%	− 0.2392*	14%	*0.0011*	--	0.0022*	0%	*− 0.0006*	--	− 0.0024*	0%	0.0794*	5%	− 0.1729*	10%	− 0.1729*	10%
Agincourt	7.618	− 1.1791*	15%	− 1.1297*	15%	− 0.0428*	1%	*− 0.0077*	--	*− 0.0260*	--	− 0.1102*	1%	0.5034*	7%	− 0.8773*	12%	− 0.7734*	10%
Kumani	3.034	− 0.2494*	8%	− 0.2368*	8%	− 0.0465*	2%	− 0.0482*	2%	− 0.0380*	1%	− 0.1110*	4%	− 0.1671*	6%	− 0.2523*	8%	− 0.2948*	10%
Belfast	2.589	− 0.1248*	5%	− 0.1226*	5%	0.0050*	0%	*− 0.0007*	--	0.0038*	0%	− 0.0117*	0%	0.0697*	3%	− 0.1016*	4%	− 0.1044*	4%
Cunningmore A	3.899	− 0.5148*	13%	− 0.5309*	14%	− 0.0426*	1%	− 0.0568*	1%	− 0.0405*	1%	− 0.0328*	1%	0.5143*	13%	− 0.1725*	4%	− 0.1699*	4%
Justicia A	2.068	− 0.3545*	17%	− 0.3591*	17%	− 0.0115*	1%	*0.0020*	--	− 0.0099*	0%	− 0.0197*	1%	0.0447*	2%	− 0.3237*	16%	− 0.3541*	17%
MP Stream	5.262	− 0.3175*	6%	− 0.2758*	5%	0.0565*	1%	*0.0121*	--	*0.0217*	--	0.1121*	2%	− 0.0495*	1%	− 0.2775*	5%	− 0.4637*	9%
Kildare A	2.440	− 0.2761*	11%	− 0.2545*	10%	0.0126*	1%	0.0370*	2%	0.0326*	1%	0.0137*	1%	0.2321*	10%	− 0.1945*	8%	− 0.1913*	8%
Xanthia	8.300	− 1.3828*	17%	− 1.3600*	16%	− 0.2993*	4%	− 0.3408*	4%	− 0.2412*	3%	− 0.6898*	8%	− 0.1283*	2%	− 1.4285*	17%	− 1.5224*	18%
Kildare B	1.717	− 0.2323*	14%	− 0.2128*	12%	*0.0035*	--	0.0099*	1%	− 0.0109*	1%	− 0.0098*	1%	0.0936*	5%	− 0.1254*	7%	− 0.1831*	11%
Ireagh A	3.944	− 0.5195*	13%	− 0.5380*	14%	0.0119*	0%	− 0.0182*	0%	0.0263*	1%	*− 0.0038*	--	− 0.1344*	3%	− 0.5221*	13%	− 0.6382*	16%
Khaya Lami	2.137	− 0.2652*	12%	− 0.2658*	12%	0.0065*	0%	− 0.0225*	1%	0.0034*	0%	0.0222*	1%	0.0435*	2%	− 0.2356*	11%	− 0.2213*	10%
Cunningmore B	6.513	− 0.8091*	12%	− 0.7945*	12%	− 0.0431*	1%	− 0.0240*	0%	− 0.0664*	1%	− 0.1181*	2%	− 0.2445*	4%	− 0.8054*	12%	− 0.9193*	14%
Huntington	4.517	− 0.7698*	17%	− 0.7521*	17%	0.0107*	0%	*0.0070*	--	*0.0020*	--	*− 0.0078*	--	0.2225*	5%	− 0.4756*	11%	− 0.6784*	15%
Croquetlawn	5.647	− 0.9457*	17%	− 0.9310*	16%	− 0.0225*	0%	− 0.0327*	1%	− 0.0187*	0%	− 0.0091*	0%	− 0.0786*	1%	− 1.1095*	20%	− 1.0913*	19%

Table 4 (continued)

Village (sorted by household size)	True household location	Geomasking within 300-m buffer zone				Geomasking within density-corrected 300-m buffer zone				Geomasking defined by *k*-anonymity				Random displacement within specific boundary					
		Random		Gaussian		Random (density)		Gauss (density)		k$_a$=5 k_b=50		k_a=20 k_b=200		Within village		Sub-village by infrastructure (roads)		Sub-village by distance to village boundary	
		(1)		(2)		(3)		(4)		(5)		(6)		(7)		(8)		(9)	
	Constant	Coef.	% diff.	Coef.	% diff.	Coef.	% diff.	Coef.	% diff.	Coef.	% diff.	Coef.	% diff.	Coef.	% diff.	Coef.	% diff.	Coef.	% diff.
Newington B	3.941	− 0.6166*	16%	− 0.5596*	14%	− 0.0273*	1%	− 0.0878*	2%	− 0.0248*	1%	*− 0.0090*	--	0.1594*	4%	− 0.4944*	13%	− 0.3915*	10%
Somerset A	4.072	− 0.6815*	17%	− 0.6894*	17%	− 0.0036*	0%	*0.0005*	--	0.0094*	0%	0.0324*	1%	− 0.0128*	0%	− 0.6869*	17%	− 0.6713*	16%
Makaringe	8.833	− 0.5032*	6%	− 0.5079*	6%	− 0.2065*	2%	− 0.1654*	2%	− 0.0685*	1%	− 0.3279*	4%	*− 0.0293*	--	− 1.8175*	21%	− 0.6876*	8%
Lillydale B	3.756	− 0.7862*	21%	− 0.7673*	20%	− 0.0103*	0%	− 0.0048*	0%	0.0480*	1%	0.0809*	2%	− 0.0258*	1%	− 0.8560*	23%	− 0.8471*	23%
Ireagh B	7.489	− 1.0306*	14%	− 1.0086*	13%	− 0.0177*	0%	0.0257*	0%	− 0.0457*	1%	0.0283*	0%	0.1475*	2%	− 1.4010*	19%	− 0.9955*	13%
Dumphries A	7.837	− 0.2048*	3%	− 0.2565*	3%	0.0275*	0%	− 0.0336*	0%	*0.0076*	--	*− 0.0006*	--	− 0.7924*	10%	− 0.7068*	9%	− 0.9323*	12%
Rholane	2.217	− 0.5186*	23%	− 0.5180*	23%	*− 0.0011*	--	*− 0.0031*	--	*− 0.0044*	--	0.0079*	0%	0.0076*	0%	− 0.5476*	25%	− 0.4892*	22%
Dumphries B	7.609	− 0.1831*	2%	− 0.2533*	3%	0.0123*	0%	0.0437*	1%	0.0607*	1%	0.0758*	1%	0.2720*	4%	− 0.2000*	3%	− 0.0549*	1%
Newington C	7.071	− 1.1077*	16%	− 1.0880*	15%	− 0.0236*	0%	*0.0079*	--	− 0.1219*	2%	− 0.1560*	2%	− 0.2826*	4%	− 1.1515*	16%	− 1.3710*	19%
Rolle C	4.242	− 0.1476*	3%	− 0.1768*	4%	0.0660*	2%	− 0.0282*	1%	− 0.0785*	2%	− 0.1387*	3%	0.1621*	4%	*− 0.0037*	--	− 0.1618*	4%
Justicia B	2.269	− 0.4573*	20%	− 0.4563*	20%	− 0.0204*	1%	− 0.0048*	0%	− 0.0079*	0%	− 0.0611*	3%	− 0.0570*	3%	− 0.4655*	21%	− 0.5176*	23%
Ireagh C	3.105	− 0.4789*	15%	− 0.4908*	16%	− 0.0135*	0%	− 0.0343*	1%	− 0.0172*	1%	− 0.0833*	3%	− 0.2194*	7%	− 0.5732*	18%	− 0.6071*	20%
Dumphries C	14.193	− 0.1694*	1%	− 0.3498*	2%	0.2190*	2%	− 0.0635*	0%	0.3207*	2%	0.1589*	1%	− 0.0998*	1%	− 0.9658*	7%	− 0.3601*	3%

Table 4 (continued)

Village (sorted by household size)	True household location	Geomasking within 300-m buffer zone				Geomasking within density-corrected 300-m buffer zone				Geomasking defined by *k*-anonymity				Random displacement within specific boundary					
		Random		Gaussian		Random (density)		Gauss (density)		k$_a$=5 k_b=50		k_a=20 k_b=200		Within village		Sub-village by infrastructure (roads)		Sub-village by distance to village boundary	
		(1)		(2)		(3)		(4)		(5)		(6)		(7)		(8)		(9)	
	Constant	Coef.	% diff.	Coef.	% diff.	Coef.	% diff.	Coef.	% diff.	Coef.	% diff.	Coef.	% diff.	Coef.	% diff.	Coef.	% diff.	Coef.	% diff.
Kildare C	2.306	− 0.3552*	15%	− 0.3480*	15%	− 0.0155*	1%	− 0.0124*	1%	− 0.0071*	0%	− 0.0383*	2%	− 0.0155*	1%	− 0.3446*	15%	− 0.3590*	16%
Somerset B	4.302	− 0.9587*	22%	− 0.9604*	22%	0.0101*	0%	− 0.0321*	1%	*− 0.0043*	--	− 0.0304*	1%	− 0.2766*	6%	− 0.9694*	23%	− 0.9837*	23%
Somerset C	3.956	− 0.3275*	8%	− 0.4214*	11%	0.0426*	1%	− 0.0349*	1%	0.0431*	1%	0.0058*	0%	− 0.2267*	6%	− 0.3752*	9%	− 0.3191*	8%

Analysis using 1-km buffer zones indicates smaller differences between true and masked estimates. Overall, results were similar to 2-km buffer zones, with density-adjusted Gaussian methods and geomasking methods determined by *k*-anonymity producing the least differences between true and masked estimates. Cells highlighted in italics are non-significant difference between reference category (true household location)

% diff. percentage of coefficient divided by the coefficient for the reference category

*Results significant at $p < 0.05$

Table 5 Estimated distance and actual *k*-anonymity between NDVI estimates provided by seven geomasking approaches relative to the NDVI estimate for true household locations

Village (sorted by household size)	Geomasking within 300-m buffer zone				Geomasking within density-corrected 300-m buffer zone				Geomasking defined by *k*-anonymity				Random displacement within specific boundary					
	Random		Gauss		Random (density)		Gauss (density)		k$_a$=5 k_b=50		k_a=20 k_b=200		Within village		Sub-village by infrastructure (roads)		Sub-village by dist. to village boundary	
	(1)		(2)		(3)		(4)		(5)		(6)		(7)		(8)		(9)	
	Dist. (m)	k_{act}	Dist. (m)	k_{act}	Dist. (m)	k_{act}	Dist. (m)	k_{act}	Dist. (m)	k_{act}	Dist. (m)	k_{act}	Dist. (m)	k_{act}	Dist. (m)	k_{act}	Dist. (m)	k_{act}
MP Stream	220	25	188	19	365	58	312	44	306	44	598	134	2047	671	974	326	1549	541
Agincourt	222	28	189	21	318	49	271	38	284	42	561	119	2633	835	1616	472	2653	833
Kumani	222	37	190	28	271	50	229	38	261	48	523	137	1786	768	690	274	1729	728
Xanthia	222	31	188	24	314	55	271	43	283	48	549	132	1626	473	1127	318	1625	470
Cunningmore A	226	51	189	38	206	44	174	33	231	54	460	167	1389	717	867	443	1263	645
Newington B	223	21	186	15	399	51	340	39	316	36	626	109	1660	455	1177	314	1608	444
Lillydale A	223	59	190	44	161	33	136	24	205	51	400	170	1235	920	1059	749	1188	868
Kildare A	223	43	188	32	229	45	192	33	243	51	480	163	1235	595	830	392	1161	549
Belfast	223	59	190	43	162	33	136	24	203	50	401	168	1029	683	836	516	957	612
Justicia A	223	57	190	43	166	34	140	24	207	51	407	166	1046	658	695	409	1001	615
Cunningmore B	222	38	191	29	222	38	186	28	235	42	462	122	1191	399	916	301	1166	392
Ireagh A	219	40	189	32	195	33	167	26	219	41	442	122	1026	422	732	280	1015	409
Croquetlawn	222	43	190	34	211	40	182	31	228	46	456	135	1068	392	718	257	1075	392
Kildare B	222	54	190	41	166	32	141	24	205	48	408	162	973	563	669	377	895	500
Huntington	220	49	187	37	189	38	160	28	218	50	432	150	1014	429	863	365	995	412
Newington C	216	25	188	20	299	41	260	33	268	35	534	99	1187	246	1151	241	1159	245
Rolle C	221	31	186	23	264	40	226	31	258	40	492	109	987	231	851	196	1000	234
Dumphries A	224	44	189	33	204	38	176	29	226	45	447	123	810	253	424	120	751	232
Ireagh B	221	41	189	32	198	34	172	27	225	42	441	113	1388	302	796	218	1314	284

Table 5 (continued)

Village (sorted by household size)	Geomasking within 300-m buffer zone				Geomasking within density-corrected 300-m buffer zone				Geomasking defined by k-anonymity				Random displacement within specific boundary					
	Random		Gauss		Random (density)		Gauss (density)		k_a=5 k_b=50		k_a=20 k_b=200		Within village		Sub-village by infrastructure (roads)		Sub-village by dist. to village boundary	
	(1)		(2)		(3)		(4)		(5)		(6)		(7)		(8)		(9)	
	Dist. (m)	k_{act}	Dist. (m)	k_{act}	Dist. (m)	k_{act}	Dist. (m)	k_{act}	Dist. (m)	k_{act}	Dist. (m)	k_{act}	Dist. (m)	k_{act}	Dist. (m)	k_{act}	Dist. (m)	k_{act}
Dumphries B	220	42	188	32	209	38	178	29	227	45	437	116	1110	274	717	194	1062	257
Somerset A	219	53	189	40	140	24	119	17	193	43	367	132	679	339	555	265	666	334
Rholane	220	53	187	39	181	37	154	28	214	51	418	155	776	315	638	264	718	292
Lillydale B	222	51	188	38	170	32	144	24	205	45	401	127	764	285	645	235	705	256
Justicia B	221	35	190	28	236	39	201	30	239	39	447	94	1076	218	501	118	1060	217
Makaringe	223	27	188	20	353	51	306	42	301	41	579	103	1655	342	1401	292	1584	322
Dumphries C	222	35	188	27	232	37	198	29	238	38	462	92	729	135	638	118	722	134
Somerset B	223	47	187	35	207	41	179	32	232	51	438	144	492	142	387	111	472	142
Khaya Lami	219	151	186	116	57	15	48	11	121	57	233	167	606	504	421	370	601	499
Kildare C	219	49	187	37	164	30	140	23	198	42	395	128	500	151	525	158	477	140
Ireagh C	215	120	186	97	67	17	57	12	129	56	239	137	410	196	386	190	386	188
Somerset C	203	45	180	41	47	8	40	6	107	24	201	43	183	39	186	38	173	37

Dist (m) average distance in meters between true household location and moved household location, k_{act} number of households located closer to true location than masked household location (averaged for each village)

Acknowledgements Early versions of this manuscript were presented at the 2019 Annual Meeting of the Population Association of America, as well as the 2019 Conference on Demographic Responses to Changes in the Natural Environment. The latter was supported in part by an R13 grant from the Eunice Kennedy Shriver National Institute on Child Health and Human Development (#HD096853). We thank those who provided useful feedback at each venue.

Funding Funding for this research was provided by the University of Colorado Boulder's Research and Innovation Office. This research has also benefited from research, administrative, and computing support provided by the University of Colorado Population Center (Project 2P2CHD066613-06), funded by the Eunice Kennedy Shriver National Institute of Child Health and Human Development. The content is solely the responsibility of the author and does not necessarily represent the official views of the CUPC, NIH or CU Boulder. Indirect support was provided by the Wellcome Trust (Agincourt Unit, grant 085477/Z/08/Z) through its support of the MRC/Wits Rural Public Health and Health Transitions Research Unit. Indirect support was also provided by the University of Colorado Boulder's Earth Lab supported by CIRES and the Grand Challenge Initiative at CU Boulder. The authors would like to thank the communities, respondents, field staff, and management of the Agincourt Unit for their respective contributions to the production of the data used in this study.

References

Abowd, J. M., & Schmutte, I. M. (2019). An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review, 109*(1), 171–202.

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

Anane-Sarpong, E (2016). Application of ethical principles to research using public health data in the Global South: Perspectives from Africa. *Developing World Bioethics*.

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

Byers, E., Gidden, M., Leclère, D., Balkovic, J., Burek, P., Ebi, K., & Johnson, N. (2018). Global exposure and vulnerability to multi-sector development and climate change hotspots. *Environmental Research Letters, 13*(5), 055012.

Cassa, C. A., Wieland, S. C., & Mandl, K. D. (2008). Re-identification of home addresses from spatial locations anonymized by Gaussian skew. *International Journal of Health Geographics. 7*(1), 1-9.

Collinson, M. A. (2010). Striving against adversity: The dynamics of migration, health and poverty in rural South Africa. *Global Health Action, 3*(1), 5080.

Elkies, N., Fink, G., & Bärnighausen, T. (2015). "Scrambling" geo-referenced data to protect privacy induces bias in distance estimation. *Population and Environment, 37*(1), 83–98.

Foody, G. M., Cutler, M. E., Mcmorrow, J., Pelz, D., Tangki, H., Boyd, D. S., & Douglas, I. (2001). Mapping the biomass of Bornean tropical rain forest from remotely sensed data published by: Blackwell Publishing Stable http://www.Jstor.Org/Stable/2665383. *Global Ecology & Biogeography, 10*(4), 379–387.

Giannecchini, M., Twine, W., & Vogel, C. (2007). Land-cover change and human–environment interactions in a rural cultural landscape in South Africa. *Geographical Journal, 173*(1), 26–42.

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.

Hunter, L. M., Twine, W., & Patterson, L. (2007). ``Locusts are now our beef'': Adult mortality and household dietary use of local environmental resources in rural South Africa1. *Scandinavian Journal of Public Health, 35*(69_suppl), 165–174.

INDEPTH Network. (2017). "About Us"http://www.indepth-network.org/about-us.

Leyk, S., Maclaurin, G. J., Hunter, L. M., Nawrotzki, R., Twine, W., Collinson, M., & Erasmus, B. (2012). Spatially and temporally varying associations between temporary outmigration and natural resource availability in resource-dependent rural communities in South Africa: A modeling framework. *Applied Geography, 34*(2012), 559–568.

Lu, Y., Yorke, C., & Zhan, F. B. (2012). Considering risk locations when defining perturbation zones for geomasking. *Cartographica: The International Journal for Geographic Information and Geovisualization* 47(3):168–78.

LSRD. (2018). Land Satelite Data System (LSDS) Science Research and Development (LSRD) Reposiory. Sioux Falls, ND. U.S. Geological Survey (USGS) Earth Resources Observation and Science (EROS) Center. `https://espa.cr.usgs.gov.

Matsika, R., Erasmus, B. F. N., & Twine, W. C. (2013). Double jeopardy: The dichotomy of fuelwood use in rural South Africa. *Energy Policy, 52,* 716–725.

Mutanga, O., & Skidmore, A. K. (2004). Narrow band vegetation indices overcome the saturation problem in biomass estimation. *International Journal of Remote Sensing, 25*(19), 3999–4014.

NASA. (2000). Measuring Vegetation (NDVI & EVI). *Measuring Vegetation (NDVI & EVI).*

Olsson, L., Opondo, M., Tschakert, P., Agrawal, A., & Eriksen, S. E. (2014). Livelihoods and poverty. In: Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Field, C.B., V.R. Barros, D.J. Dokken, K.J. Mach, M.D. Mastrandrea, T.E. Bilir, M. Chatterjee, K.L. Ebi, Y.O. Estrada, R.C. Genova, B. Girma, E.S. Kissel, A.N. Levy, S. MacCracken, P.R. Mastrandrea, and L.L.White (Eds.), Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 793–832.

Paumgarten, F., & Shackleton, C. M. (2011). The role of non-timber forest products in household coping strategies in South Africa: the influence of household wealth and gender. *Population and Environment, 33*(1), 108.

Roerink, G. J., Menenti, M., Soepboer, W., & Su, Z. (2003). Assessment of climate impact on vegetation dynamics by using remote sensing. *Physics and Chemistry of the Earth, 28*(1–3), 103–109.

Ruggles, S., Fitch, C., Magnuson, D., & Schroeder, J. (2019). Differential privacy and census data: Implications for social and economic research. *AEA Papers and Proceedings, 109,* 403–408.

Sumner, D., Christie, M. E., & Boulakia, S. (2017). Conservation agriculture and gendered livelihoods in Northwestern Cambodia: Decision-making, space and access. *Agriculture and Human Values, 34*(2), 347–362.

Sweeney, L. (2002). *k*-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10*(5), 557–570.

Tlou, B., Sartorius, B., & Tanser, F. (2017). Space-time patterns in maternal and mother mortality in a rural South African population with high HIV prevalence (2000–2014): Results from a population-based cohort. *BMC Public Health, 17*(1), 543.

Tucker, C. J. (1979). *Red and Photographic Infrared l,Lnear Combinations for Monitoring Vegetation*. Vol 8.

Wang, J. & Rich, P. M. (2008). Geocarto International Relations between NDVI, Grassland Production, and Crop Yield in the Central Great Plains.

Wang, H., & Reiter, J. P. (2012). Multiple imputation for sharing precise geographies in public use data. *The Annals of Applied Statistics, 6*(1), 229–252.

Warren, J. L., Perez-Heydrich, C., Burgert, C. R., & Emch, M. E. (2016). Influence of demographic and health survey point displacements on distance-based analyses. *Spatial Demography, 4*(2), 155–173.

Wessels, K. J., Prince, S. D., Frost, P. E., & Van Zyl, D. (2004). Assessing the effects of human-induced land degradation in the former homelands of Northern South Africa with a 1 km AVHRR NDVI time-series. *Remote Sensing of Environment, 91*(1), 47–67.

Wisely, S. M., Alexander, K., & Cassidy, L. (2018). Linking ecosystem services to livelihoods in southern Africa. *Ecosystem Services, 30,* 339–341.

Zandbergen, P. A. (2014). Ensuring confidentiality of geocoded health data: Assessing geographic masking strategies for individual-level data. *Advances in Medicine,* 1–14.

Zhou, F. D., & Louis, T. A. (2010). A smoothing approach for masking spatial data. *The Annals of Applied Statistics, 4*(3), 1451–1475.

Authors and Affiliations

Lori M. Hunter[1,2] · **Catherine Talbot**[1,2] · **Wayne Twine**[3] · **Joe McGlinchy**[5] · **Chodziwadziwa W. Kabudula**[4] · **Daniel Ohene-Kwofie**[4]

[1] CU Population Center, Institute of Behavioral Science, University of Colorado Boulder, Boulder, USA

[2] Department of Sociology, University of Colorado Boulder, Boulder, USA

[3] School of Animal, Plant and Environmental Sciences, University of the Witwatersrand, Johannesburg, South Africa

[4] MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[5] Earth Lab, Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, USA