Seminar: Anonymization of Georeferenced Microdata

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, ·oe McGlinchy, Chodziwadziwa W. Kabudula, Daniel Ohene-Kwofie

Presenting: Divya Prima Crasta
Lecturer: Prof. Dr. Rainer Lenz

Table of Contents

Introduction

- Climate change is one of the most discussed topic in today's world.
- The individuals living with close proximity to environment are directly effected by climate changes. The effects can be understood by population-health-environment research.
- Such understandings are essential to inform programs and policies.
- It usually requires the geographical location of households.
- But acquiring such information violates privacy.
- Therefore various geographic masking techniques have been developed.

- Here we consider the data from South Africa's Agincourt Health and Socio-Demographic Surveillance Site.
- The population in this region is highly reliant on natural resources.
- As an measure that reflects natural resources, we consider Normalized Difference Vegetation Index(NDVI).
- We are interested in how anonymization effects the estimates of environmental measures.
- In this project, we compare the values of NDVI values calculated from true locations and that calculated after geomasking.
- It is also aimed to examine the trade offs between accuracy and privacy.
- It must be noted that the analyses presented here are context-specific.

Data

- The data is collected from Agincourt Health and Socio-Demographic Surveillance System(AHDSS).
- The AHDSS provided physical locations of 22,708 households which belong to 31 separate villages.
- NDVI values are derived from Landsat 5, 7, and 8 missions and processing completed by USGS Earth Resources Observations and Science (EROS) Center.
- In this project we consider data from March 1997 to December 2017.
- To deal with seasonality, 200 estimates for each households for average 10 months per year, 1997 to 2017 was incorporated.

- Additional preprocessing:
  - “Cloud Shadow”, “Cloud” or “Water” were excluded.
  - Areas within village boundaries were excluded as they do not represent areas of resource collection.
  - Neighbouring game reserves and parkland were excluded since village residents do not have access to them.

NDVI:

- Positively correlated with vegetation biomass and productivity.
- Values range from -1 to +1.
- Low values (≤ 0.1) : indicate barren land, rock, sand or water.
- Moderate values (0.2-0.3) : indicate shrublands or grasslands
- High values (0.6-0.8) : tropical rainforests.

- NDVI associated with a household point location is obtained within 2km buffer zone.
- 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.
- Measure of household resource availability that is the sum of the NDVI values divided by the number of household in each individual household buffer zone.

Methods

- We employ nine geomasking techniques. Four of these are “donut” approach.
- Donut approach is where displacement occurs within a minimum and maximum distance. They are,
    1. Random displacement
    2. Offsets that represent Gaussian distributions of displacement
    3. Gaussian random displacement
    4. Gaussian displacement with a “distance/density factor”
- Here displaced between 150 m to 300 m.

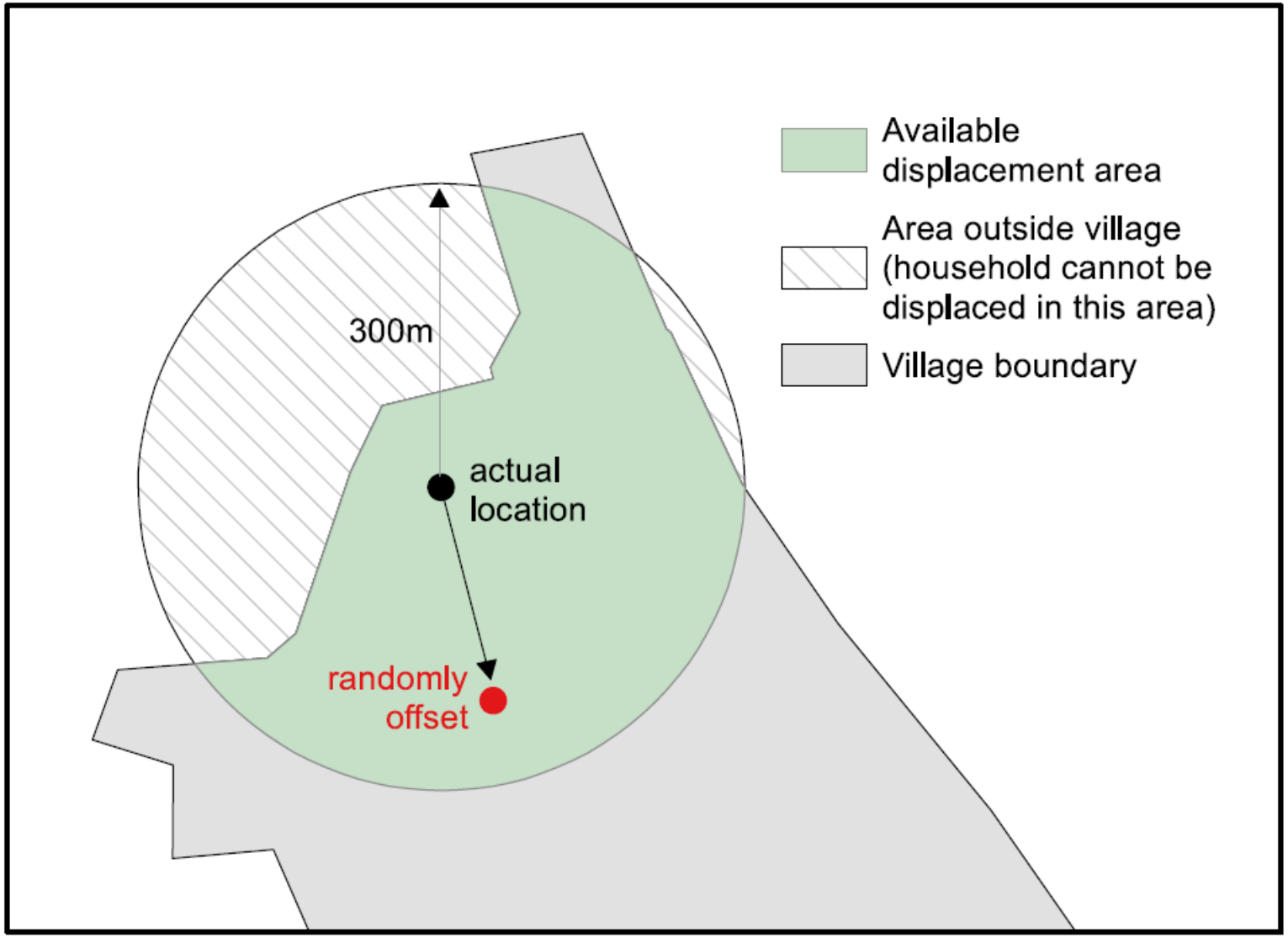

Fig. 1 Example of a simple random offset for a households with available displacement area constrained by village boundary

Displacement distance adjustment:

Distance/density factor:

- Density is taken into account for weighted displacement approaches.
- More dense villages require less displacement for privacy protection.

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

Distance adjustment with k-anonymity consideration:

- k-anonymity: It is the number of individuals(or households) needed within buffer to preserve confidentiality. It is defined as number of households that are closer to true location than masked.

- Minimum and maximum displacement distance is defined by the density of households and user-defined levels of k-anonymity as:

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

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

where $N_i =$ number of households in each village, $A_i =$ Area
$k_a =$ minimum displacement threshold and $k_b =$ maximum displacement threshold.

Typically, $k_b = 10 * k_a$

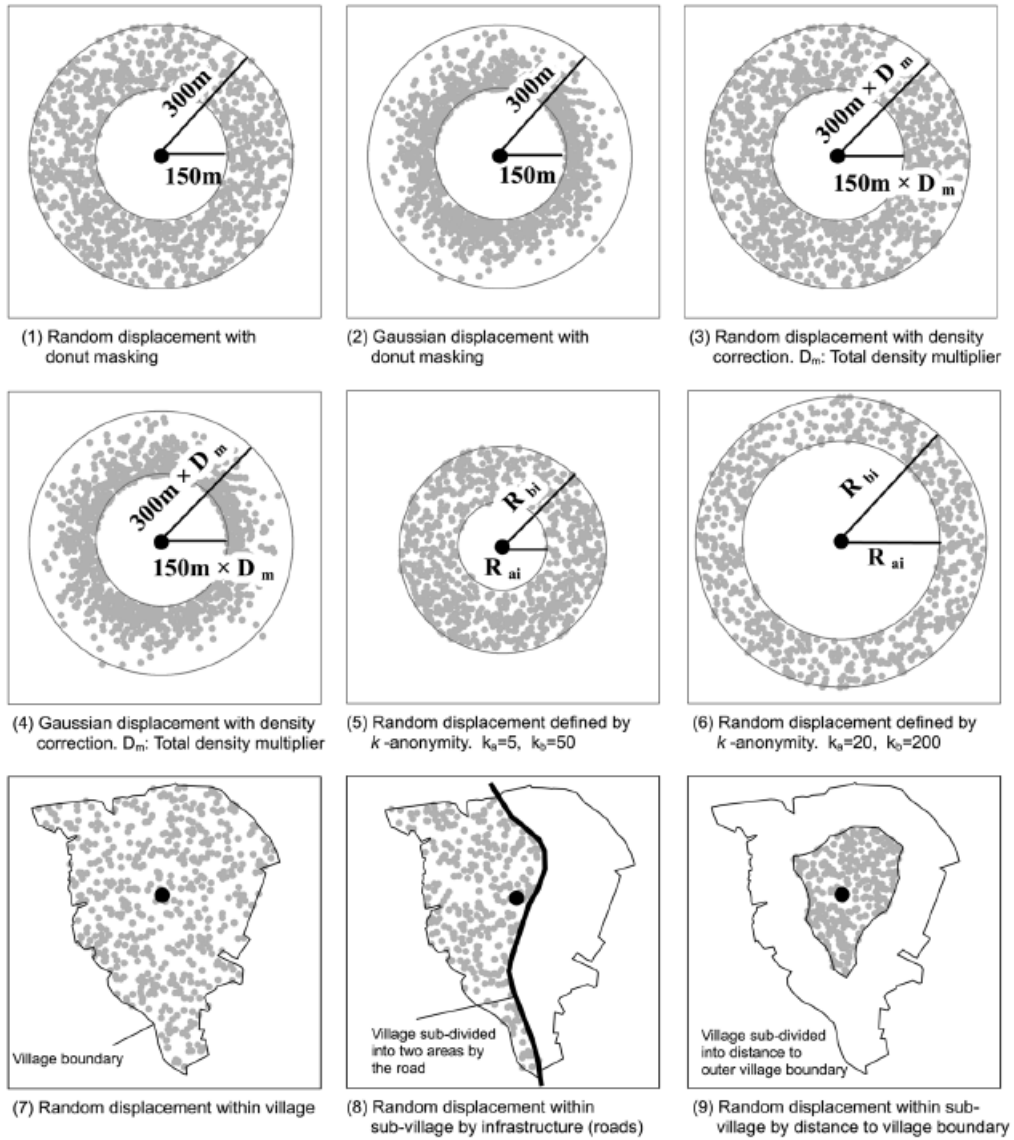

Fig. 3 The nine illustrative geographic masking techniques examined. Black circle represents the original location; gray dots represent simulated possible locations using each masking method with a radius of 300 m and exclusion zones of 150 m for masking approaches 1 and 2. For approaches 3 and 4, the 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

- Other approaches : Displacement within-village geographic clusters such as:
    - The entire village
    - Sub-village areas defined by physical boundaries such as roads, rivers or railroads.
    - Sub-village boundaries defined by distance from edge of village boundary in 100m buffer zones.

<u>Village descriptive profiles</u>

- Large distance not practical for small geographic area(eg. Somerset C)
- Also, low population is challenge to privacy.
- West-east greenness gradient in the study area.

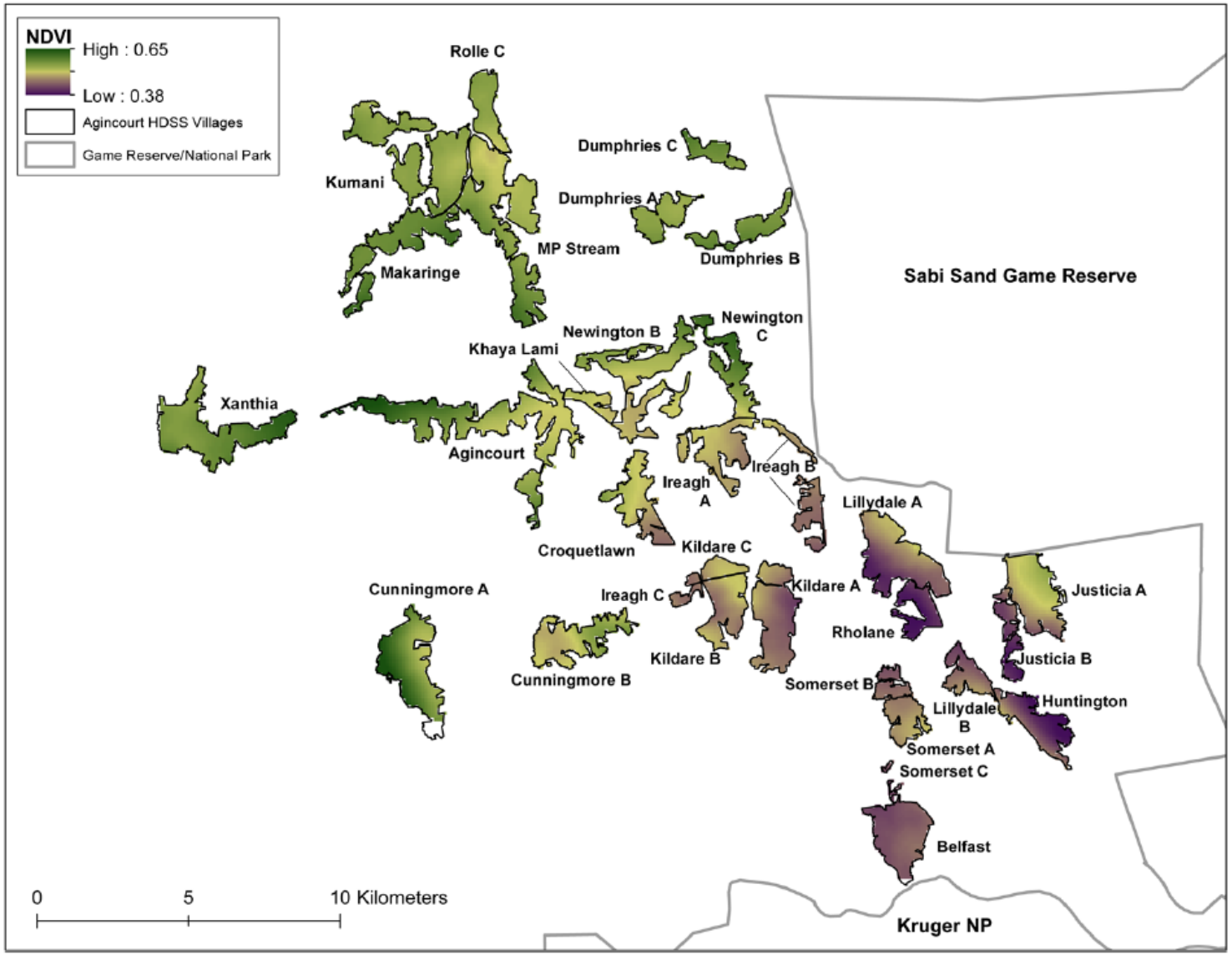

Fig. 4 NDVI values (Jan 2010) for true household locations, 2 km buffers, MRC/Wits-Agincourt Unit

Understanding implications of geomasking

- Significance difference in NDVI estimates obtained from true household locations and that from masked locations are examined.
- Sum and median NDVI values are considered for comparison.
- Only growing season months September to April are considered.
- The NDVI estimated from true household varied widely in villages in western portion. These villages are relatively large and less dense.
- Even so, Xanthia, the village with highest density has especially high NDVI likely due to topography and proximity to water sources.

- More thorough sense of significance of difference is obtained by using linear mixed regression models with fixed effects.
- If coefficient is significantly different, that implies important differences in NDVI estimates.
- Analyses were conduced for each village to estimate variation of randomization methods at the village scale. This helps us understand potential sources of deviations.

<u>Important findings</u>

1. Methods that account for density(random, Gaussian, and k-anonymity) provide results most similar to median NDVI estimates of true household location as demonstrated by consistent lack of statistically significant differences.

- There are greater number of statistically significant differences for sum NDVI estimates (but the differences are small(< 4 %)).
- This could be due to large variations of sum estimates. Sum is more sensitive to position relative to village boundary.

2. Few villages, notably Rholane, Somerset B, and Lilydale B simply miss the mark for all of the geomasking approaches.
    - These patterns of error can be spatially represented as in Fig. 5
    - High degree of variation depending on method used.
    - central and eastern villages show large differences.

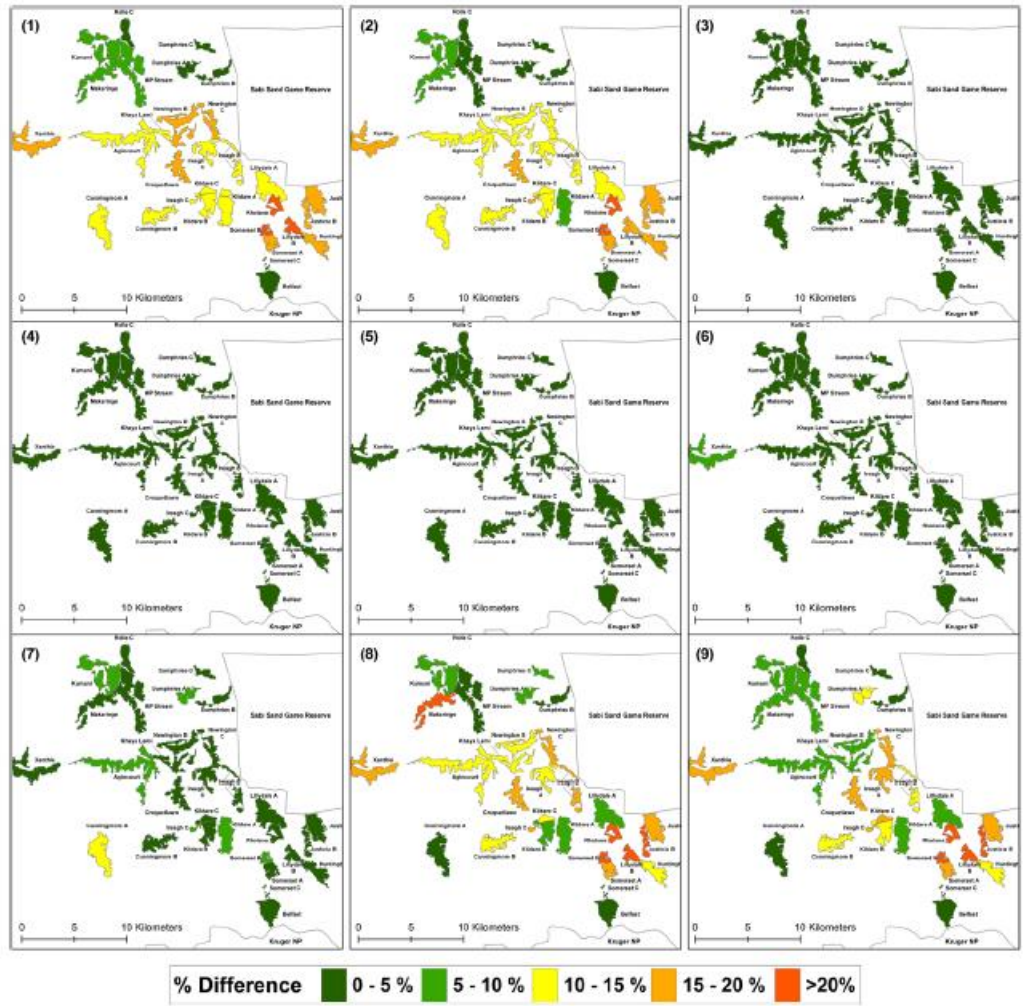

Fig. 5 Spatial distribution of difference in sum of NDVI values/number of households, true vs. geo-masked locations, MRC/Wits-Agincourt Unit (1) Random, donut, (2) Gaussian, donut, (3) Random (density), donut, (4) Gaussian (density), donut (5) ka=5 kb=50, donut, (6) ka=20 kb=200, donut, (7) Random within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

Further explorations:

- Underlying distributions of NDVI, population, and density as well as village proximities were examined.
- Irrespective of masking method, smaller and densely populated villages tend to have smaller variation. This is because they tend to have relatively low displacement distances.
- Some villages in northern part have less differences. These generally exhibit higher levels of NDVI compared to middle part of the study.
- Two southern villages, Belfast and Somerset C also show better agreement. These are relatively small and have small ranges of NDVI. Also, Belfast is located at the edge of study site.

In all, spatial variation in error is influenced by NDVI variability, population, and/or population density, household proximity to village edges, other households, other villages and protected areas.

Understanding their influence is important to decide which geomasking method is most appropriate for a particular research project.

Tradeoffs between anonymity and accuracy

- Illustrated using calculated average displacement between true and masked locations and estimated k-anonymity for each location averaged within villages.

  Recall: k- anonymity is the number of households that are closer to the true location than masked.

- Summarized in Fig. 6
- Approaches with lower displacement distances have lower k-anonymity. Density informed methods provide smaller displacement distances and lower levels of k-anonymity.
- K-anonymity methods have medium levels.

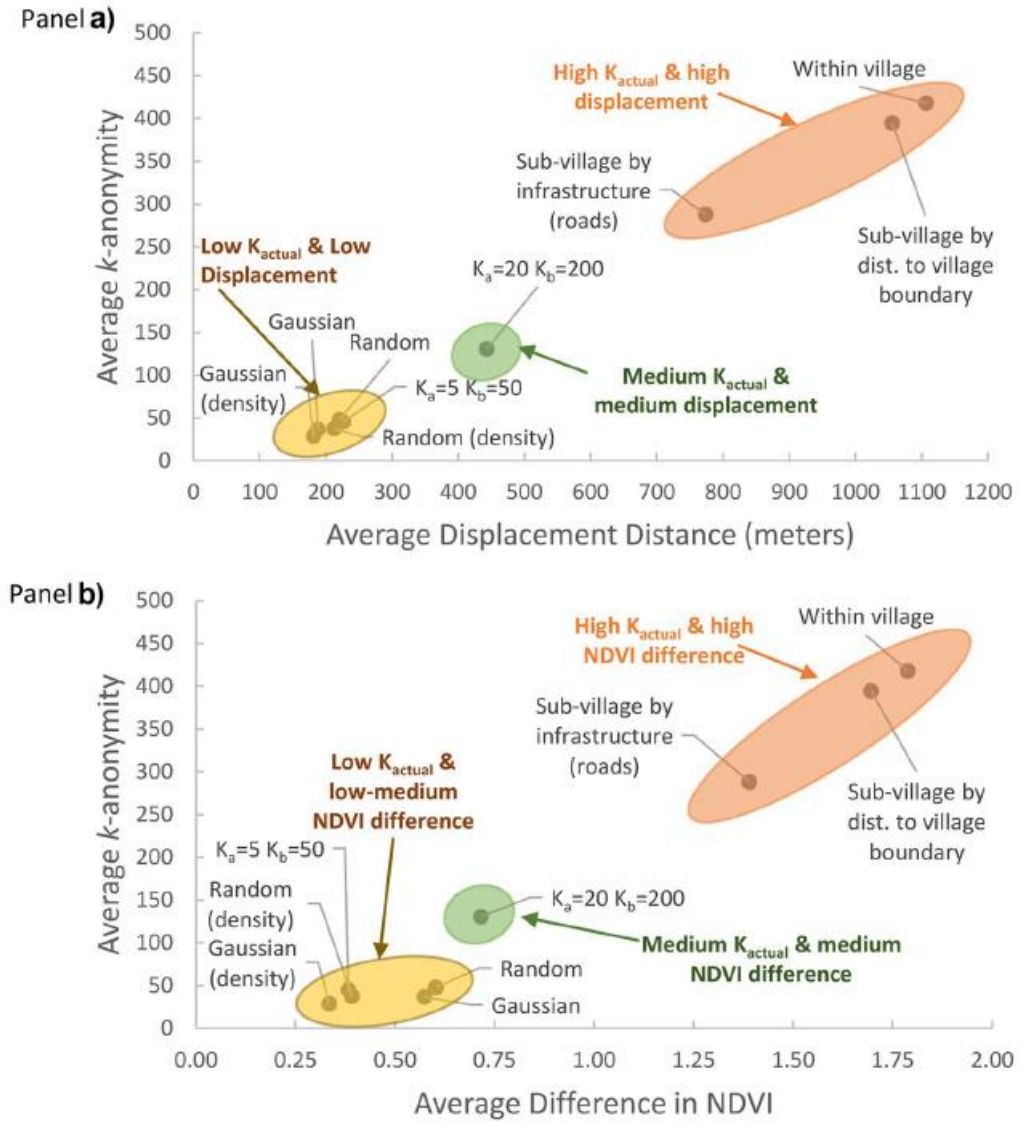

Fig. 6 Evaluation of accuracy and privacy for true vs. geomasked locations, MRC/Wits-Agincourt Unit. **a** Displacement distance versus *k*-anonymity. **b** *k*-Anonymity versus absolute difference in NDVI

- Second plot shows average k-anonymity versus the average difference in the sum of NDVI between true and masked household locations.
- Methods with lower k-anonymity have smaller differences. These methods are random and Gaussian displacement.
- Methods that randomize locations within village or sub-village boundary indicate higher levels of k-anonymity and large differences in NDVI.
- Method incorporating both household density and k-anonymity have high levels of k-anonymity and small differences in NDVI.

Further extension of analysis

- Setting a threshold of k-anonymity.
- It was seen at most 1.2 % of households do not meet privacy standard of 5 households.
- For threshold 10, only $K_a = 20$ $K_b = 200$ provides high anonymization with 99.8 % meeting this privacy standard, while each method has at least one household exposed.
- It can dealt by removing households that fail to meet desired levels of anonymity, swap few households prior to displacement or values within households.
- Such approach should be in consideration of data, research situation.
- More detailed data requires larger k-anonymity.

Summary

- Importance of population-health-environment research.
- Need for geomasking approaches to protect privacy.
- The influence of geomasking on NDVI estimates in Agincourt Health and Socio-Demographic Surveillance System in rural South Africa.
- Trade off between accuracy and privacy.
- It was seen that approaches that use buffers and account for population density produce more accurate results.
- The choice of method is context-specific.
- There is a vast potential for impact in understanding connections between health and environmental change (HDSS).
- Possible further investigations are examination of error introduced within analyses when using NDVI estimates from anonymized locations.

Thank you