50x2030
DATA-SMART AGRICULTURE

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, Statistician, FAO

September , 2023

Outline

1. Presentation of the 50x2030 initiative
2. Collection of geo-referenced data in the 50x2030 Initiative
3. DHS masking methods applied in agricultural households ‘survey (test with Senegal data)
4. Ongoing test of SDC risk assessment on spatial variables dissemination

The 50x2030 Initiative

» Promotes data-smart agriculture to address food crises, climate vulnerabilities, improve rural livelihoods, create jobs & build resilience

The Challenge

The scarcity of high-quality, regular, and relevant agricultural data makes it extremely difficult for policymakers to make sound decisions to drive their country's economic growth and reduce poverty.

The Opportunity

Produce effective data and make it accessible and available to all stakeholders so they can build capacity to enable data-driven policies and decision-making. Support countries in addressing food security, sustainability, and climate change.

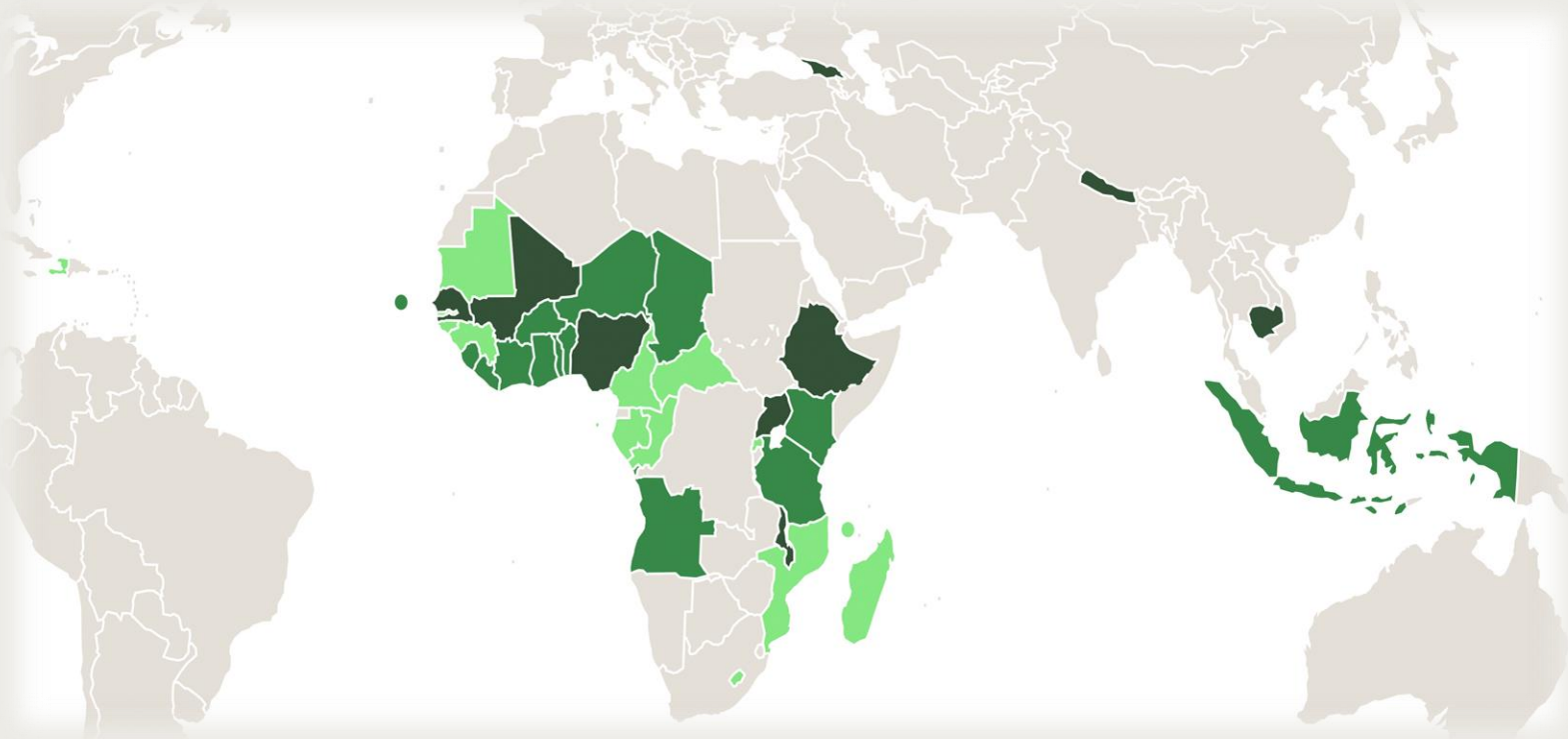
50x2030 Countries

Program Countries in FY23: 14

Angola
Benin
Burkina Faso
Cabo Verde
Chad
Côte d'Ivoire
Ghana
Indonesia
Kenya
Liberia
Niger
Sierra Leone
Tanzania
Togo

Already engaged: 9

Cambodia Ethiopia Georgia Malawi Mali
Nepal Nigeria Senegal Uganda



Program countries in FY24: 15

Cameroon
Central Africa
Comoros
Congo
Gambia
Gabon
Guinea
Guinea bissau
Haïti
Lesotho
Madagascar
Mauritania
Mozambique
Rwanda
São Tomé e Principe

Financing secured for data collection in 24 countries for 3 to 5 years through 4 IDA financed statistics & 2 ag./environment projects in Africa and 1 in Nepal

Collection of geo-referenced data in the 50x2030 Initiative

Data Covered by the 50x2030 Survey Programme

CORE: Crops, Livestock, aquaculture, fisheries, forestry production

ILP: Agricultural income, agricultural labor and productivity, land tenure, gender decision-making

PME: Production, Methods and environment, Agricultural sustainability

MEA: Assets, Machinery, Equipment

Recommendations on Collection of geo-referenced data

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

What are the tradeoff risk/utility associated with the anonymization and dissemination of these geo-referenced data?

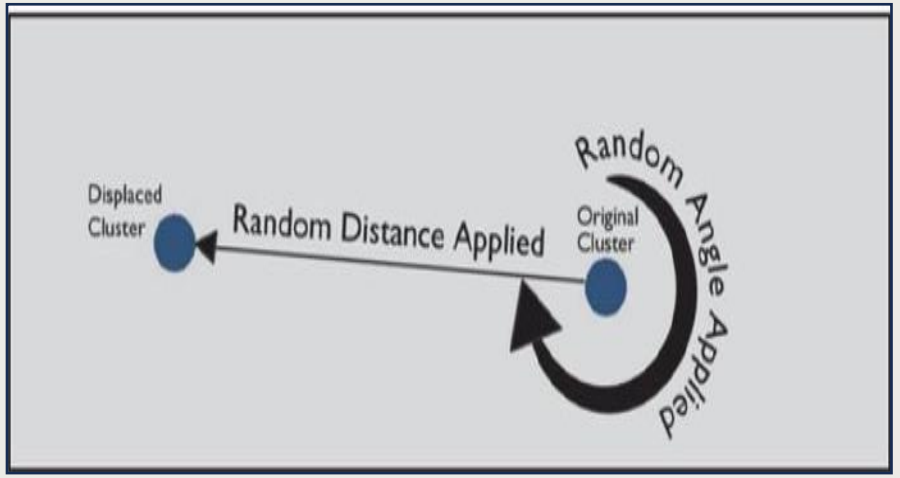
Anonymization of geo-referenced data from agricultural integrated survey: test of applicability

- The Dissemination unit of the FAO AgriSurvey Team has started a series of test on the applicability of spatial anonymization on agricultural survey data with Senegal as starting country

Anonymization method

The Geomasking Method of the Demographic and Health Survey has been considered for this test

- Urban clusters are displaced at a distance of 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.



✓ Very well-documented methods and have been adopted by the LSMS-ISA Team of the World Bank

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

- **Area-based measures** ⇒ Anonymizing plot location using the DHS masking methods can lead to significant utility loss and make the anonymized coordinate less useful

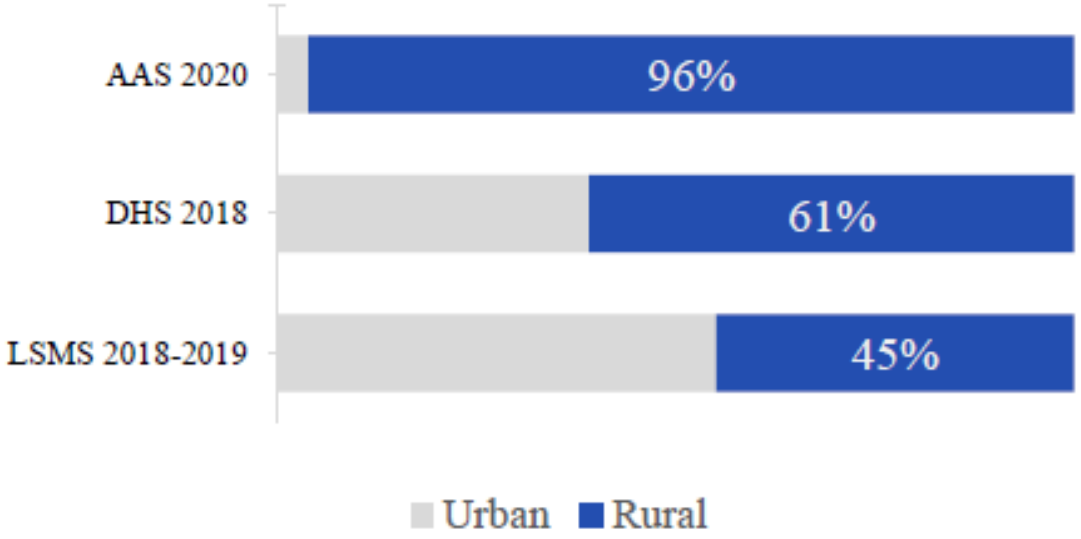
- **Commercial farm sector**⇒ Non-household sector is not covered in this test. Additional consideration may be needed.

- **Dissemination policy** ⇒ Need to enhance countries' dissemination policy/protocol

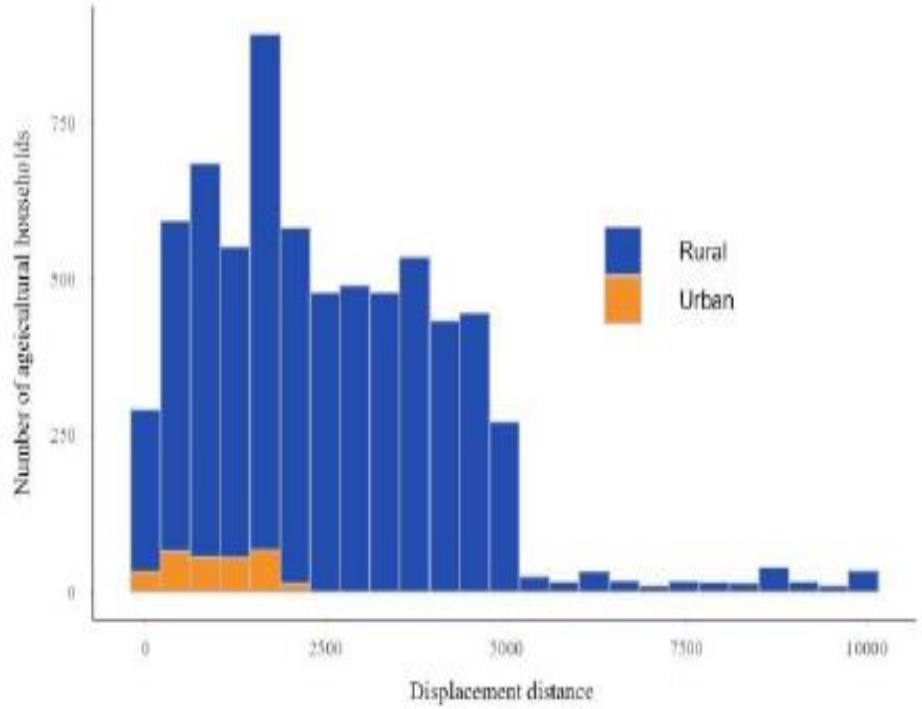
Comparative analysis of displacement distance between the AAS, DHS, and LSMS of Senegal

- The rural/urban distribution of the sample, which highly depends on the survey characteristics, may lead to a higher impact on the displacement distance in the AAS than in the DHS or LSMS of Senegal

Distribution of survey samples



Distribution of the displacement distance



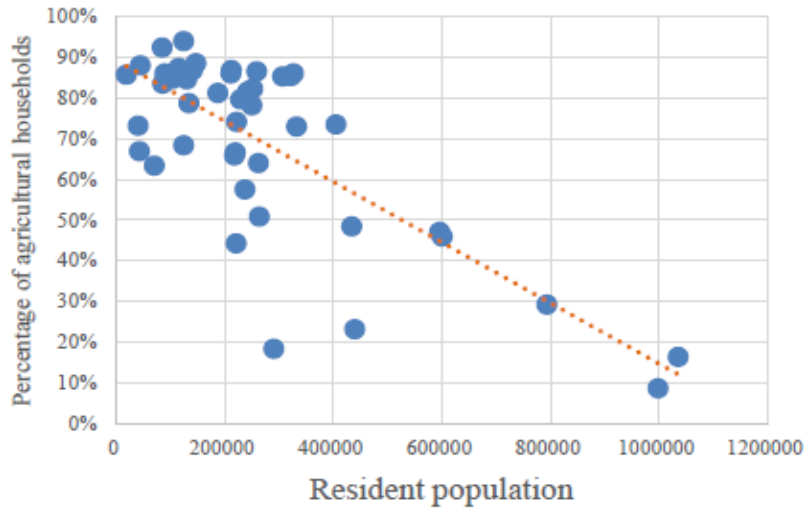
Suitability of standard disclosure risk for geo-anonymization: Spatial k-anonymity

» 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.

- Population raster is suitable to be used in spatial k-anonymity for households surveys where the target population in all households

- For the agricultural survey of Senegal, the target population is not all households but agricultural households only

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



Data: ANSD, population and housing census 2013

Suitability of standard disclosure risk for geo-anonymization: Spatial k-anonymity

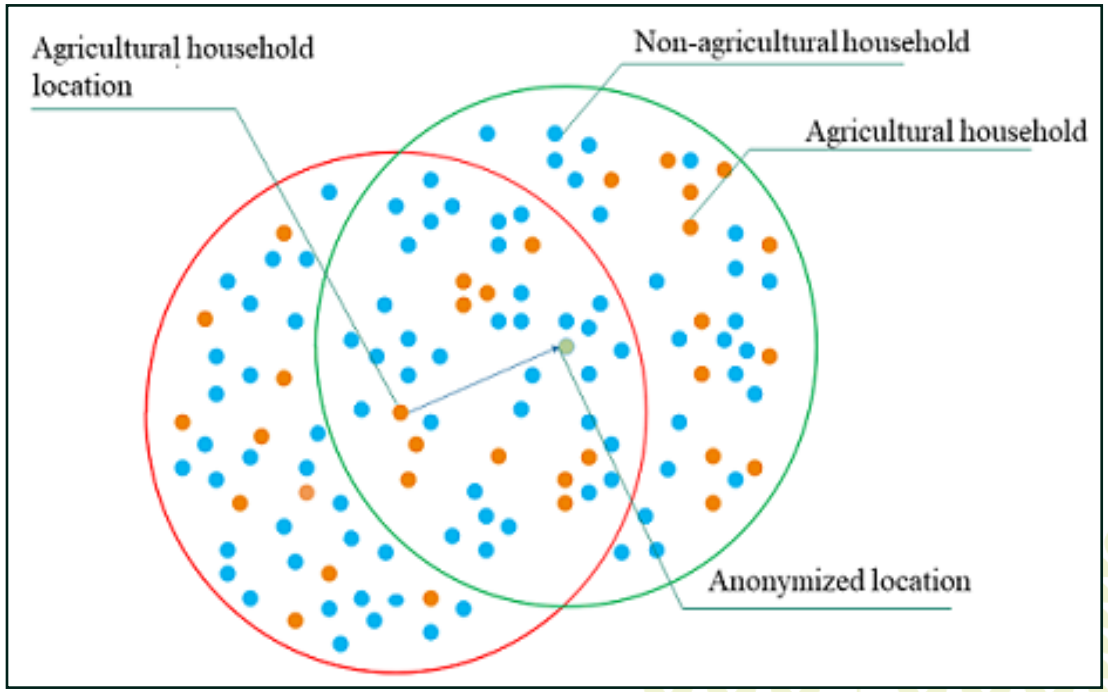
- Spatial k-anonymity based on spatial feature correlated with the number of the household such as population raster, and building footprint may not be appropriate to estimate spatial k-anonymity for agricultural households.

K= Total number of households

K'= Number of agricultural households

$\frac{1}{K} \leq \frac{1}{K'}$: using spatial k-anonymity based on population raster may under-estimate

Need to find an appropriate spatial dataset/methods to compute spatial k-anonymity in the context of agricultural survey



Sufficiency of the displacement: Community disclosure

» 40% of agricultural households from AAS can be still linked to their original village after anonymization through a spatial joint.

Attribute disclosure consists of discovering some characteristics of an individual without identifying the associated data record (Thijs & Matthew, 2019).

- Village locations of Senegal from the National Statistical Office have been used.
- Spatial joint between the anonymized location and village location using the nearest village criteria

	Number of ag, hh	Percentage of ag, hh
Urban	249	87%
Rural	2401	36%
All	2650	40%

Summary of the test of applicability

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

Alternative data product

Spatial covariates/variables

- 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.

Statistical disclosure ⇒ Statistical disclosure can occur after the dissemination of spatial variables, through the exploration of their spatial pattern/signature,

Spatial signature ⇒ the correspondence between any XY location in geographic space and the landscape configuration represented by any spatial feature that derives from the used spatial covariates

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 1:

Geographic entity disclosure (attribute/community 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.

On-going test activities: Spatial disclosure risk metrics

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

Conclusion

- DHS geomasking method has gained prominence as the prevailing standard for geomasking household survey data.
- Its applicability in the context of agricultural surveys, tested with Senegal data, showed the need to inquire for suitable spatial datasets/measures to carefully assess the associated disclosure risk.
- 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

THANKS

50x2030
DATA-SMART AGRICULTURE

Food and Agriculture Organization of the United Nations

SUPPORTED BY
THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

IFAD
Investing in rural people

USAID
FROM THE AMERICAN PEOPLE

BILL & MELINDA GATES foundation

Australian Government

Cooperazione Italiana allo Sviluppo
Ministero degli Affari Esteri e della Cooperazione Internazionale

Federal Ministry for Economic Cooperation and Development