UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert Meeting on Statistical Data Confidentiality
26-28 September 2023, Wiesbaden

Dissemination of agricultural geo-referenced data within the context of the 50x30 initiative: an overview of the tradeoff between disclosure risk and data utility

Amsata Niang (United Nations Food and Agriculture Organization, FAO)
amsata.niang@fao.org

Abstract

The dissemination of geo-referenced or geospatial data from agricultural surveys is increasingly requested by users given the range of potential applications, including spatial statistical analysis, geographic cluster modeling, integrating farm data and remote sensed information, etc. However, location data are very disclosing, and the risk of attribute disclosure increases when geo-referenced data are disseminated. The Demographic Health Survey (DHS) has pioneered a geo-masking method to anonymize household locations. The method involves random displacement of the surveyed Enumeration Area (EA) centroid by a chosen angle and distance within a fixed offset. The displacement distance for urban clusters is up to 2km, while for rural clusters, it is up to 5km, with a randomly selected 1% of rural clusters displaced up to 10km. The method has been adopted by the World Bank's Living Standards Measurement Study - Integrated Surveys on Agriculture (LSMS-ISA) team, with a randomly selected 10% of rural clusters displaced up to 10km. The Data production component of the 50x30 Initiative, led by the United Nations Food and Agriculture Organization (FAO), supports countries to produce and disseminate agricultural data including anonymized survey microdata files. The 50x2030 Initiative aims to increase the capacity of 50 low and lower-middle-income countries by 2030 to produce, analyze, interpret, and apply data to decisions in the agricultural sector that support rural development and food security. Within this context, FAO is exploring the possibility to support 50x2030 countries to disseminate geospatial data in compliance with national legislation related to data privacy. The paper discusses the results of the FAO test of the applicability of DHS geo-masking methods to agricultural holdings. It notes an underestimation of the real disclosure risk due to the nature of measures and the type of geospatial data used. The paper highlights the difficulties in designing a risk metric that accounts for both the spatial characteristics of agricultural holdings location and non-geospatial quasi-identifiers from the survey microdata. In the context of the agricultural survey, it proposes an alternative approach of disseminating spatial variables instead of anonymized coordinates, with a framework for Grid cell re-identification (spatial signature) disclosure risk assessment.

Keywords: geo-referenced data, agricultural survey, disclosure risk, spatial variables, spatial signature disclosure risk.

1 Introduction

The 50x2030 initiative aims to improve the capacity of 50 low- and lower-middle-income countries to produce, analyze, interpret, and apply data to decisions in the agricultural sector by 2030.

Countries that partake in the 50x2030 initiative collect geo-referenced data at different levels including households/commercial farm level, plot level, and parcel level. The dissemination of geo-referenced or geospatial data from agricultural surveys is increasingly requested by users given the range of potential applications, including, but not limited to, spatial statistical analysis, geographic cluster modelling, integrating farm data and remotely sensed information, development of small area estimation data products, etc. However, location data are very disclosing, and the risk of identity and attribute disclosure increases when geo-referenced data are disseminated.

A range of methods has been developed to assess location disclosure risk and to anonymize geo-referenced locations. The Demographic Health Survey (DHS) has pioneered a geomasking method that involves random displacement of the surveyed Enumeration Area (EA) centroid by a random angle and distance within a fixed offset. The DHS displacement method has been adopted widely since its implementation and extensively studied and accounted for by analysts using DHS GPS data (Inter-Secretariat Working Group on Household Surveys, 2021)

With the increasing demand for the dissemination of geo-referenced data from the 50x230, FAO is exploring the possibility to support 50x2030 countries to disseminate geospatial data in compliance with national legislation related to data privacy. The test of the applicability of DHS geo-masking methods to agricultural households in Senegal has revealed an underestimation of the real disclosure risk due to the nature of measures and the type of geospatial data used in the risk assessment. The test shows the difficulties in designing an individual risk metric that accounts for both the spatial characteristics of agricultural holdings location and non-geospatial quasi-identifiers from the survey microdata.

In the context of the agricultural survey, an alternative data product is being explored. It consists of disseminating spatial variables instead of anonymized coordinates. This alternative does not have a zero-disclosure risk. It can be possible to disclose geographic attributes based on the spatial signature of spatial variable records.

2 Geo-referenced data collection in the 50x2030 Initiative

The 50x2030 initiative to close the agricultural data gap is a multi-partner program that seeks to bridge the global agricultural data gap by transforming country data systems in 50 countries in Africa, Asia, the Middle East and Latin America by 2030. The Data production component of the 50x30 Initiative, led by the United Nations Food and Agriculture Organization (FAO), supports countries to produce and disseminate agricultural data including anonymized survey microdata.

The 50x2030 Initiative uses integrated survey models that are tailored to each country's existing agricultural survey program. Partner countries select a survey program based on their needs, capacity, and potential for technical and financial takeover. The surveys contribute to the data needs of Sustainable Development Goal 2 (SDG2) and the Comprehensive Africa Agriculture Development Programme (CAADP) by addressing 8 SDG indicators and 9 CAADP indicators. The Initiative aims to promote sustainable agriculture, end hunger, achieve food security, and improve nutrition.

The 50x2030 survey program is composed of a Core module that implements annual survey program and selected rotation farm modules. The rotative modules are:

- ILP: farm Income, Labour, and Productivity
- PME: Production Methods and the Environment
- MEA: Machinery, Equipment, and Assets
- ILS-HH: Income and Living Standards-Households.

For each survey round and depending on the type of agricultural sector, geo-referenced data are collected at different levels such as interview location, cultivated plots, and agricultural parcels.

Table 1: Recommendation for collecting georeferenced and GPS-based data in the 50x2030 Initiative.

Type of GPS measurement	Household Sector			Non-HH Sector
	2 visit structure: Post Planting Visit	2 visit structure: Post Harvest Visit	1 visit structure	1 visit structure
GPS-based area measurement with saved outlines	Cultivated plots; Agricultural parcels*	N/A	Agricultural parcels*[1]	N/A
Coordinate collection (directly in tablet)	Cultivated plots (center point); Interview location	Interview location	Interview location	Interview location

**For the CORE, ILP, and MEA this includes fully or partially cultivated parcels. For PME this includes all parcels owned or operated (including pasture, etc.).*

3 DHS geomasking methods for household survey

DHS geo-referenced data collection and dissemination

The DHS project started georeferencing coordinate data of cluster locations in the late 1980s and began making georeferenced GPS datasets available to the public in 2003. The georeferenced datasets can be linked to individual records in DHS household surveys through unique identifiers; however, the georeferenced datasets are kept separate from the main household data files and are available only by special permission. (Burgert, et al., 2013)

DHS displacement method

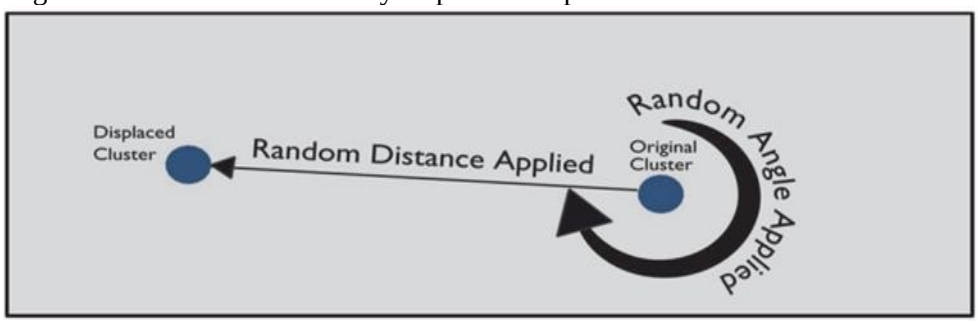
In DHS household surveys, a GPS coordinate displacement process is carried out as follows:

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

According to (VanWey, et al., 2005), urban and rural clusters are treated differently due to the varying population densities. The clusters in rural areas, which have lower population densities, require larger displacement distances to achieve the same level of reduced disclosure risk as clusters in densely populated urban areas.

GPS coordinates are displaced according to the "random direction, random distance" method with the constraint that (i) the displaced points should not be outside national geographic boundaries. and (ii) the displaced points should not be outside the second administrative boundaries.

Figure 1: DHS household survey displacement process



Source: (VanWey, et al., 2005)

[1] Depending on the timing of the visit, for single visit surveys, it is likely not be possible to conduct plot level GPS area measurements (if crops are out of the field and the boundaries are no longer apparent)

4 Application of the DHS geomasking method in agricultural integrated survey: challenges and limitations

Agricultural integrated surveys (AGRIS) generally cover agricultural households/family holding and non-households holding like commercial farms. Unlike DHS surveys, AGRIS Surveys collect GPS-based areas for agricultural parcels and cultivated plots for the households sector and GPS location for plot centre point and interview location for both households sector and non-household sector.

- **Limitation in anonymizing parcel and plot location or area**

Anonymizing plot location using the DHS geomasking methods can lead to significant information loss and make the anonymized coordinate less useful. In fact, one of the main uses of plot location data is its integration to remotely sensed data related to agriculture and the displacement can have a significant impact in the remotely sensed data that will be extracted in the anonymized location of plots, especially if the remote sensed data resolution is high. Furthermore, the DHS geomasking may not be appropriate to anonymize GPS-base area of cultivated plots and agricultural parcels and any displacement of this area lead to significant information loss when combined with remotely sensed data.

When considering the location of agricultural households/family holdings and non-household sectors, DHS masking emerges as a technically conceivable solution. It is worth noting that some concerns have arisen regarding the utilization of DHS geomasking in AGRIS survey location data for the agricultural household sector or non-household sector. This prompts a closer examination of several key aspects.
Firstly, it is essential to evaluate whether the DHS displacement parameters are truly optimal in terms of the classical risk-utility tradeoff, in other words, whether these parameters strike an appropriate balance, effectively managing the associated risks while maximizing utility.

Secondly, the suitability of standard disclosure risks for agricultural households and the non-agricultural sector should be thoroughly assessed. This evaluation seeks to determine whether the existing disclosure risk measures adequately address the unique characteristics and sensitivities of these sectors, ensuring the protection of their data while facilitating necessary information dissemination.

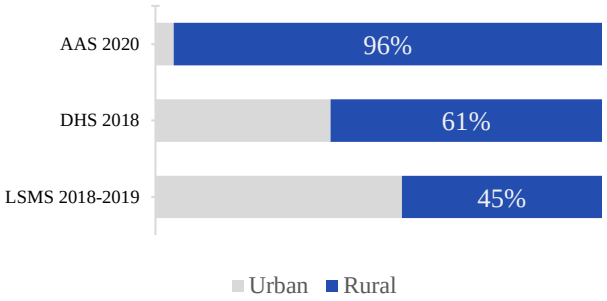
Moreover, it is important to investigate whether the microdata dissemination policies of 50x2030 countries provide a designated space that guarantees a legal framework for disseminating anonymized location information. The presence of such a space is crucial to ensure compliance with relevant laws and regulations, safeguarding privacy and confidentiality while allowing for the responsible sharing of anonymized location data.

Optimality of DHS geomasking in AGRIS survey location: utility

The survey objectives, sampling design and target population can have an impact in the magnitude of displacement based on DHS methods, through the share of rural households in the survey sample.

Around 96% of agricultural households of the 2020 Annual Agricultural Survey of Senegal (AAS) are located in rural Areas. This is not surprising as most of the target population (agricultural households) are located in rural areas. In contrast to AAS, the sample of the 2018 DHS of Senegal and the sample of the 2018-2019 LSMS survey have respectively 61% and 45% of rural households.

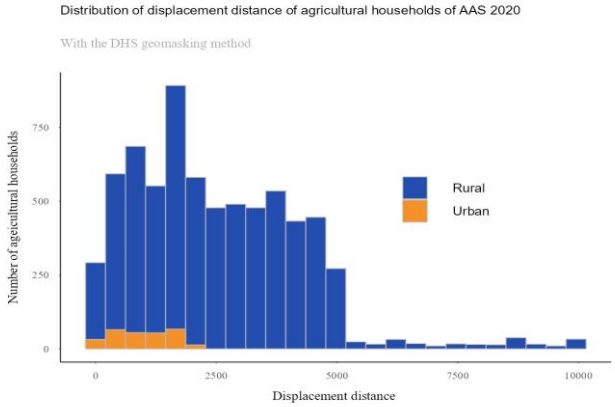
Figure 2: Distribution of the sampled households by urban-rural



Based on the DHS geomasking methods applied to AAS, 286 of the urban households will be displaced to a distance between 0 and 2 kilometers away from the original location, 6618 households will be relocated between 0 and 5 km away and 66 households will be displaced to between 0 and 10 km.

Table 2: DHS geomaking parameters applied to the AAS.

Agricultural households	Number	Displacement buffer radius
Urban households	286	2 km
99% of rural households	6618	5 km
1% of rural households	66	10 km



After applying the DHS masking methods, half of the agricultural households are displaced to a distance higher than 2.1 km, and out of 4 agricultural households, one is displaced by more than 3.6 km. Furthermore, 5% of the agricultural households are displaced to a distance higher than 4.9 km.

Based on the urban-rural distribution of the Sample, the DHS and LSMS can have lower values for the bellow-mention indicators (share of households below the median, above the 3rd quantile, and above the 95th percentile of displacement distance). Hence, the impact of the urban-rural distribution, which highly depends on the survey characteristics, has a higher impact on the displacement distance in the AAS than in the DHS or LSMS of Senegal.

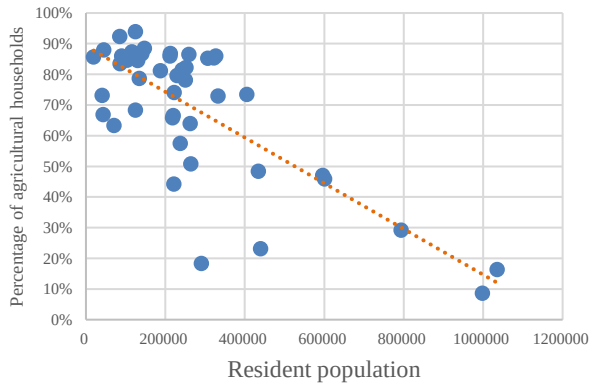
Suitability of standard disclosure risk for geo-anonymization

- **Individual risk**

After having applied masking in household locations, the intruder has some information they can explore in the process of re-identification: The original household is located in the buffer around the anonymized location. In the case of the DHS methods, the radius of the buffer depends on the rural/urban location of the households. One possibility to assess the disclosure is to count how many households are present in the buffer. The higher the number of households, the more difficult it is to re-identify the original household of the anonymized location. This is formalized as the notion of Spatial k-anonymity. The metric $\frac{1}{k}$ is considered as an estimation of the

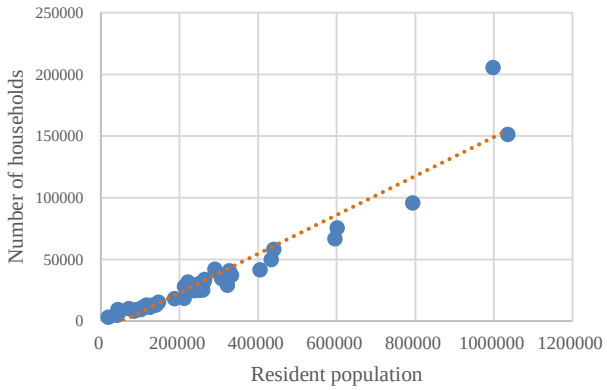
probability of re-identification of the original household. To compute this quantity, the number of households within the buffer should be known, which is not always the case. Proxy datasets are used to assess the number of households. It includes, but is not limited to, Population raster, Building footprint raster, etc.
According to (VanWey, et al., 2005) threats to data security also depend on heterogeneity in the spatial distribution of the sample clusters. If there are regional or rural–urban patterns or other systematic patterns in the spatial distribution of the clusters, however, additional care must be taken. Like the DHS cluster, the AAS cluster can exhibit urban-rural pattern but not in the same direction than in another household survey. In fact, the density of the target population of DHS (all household) are higher in urban area than in rural area, the reverse situation is observed for the target population of the agricultural households. The figure 3 shows that Senegalese department (second level administrative subdivision) with high population are characterized by lower proportion of agricultural households.

Figure 3: relation between resident population and share of agricultural households.



Data: ANSD, population and housing census 2013

Figure 4: Relationship between resident population and number of households.

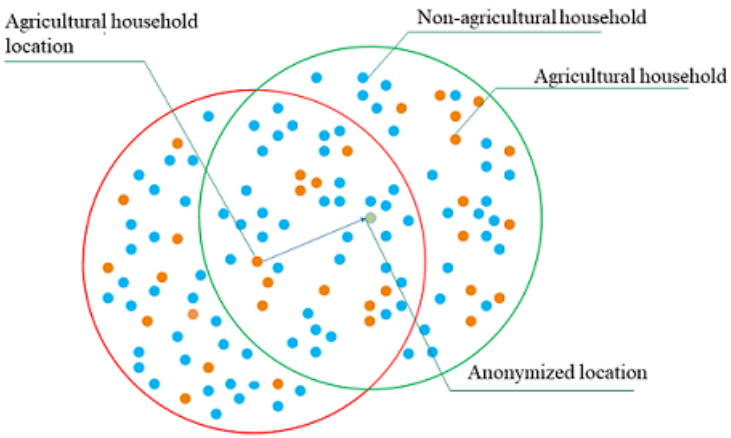


Data: ANSD, population and housing census 2013

This systematic pattern of the distribution of agricultural households makes it difficult to use population density or building footprint to assess the spatial disclosure risk as their high density does not necessarily mean high density of agricultural households, holding or commercial farms, even if they may mean high density of households, regardless their attributes. If we let K be the total number of households in the buffer and K' the total number of agricultural households in the buffer. The standard spatial k-anonymity (using population dataset, building footprints, etc.) aims to approximate the quantity $\frac{1}{K}$. When applying to agricultural location, the real risk seems to be underestimated because the real risk in the context of agricultural households' location should consider K' instead K.

$$\frac{1}{K} \leq \frac{1}{K'}$$

.



Spatial k-anonymity based on spatial feature correlated with the number of the household such as population raster, and building footprint may not be appropriate to estimate spatial k-anonymity for agricultural households.

This approach under-estimate the real disclosure risk associated with the dissemination of anonymized locations for agricultural households. Furthermore, these proxy data are not suitable for non-household farms.

Table 3: Spatial k-anonymity assessment using population dataset.

spatial k-anonymity thresholds	Rural	Urban
k<=1000	652 (9%)	0 (0%)
k<=2000	1109 (17%)	0 (0%)
k<=5000	2199(33%)	0 (0%)

Population data: World pop

The spatial dataset to be used, in the context of agricultural surveys, should be correlated with the number of agricultural households in the population.

- **Attribute disclosure**

Attribute disclosure consists of discovering some characteristics of an individual without identifying the associated data record (Thijs & Matthew, 2019). Beyond household disclosure, there is other re-identification that can occur via the dissemination of geo-referenced information such as re-identification of community, town, or other geographic attributes. The Inter-Secretariat Working Group on Household Surveys (Inter-Secretariat Working Group on Household Surveys, 2021) has tested the use of some proxy datasets that represent community features to assess the risk of community disclosure. Those datasets include villages' locations, the location of populated public places, and small settlement areas. Applying the notion of spatial k-anonymity with a threshold of 5 counts, it has come out that this proxy dataset would not support an accurate estimation of community-level reidentification risk, particularly in the urban context, and there is a need for additional data exploration.

In this exercise, the village location of Senegal from the National Statistical Office has been used. The approach used to assess the disclosure of this geographical feature consists of performing a spatial joint between the anonymized location and village location using the nearest village criteria. Afterward, the metric used is the percentage of households which village is re-identified through spatial joint. It comes out that 40% of agricultural households from AAS can be still linked to their original village after anonymization through a spatial joint. This indicator is 87% for urban households and 36% for rural agricultural households.

Table 4: Number and percentage of households with re-identified village by urban/rural

	Number	Percentage
Urban	249	87%
Rural	2401	36%
All agricultural household	2650	40%

Spatial disclosure risk combining location and non-georeferenced quasi-identifier.

In addition to the aforementioned limitation regarding the use of spatial k-anonymity in the agricultural survey context, there are other intrinsic limitations that should be taken into consideration. One such limitation is that spatial k-anonymity solely focuses on spatial characteristics and does not account for non-spatial quasi-identifiers present in the microdata. Consequently, spatial k-anonymity inherently underestimates the true disclosure risk associated with the dissemination of anonymized location data, regardless of the suitability of the spatial dataset used in the computation.

This emphasizes the need to explore additional techniques or frameworks that address the challenges posed by non-spatial quasi-identifiers, thereby ensuring a more accurate and comprehensive assessment of disclosure risk in spatial anonymization processes.

Table 5: Comparative difference in spatial anonymization between DHS and AGRIS survey

Characteristics	Survey		The implication in DHS geomasking with AGRIS data
	DHS	AAS	
Share of rural households in the sample	60%	95%	• The majority of households will be highly displaced according to the DHS displacement methods.
Target population	All households	Agricultural household	• Population density cannot be used to assess disclosure risk. Need to find the best geographic feature to assess disclosure risk
GPS data collected	Households	Households, cultivated plots, and parcels	• Displacement of plots and parcels leads to higher information loss, especially when combined with remotely sensed data. • DHS displacement methods may not be appropriate to anonymize parcels and plots. • DHS displacement is not suitable to anonymize GPS-base area (plot or parcel boundaries)
Release type of anonymized location	Special permission	Not feasible for PUF or SUF, the usual release type of 50x2030 microdata	• Special release type is needed. • Some countries may need to update their microdata dissemination policy

5 Spatial covariates as an alternative to dissemination anonymized location in the agricultural integrated survey

The spatial covariates dataset or spatial variables refers to a set of spatial variables like temperature, population, and precipitation, etc., extracted at the location/buffer of the survey unit. For statistical disclosure consideration, this information is often extracted from the anonymized location/buffer of the survey unit.

The dissemination of Spatial covariates, even if it is done with the anonymized location, has a disclosure risk associated with it. Remote sensed data and other spatial information extracted as spatial covariates can exhibit spatial patterns. Considered alone, a spatial covariate is not likely to identify a location, but when combined together, it can be possible to identify a location through its spatial signature. The notion of spatial signature, in this case, is close to the definition given by (Fuentes, et al., 2022). Within the framework of the construction of a raster-based classification algorithm, (Fuentes, et al., 2022) define spatial signature as a relative measurement of the correspondence between any XY location in geographic space and the landscape configuration represented by a classification unit. In the context of spatial covariates, the landscape configuration can be represented by any spatial feature that derives from the used spatial covariates.

Disclosure scenarios

In order for the re-identification of households from spatial covariates to be possible, the intruder should have at its disposal some information and proceed to attack using adequate methods. Therefore, it is important to consider some disclosure scenarios before proposing any risk assessment framework from spatial variables. In the exercise of developing a spatial signature risk assessment framework, the two below disclosure scenarios are proposed. From these scenarios, risk assessment methods and metrics could be later suggested.

Disclosure scenario 1: Record-level re-identification (identity disclosure)

The intruder has XY location of the statistical unit from the survey sample and tries to link this unit with a record in the spatial covariate datasets.

- **Hypothesis 1:** The intruder has access to all the raw data of spatial variables (variables like temperature, population, and precipitation) used to extract the covariate at the survey location.

- **Hypothesis 2:** The intruder can extract the exact spatial covariate information at the locations he/she has.
- **Hypothesis 3:** The intruder proceeds to the re-identification by taking the spatial covariate record which is more similar in terms of spatial signature than the XY location he/she has.
-

Disclosure scenario 2: Geographic entity disclosure (attribute disclosure)

The intruder wants to disclose geographic entity information which has not been disseminated. Those entities can be a lower level of administrative boundaries, villages, etc.

- **Hypothesis 1:** The intruder has access to all the raw data of spatial variables (variables like temperature, population, and precipitation) used to extract the covariate at the survey location.
- **Hypothesis 2:** The intruder can extract the exact aggregate of spatial covariate information at the geographic entities' unit he/she has.
- **Hypothesis 3:** The intruder proceeds to the re-identification by taking the spatial covariate record which is more similar in terms of spatial signature than the geographic entity one.

The upcoming phase of the exercise involves a comprehensive exploration of appropriate disclosure risk assessment methods for Spatial covariates through spatial signature. This will take into account the above-mentioned disclosure scenario. This approach promises to enhance the effectiveness and reliability of safeguarding location information and other geographic information during spatial covariates release.

6 Conclusion

Since its inception in the early 2010s, the DHS geomasking method has gained prominence as the prevailing standard for geomasking household survey data. Its applicability in the context of agricultural surveys has been tested, providing the FAO AGRIS survey team with valuable insights into the advantages and disadvantages of using this displacement method specifically for geo-referenced agricultural data dissemination. However, further research and testing are required to ensure the safe adoption of the DHS displacement method in AGRIS surveys.

As an alternative to the dissemination of anonymized households' location, the dissemination of spatial covariates has emerged as a conceivable alternative. Nevertheless, it is crucial to correctly evaluate the disclosure risk associated with the dissemination of spatial covariates through spatial signature. Adequate methods should be implemented to effectively mitigate the associated risk. This comprehensive assessment and risk reduction approach are necessary to safeguard the confidentiality and privacy of the data collected in agricultural surveys.

7 References

ANSD, 2020. Sénégal : Enquête Démographiqueet de Santé Continue (EDS-Continue) 2019, s.l.: s.n.

ANSD, 2021. Enquête harmonisée sur les Conditions de Vie des Ménages (EHCVM) au Sénégal, s.l.: s.n.

Burgert, C. R., Josh, C., Thea, R. & Blake, Z., 2013. Geographic displacement procedure and georeferenced data release policy for the Demographic and Health Surveys.DHS Spatial, Calverton, Maryland, USA: ICF International.

Fuentes, B. A., Dorantes, M. J. & Tipton, d. J. R., 2022. rassta: Raster-Based Spatial Stratification. The R Journal, 2 June.Volume 14.

Insee - Eurostat, 21018. Handbook of Spatial Analysis,theory and Application with R. INSEE Méthodes, Octobre, pp. 349-367.

Inter-Secretariat Working Group on Household Surveys, 2021. Spatial Anonymization. s.l.:UNSC.

THE 50x2030 INITIATIVE, 2022. A guide to the 50x2030 data collection approach: Questionnaire design, s.l.: 50x2030 technical notes series.

Thijs, B. & Matthew, W., 2019. Statistical Disclosure Control for Microdata: A Practice Guide for sdcMicro. s.l.:International Household Survey Network Revision.

VanWey, L. K. et al., 2005. Confidentiality and spatially explicit data: Concerns and challenges. Proceedings of the National Academy of Sciences (PNAS), 102(43), p. 4.