A GRID-BASED SAMPLING DESIGN FOR HOUSEHOLD SURVEYS

IN THE ABSENCE OF ACTIONABLE SAMPLING FRAMES

Édith Darin[1], **Gianluca Boo**[1], **Dana R Thomson**[2], **Andrew J Tatem**[1,2]

[1] WorldPop Research Group, Department of Geography and Environmental Science, University of Southampton, UK

[2] Flowminder Foundation, Stockholm, Sweden

INTRODUCTION

Household surveys are a cost-effective data source for **estimating health and demographic characteristics** in low- and middle-income countries. These surveys involve sampling from a **frame listing all the members of the population of interest**. However, the listing needs to be complete, accurate, and up-to-date to draw samples that can be generalized to the entire population [1].

A sampling frame generally consists of the enumeration areas established during the last national census, together with the associated population counts. However, these counts are often inaccurate because the national census is carried out on a decade basis. **Outdated sampling frames are particularly critical in Africa,** where five countries had their last census more than fifteen years ago [2].

To tackle the absence of actionable sampling frames, we propose an **original grid-based sampling design**, embedding spatial sampling concepts into household surveys. We demonstrate our framework with a case study developed in the **western part of the Democratic Republic of Congo**. This country had its last national census in 1984, and the resulting sampling frame is still currently used in household surveys [3].

CONCEPTUAL FRAMEWORK AND CASE STUDY

The proposed grid-based sampling design draws from the formal description of spatial sampling, where we adopt the **concept of random field to describe the population of interest** [4]. In this context, we regard the geographic distribution of population as characterized by **first- and second-order spatial non-stationarity** because settlements are highly clustered and of heterogeneous sizes and densities. We also consider the **discrete nature of populated places**, by adapting spatial sampling designes concieved for environmental phenomena to household surveys. Our conceptual framework consists of three essential components illustrated in an **EXTRACT FROM THE STUDY AREA**:

1. GRIDDED SAMPLING FRAME enables capturing the geographic distribution of the population of interest through a regular spatial structure. The use of a **gridded settlement layer**, extracted using remote sensing techniques, provides a proxy for the **discrete distribution of populated places**. Only **settled grid cells** are therefore considered as **the sampling units**. In our case study, we could not discard non-residential buildings because the settlement layer did not include this information.

2. CONTEXTUAL STRATIFICATION accounts for first- and second-order spatial non-stationarity. For instance, distance to main roads and conflict areas as well as elevation and precipitation are **contextual factors** driving the **diversity of populated places**. Contextual datasets can be combined through a **principal component analysis** to extract the most relevant linear components, which are then grouped using a **k-means clustering algorithm**. In our case study, the resulting clusters visually corresponded to urban, peri-urban, and rural settings.

2. POPULATION-WEIGHTED SAMPLING also seeks to capture second-order spatial non-stationarity manifested through the presence of **few highly dense populated areas**. Under this sampling design, the **optimal sample size** can be estimated by simulating different sample sizes using GridSample [5] and, based on the **Kolmogorov–Smirnov distance**, assessing how well the sampled population approximates the distribution of a baseline gridded population. The **sampling weights can also be incorporated the estimator**. To account for the spatial non-stationarity, sample size estimation has been carried out independently for each stratum.

DISCUSSION

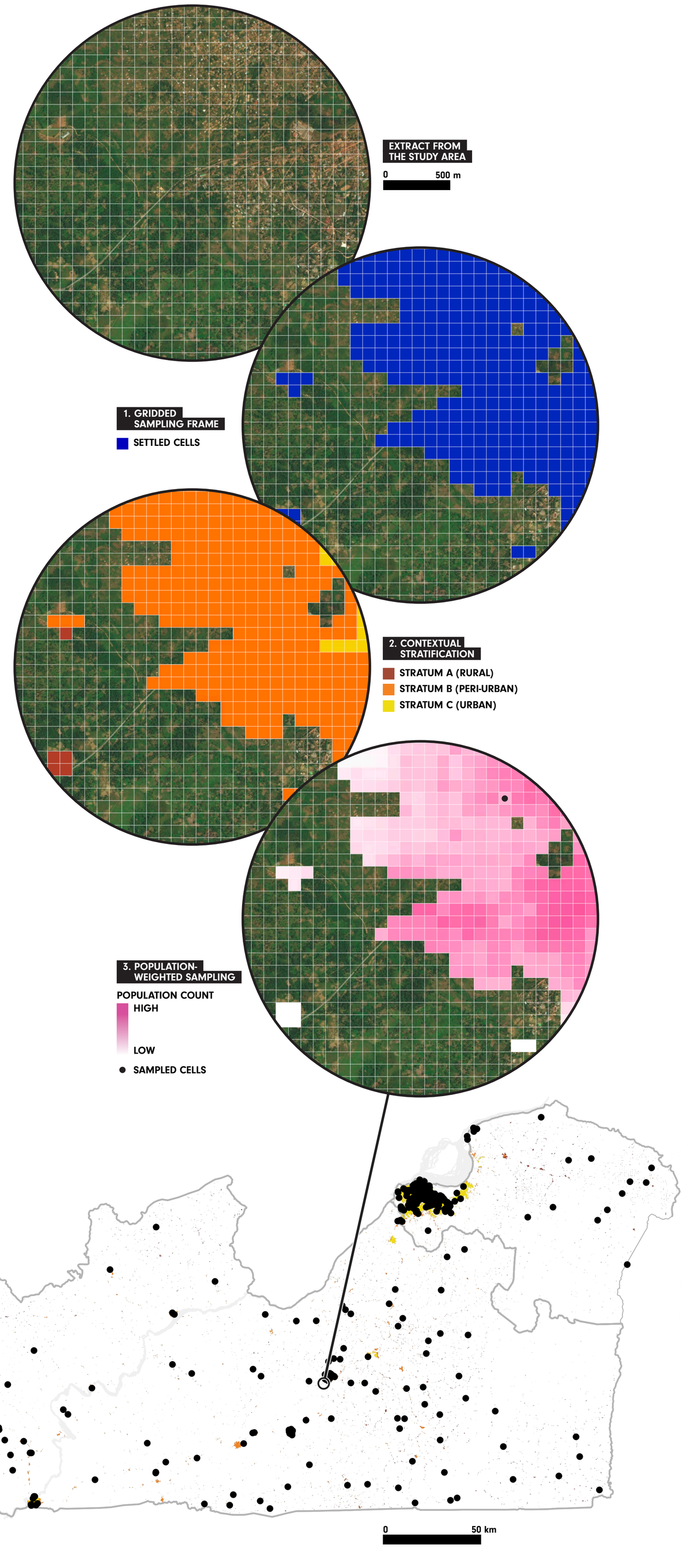
The presented case study enables us to highlight four main considerations related to the implementation of our conceptual framework.

A) **Gridded settlement data** can provide an alternative sampling frame which is complete, accurate, and up-to-date. However, these requirements need to be met by the settlement layer to draw generalizable samples, where the **feature extraction process is critical** to the sampling design.

B) **Contextual data** can enable to refine the traditional urban/rural stratification by better accounting for the diversity of populated places. Dealing with **high dimensional contextual information is a major challenge** to produce meaningful contextual strata.

C) **Gridded population data** can allow to sample densely populated places through population-weighted sampling. Given that this design under-samples low density areas, this design could be **complemented by other strategies** (e.g., spatial oversampling).

D) **Gridded population data** can also be used to estimate the optimal sample size. Given the high spatial resolution of available gridded population data, it is possible to aim at capturing the **distribution of the entire population of interest** and not only its mean.



REFERENCES

[1] Turner AG. 2008. Sampling frames and master samples. In "Designing household survey samples: practical guidelines". New York, USA: United Nations.

[2] United Nations Statistics Division (UNSD). 2019. The 2020 World Population and Housing Census Programme. [Online]. https://unstats.un.org/unsd/demographic-social/census/censusdates.

[3] Ministère du Plan et Suivi de la Mise en œuvre de la Révolution de la Modernité (MPSMRM), Ministère de la Santé Publique (MSP), and ICF International. 2014. Enquête démographique et de santé en République Démocratique du Congo 2013-2014. Rockville (MA), USA: MPSMRM, MSP and ICF International.

[4] Wang J-F., Stein A., Gao B-B., and Ge Y. 2012. A review of spatial sampling. Spatial Statistics 2 (1).

[5] Thomson DR., Stevens FR., Ruktanonchai NW., Tatem AJ., and Castro MC. 2017. GridSample: an R package to generate household survey primary sampling units (PSUs) from gridded population data. International Journal of Health Geographics 16 (1).

AKNOWLEDGEMENTS

This work was funded by the Bill and Melinda Gates Foundation and the United Kingdom Department for International Development (OPP1182408).

E.C.Darin@soton.ac.uk www.worldpop.org