TU DORTMUND

SEMINAR: ANONYMIZATION OF GEOREFERENCED DATA

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

Authors: Lori M. Hunter, Catherine Talbot, Wayne Twine, Joe McGlinchy, Chodziwadziwa W. Kabudula, Daniel Ohene-Kwofie

Lecturer:
Prof. Dr. Rainer Lenz

Divya Prima Crasta

February 28, 2024

Contents

1 Introduction 1

2 Background 1

3 Data and Methods 2

3.1 Data 2

3.2 Methods 3

3.2.1 NDVI 3

3.2.2 Geographic masking approaches 3

4 Village descriptive profiles 5

5 Understanding implications of geomasking 6

6 Tradeoffs between privacy and accuracy 8

7 Summary 9

1 Introduction

Population-health-environment research is gaining importance these days due to rapid global climate change. Specifically in the areas where people live with close proximity to environment, the research that links environment and health is important. But conducting such a research requires to access the geographic location of the respondents. This can risk their privacy. Therefore it was required to develop geographic masking techniques. The physical locations of households are displaced to protect privacy. There are several methods used to for this purpose. It is intended to explore some of these in this project. The geographic masking techniques should be such that they protect privacy of respondents and provide sufficiently accurate results at the same time. This is a trade off between accuracy and privacy. Which is also examined in this project.

The data was collected from South Africa's Agincourt Health and Socio-Demographic Surveillance Site(AHDSS). Here Normalized Difference Vegetation Index (NDVI) was used as environmental measure to check for tradeoff between accuracy and privacy.

It was found out that the methods that use buffers with consideration of population density provide accurate results for NDVI values. High accuracy resulted in lower privacy protection and low accuracy resulted in high privacy protection.

In the second section of this report background of surveillance site and privacy policies are presented. In the next section research setting is introduced and various geomasking techniques are discussed. Following this, some peculiarities of the villages are discussed. In the next section, the results obtained after employing these techniques is analysed. Next, the tradeoffs between accuracy and privacy is discussed for various geomasking approaches. Finally, the report is summarized with presentation of central results and further investigation topics.

2 Background

In many areas of Global South, people live with close proximity to nature. Their livelihoods depend highly on natural resources. Which is why changes in climate can effect their health directly. Due to this understanding the impact of climate change on lives of people is important so that researches can inform policy makers about this. Research linking environment and health can help make the policies sustainable. But conducting such research requires using very personal information such physical location, age, gender, ethnicity, health status etc. This is a challenge to privacy of repondents. The examples of such situations include environmental research connecting weather conditions and household locations, disease mapping etc. In this project, the focus is on NDVI values which reflects the natural resource availability in a particular location.

There are several approaches used to protect data confidentiality. At some instances data privacy is protected through highly confidential agreements that particular researchers have to sign. They also need to travel to data centers for access and publication is also subject to constraints. US Census Bureau's network of Federal Statistical Research Data Centers (FSRDC) is an example for this. Other approaches includes making data available with less precision and aggregation. But aggregation can compromise with the quality of analyses. The data can also be represented by weighted averages. It is called as spatial smoothing. Other methods include multiple imputation, linear programming, data swapping and synthetic data. In linear programming noise is added based on some probabilities. In differential privacy additional simulated rows are added to the data. This can misrepresent the data and thereby risk the privacy of minorities.

In this project, the focus is on geomasking approaches where the physical locations are displaced. These approaches have high potential in demographic analysis.

3 Data and Methods

3.1 Data

The data was collected from Agincourt Health and Socio-Demographic Surveillance System(AHDSS). It includes physical location of approximately 22,708 households which belong to 31 separate villages. The residents in these locations live in close proximity to their surroundings. They are highly dependent on natural resources. Which is why changes in environment and availability in natural resources has direct impact on health of respondents in this area. Therefore, exploring availability of resources in this site is important. And this requires accessing physical locations of the households. In this project we aim to examine various geographic masking techniques and understand it's implications on values of measure used to quantify natural resources availability. It is also intended to analyse the trade offs between accuracy and privacy for various techniques used. It shall be noted that, these results are specific to this context.

3.2 Methods

3.2.1 NDVI

Normalized Difference Vegetation Index(NDVI) is a measure used to quantify natural resource availability. Vegetation biomass and productivity are positively correlated with NDVI. It's values ranges from -1 to 1. If the value is ≤ 0.1, it implies barren land, rock, sand or water. Values between 0.2 and 0.3 indicates shrublands or grasslands. High values (0.6-0.8) indicates temperate or tropical rainforests.

The NDVI values were obtained from Landsat 5,7 and 8 missions. Data was processed by USGS Earth Resources Observation and Science (EROS) Center. There were QA pixel files available. For this project, data from March 1997 to December 2017 is considered. Regions covered by "Cloud", "Cloud Shadow" or "Water" was excluded from analyses. Households in village boundaries do not represent the area where resource collection occurs. Therefore they were not included. The area near reserves were also not included as villagers do not have access to them.

In this project we consider sum and median NDVI values. A buffer zone of 2 km was used to estimate the NDVI value at at location. The data comprises of NDVI values from March 1997 to December 2017. Median NDVI values of a buffer zone are taken as measure of central tendency for the household. Also, sum of NDVI values of the households within a buffer zone divided by number of households is considered. The aggregations are performed for average 10 months per year. Therefore there are around 200 estimates for each of the 22,708 households.

3.2.2 Geographic masking approaches

In this project we implement nine geographic masking techniques. Four of these are "donut" approach. In donut approach the locations of household are displaced within a minimum and maximum radius. These approaches are random displacement, offsets that represent gaussian distributions of displacement, gaussian random displacement and gaussian displacement with density/distance factor. Here, these approaches are carried out with minimum distance 150m and maximum distance 300m. The minimum and maximum distance are constrained to village boundaries. That is, if any household's displaced location is outside the village boundary, it is displaced to village boundary.

Due to this, households in smaller and long, narrow villages have smaller potential for displacement.

We also consider density factor for the displacement. The density factor is multiplied to adjust the amount displacement. The total density multiplier is given by:

$$Totaldensitymultiplier = \frac{Averagetotalhouseholddensity}{Villagehouseholddensity}$$

There is also another approach where k anonymity factor is included. k-anonymity is the number of individuals needed within buffer zone to preserve confidentiality. For a village with area A_i and number of households N_i the minimum distance(R_{ai}) and maximum distance(R_{bi}) in this approach is given by,

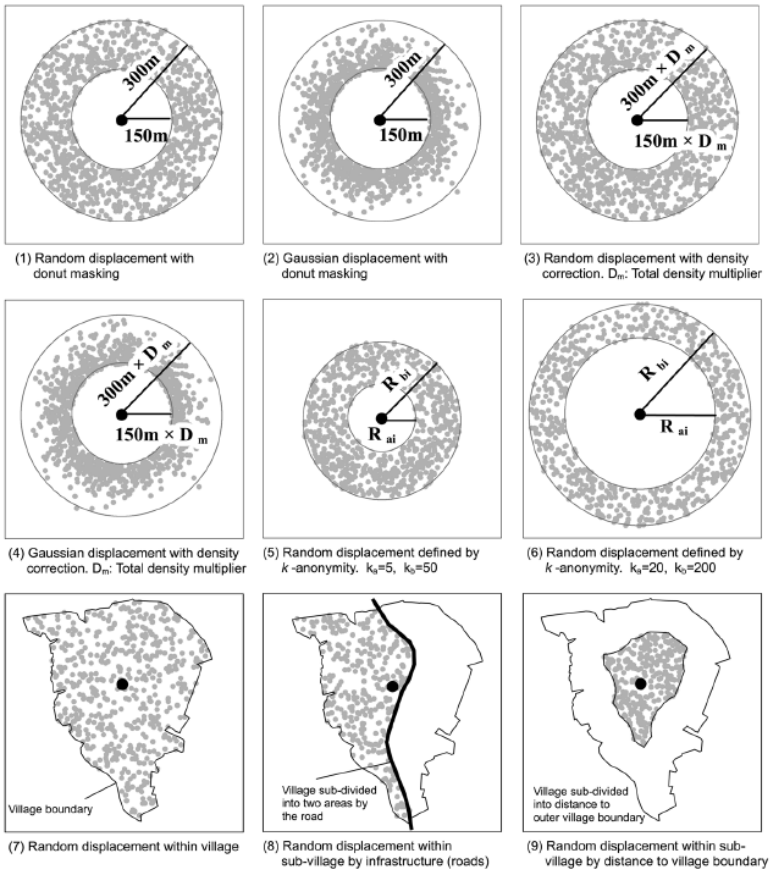
$$R_{ai} = \sqrt{(\frac{A_i}{\pi} * \frac{k_a}{N_i})}$$

$$R_{bi} = \sqrt{(\frac{A_i}{\pi} * \frac{k_b}{N_i})}$$

Here, k_a is the minimum number of individuals that are in close proximity with true location than displaced location. It was decided to take $k_b = 10 * k_a$. In this study $k_a = 5$ and $k_a = 20$ are considered.

Other approaches include displacement within entire village, displacement within regions of villages separated by geographical boundaries such as water, roads etc. and displacement within region formed by taking 100m distance from edge of village boundary.

Various approaches are summarized in Figure 1.


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

Figure 1

4 Village descriptive profiles

It can be noted that there are variations in size and density of villages. For example Somerset C has relatively low geographical area. Displacement using large distance is not practical for villages with small area. It's population is also low. Thus, using small distance also holds risk to privacy.

On the other hand, the villages in western portion have higher NDVI values compared to villages in eastern portion. The lower NDVI could be due to sharing boundary with fenced reserves. The west-east greenness gradient is depicted in the Figure 2.

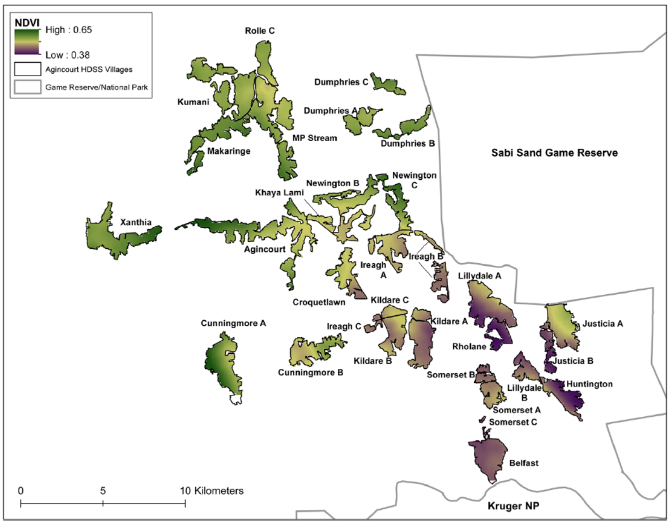


Fig. 4 NDVI values (Jan 2010) for true household locations, 2 km buffers, MRC/Wits-Agincourt Unit

Figure 2

5 Understanding implications of geomasking

To understand the implications of geomasking techniques, the NDVI values estimated from displaced location are compared with NDVI values obtained from true location. It was also necessary to explore the variation of NDVI values within and between villages. It is noted that villages in western region of study site have higher variation compared to those in easter region. The villages in western regions are mostly large and less dense. But there is no established strong correlation between variation of NDVI values within villages and their density. It can also be noted that village Xanthia has especially high NDVI value even though it has highest density.

Further, linear mixed regression models with fixed effects representing different methods of randomization was performed. If the estimated coefficients of fixed effects are significant, then we can say that there are important differences in values of NDVI. In this method, it is possible to determine the source of deviation of NDVI values. The sources of deviation could be high or low NDVI values or variation in NDVI values within and between villages.

It was found out that there was consistent lack of significant differences for three methods. The significant differences in coefficients was relatively low for median NDVI estimates than sum NDVI estimates. But the differences are less than 4%. Higher number of significant differences for sum NDVI estimates could be due to higher sensitivity of sum to displacement. For instance, the households close to village boundary have higher sum NDVI than those far from boundary. The methods for which NDVI estimates are more similar are, random, Gaussian and k-anonymity with density factor. It was also found out that density factor is more limiting factor that contributes to differences in estimates than proximity to village boundary.

It is also found out that villages such as Rholane, Somerset B, and Lilydale B simply miss the mark for almost all masking techniques. Moreover, the villages at central and eastern portions of study site have higher differences in estimates compared to those in west. All these findings is summarized in the Figure 3.

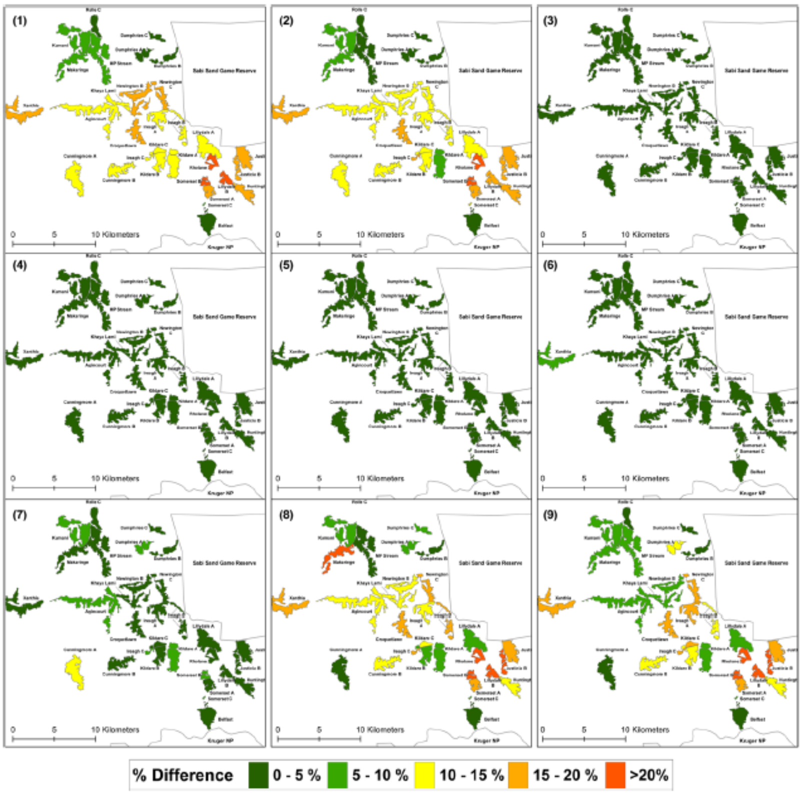


Fig. 5 Spatial distribution of difference in sum of NDVI values/number of households, true vs. geo-masked locations, MRC/Wits-Agincourt Unit (1) Random, donut, (2) Gaussian, donut, (3) Random (density), donut, (4) Gaussian (density), donut (5) ka=5 kb=50, donut, (6) ka=20 kb=200, donut, (7) Random within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

Figure 3

To explore the factors that can contribute to differences in estimates, underlying distributions of NDVI, population and density were examined. Also, other potential factors such as proximity to village boundary, resources and edge of study site were considered. It is seen that the villages with high NDVI values have smaller differences in

NDVI estimates. Whereas, the villages with less density have higher differences for the approaches with density factor. This could be because, households with lower density needs to be displaced at larger distance to protect privacy. And displacement to larger distance reduces accuracy. The villages with smaller range of NDVI values also show better agreement. Belfast and Somerset C are the examples of villages which are small and have small ranges of NDVI. These villages have similar values for NDVI estimates. Another reason Belfast shows better agreement is because it located at the edge of study site.

Thus based on above inferences, it can be said that population, density, range and value of NDVI, proximity of villages to boundary, resources plays a role in accuracy after displacement. The degree of influence of these factors is different for each of the geomasking methods. Thus the choice of geomasking method must be made with consideration of these factors. Hence the evaluation of various geomasking approaches is specific to the research context.

6 Tradeoffs between privacy and accuracy

The tradeoff between privacy and accuracy was explored using k-anonymity. k-anonymity is used to quantify the privacy. Higher k-anonymity indicates higher privacy and lower k-anonymity indicates lower privacy. The average displacement distance was plotted against average k-anonymity which is given in Figure 4 a). Additionally, the average difference in NDVI was plotted against average k-anonymity which is given in Figure 4 b). It can be seen that the approaches with high average displacement distance have high k-anonymity and those with high k-anonymity have high difference in NDVI. Likewise, the approaches with low average displacement distance have low k-anonymity and those with low k-anonymity have low differences in NDVI. Whereas, the for approach with $K_a = 5 and K_b = 50$ and $K_a = 20$ and $K_b = 200$ the k-anonymity is medium. But the average NDVI differences is relatively low.

It is important to note that the best approach is specific to research context. In this project it is only aimed to explore the process of determining most suitable approach. The analyses can be further extended by setting up threshold a for k-anonymity. It is seen that if 10 is the threshold, for approach with $K_a = 20$ and $K_b = 200$, 99.8% of the households meet this privacy policy and it provides highest anonymization. For other approaches at least one household does not meet this criteria. Exposure of just one

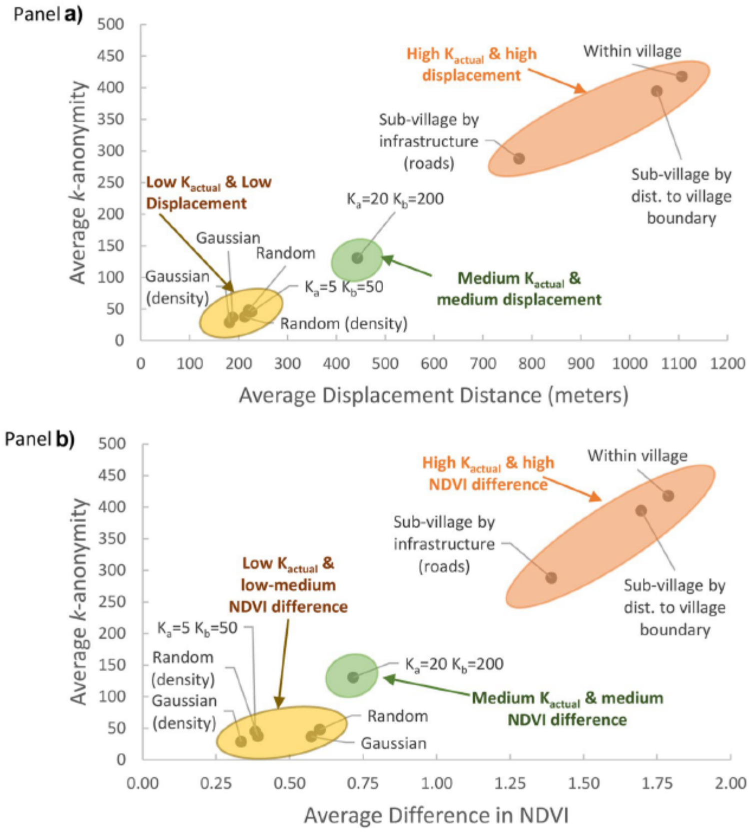


Fig. 6 Evaluation of accuracy and privacy for true vs. geomasked locations, MRC/Wits-Agincourt Unit. **a** Displacement distance versus *k*-anonymity. **b** *k*-Anonymity versus absolute difference in NDVI

Figure 4

household is very risky. Other methods can be employed to protect privacy. Other methods are swapping locations before displacement or swapping values for few observations. The choice of privacy policy is also specific to the context.

7 Summary

In this project it was intended to implement geomasking techniques to physical locations of households in Agincourt Health and Socio Demographic Surveillance System in South Africa. NDVI was environmental measure used to explore the tradeoff between privacy and accuracy.

Four donut approaches, density based approach, approach with sub village boundary, entire village were the geomasking techniques that were explored. It was seen that the approach that accounts for density and k-anonymity provide sufficient(in this context) anonymity with lower differences in estimated NDVI values. The tradeoff between privacy and accuracy is critical. Also, the population, density, distribution of values, proximity to boundary, resources etc. influences the choice of most suitable geomasking approach. The most suitable approach and evaluation are specific to context of the research.

The approaches discussed here have vast potential in researches related to population health. As further investigations, the error of NDVI estimates can be evaluated for various approaches. The results and analyses presented here serve as first step to further opportunities in anonymization of georeferenced data.