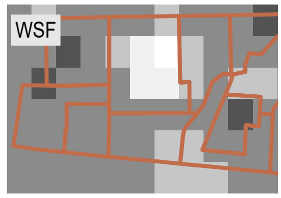
PREDICTING POPULATION IN DATA-SCARCE CONTEXTS

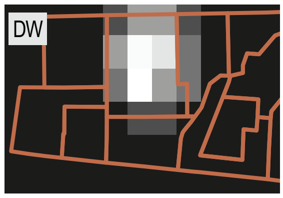
How Bayesian modelling could use the Big Tech competition in producing built-up maps

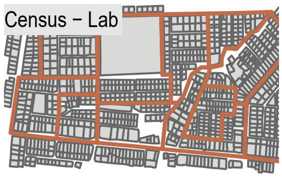
Edith Darin

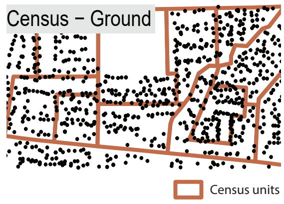
Satellite-imagery derived products represent an exciting opportunity to map and estimate current population with high spatial precision in contexts where traditional demographic data is not available. Recently several high-profile institutions have produced global-scale built-up maps.

How do they compare with human-made maps? What are their impacts on population models? And finally, how could we articulate them together?









Data context

Mali has conducted in 2019 a census cartography, digitising manually all buildings in the country (Census - Lab), collecting population count for 85% of it and collecting building GPS points for 20% of it (Census - Ground). Parallelly, the Joint Research Centre produced a 2020 Global Human Settlement Layer (GHSL) mapping built-up area at ~100m resolution based on Sentinel 2 imagery, as well as a correction for residential area (GHSL - res). The German Space Agency also based their World Settlement Footprint on Sentinel 2 data and created a binary map at 10m resolution for 2019 (WSF). Google computes with same input a land cover classification at 10m resolution near real-time called Dynamic World (DW). Finally Ecopia.AI extracted building footprints from Maxar Imagery as part of the Digitize Africa (DA) initiative that can be summarised either by count or the area sum. Microsoft and Google building footprints do not cover completely the Malian territory.

Population model

We want to investigate the impact of the selected built-up maps on a Bayesian Hierarchical model developed by Leasure et al. to estimate population in data-scarce contexts and applied across sub-saharian Africa.

The model base consists in decomposing the observed population count (POP) between an observed built-up metric (BUILT) multiplied by a population density. Population density distribution is modelled around a location parameter determined by a baseline plus a settlement type-wide variation, and a regional-wide variation.

$$POP \sim Poisson(pop\_density * BUILT)$$
$$pop\_density \sim LogNormal(\alpha_{r,t}, \sigma)$$

$$\alpha_{t,r} = alpha + delta_t + delta_{t,r}$$
$$alpha \sim Normal(5,1)$$
$$delta_t \sim Normal(0, \tau^t)$$
$$delta_{t,r} \sim Normal(0, \tau^r)$$
$$\tau^t \sim HalfNormal(0,1)$$
$$\tau^r \sim HalfNormal(0,1)$$

$$\sigma \sim HalfNormal(0,2)$$

References

Model

Leasure, et al. 2020. "National Population Mapping from Sparse Survey Data: A Hierarchical Bayesian Modeling Framework to Account for Uncertainty." Proceedings of the National Academy of Sciences 117 (39): 24173–79.

Built-up maps

Brown, et al. 2022. "Dynamic World, Near Real-Time Global 10 m Land Use Land Cover Mapping." Scientific Data 9 (1): 251.

Ecopia.AI, and Maxar Technologies. 2019. Digitize Africa Data. http://digitizeafrica.ai.

Marconcini, et al. 2021. "Understanding Current Trends in Global Urbanisation - The World Settlement Footprint Suite." GI_Forum 2021, Volume 9, (June): 33–38.

Pesaresi, et al. 2022. "GHS-BUILT-S R2022a - GHS Built-up Surface Grid, Derived from Sentinel2 Composite and Landsat, Multitemporal (1975-2030)," June.

Fundings

Bill and Melinda Gates Foundation and the United Kingdom Foreign, Commonwealth & Development Office (OPP1182425). Leverhulme Centre for Demographic Science Graduate Scholarship

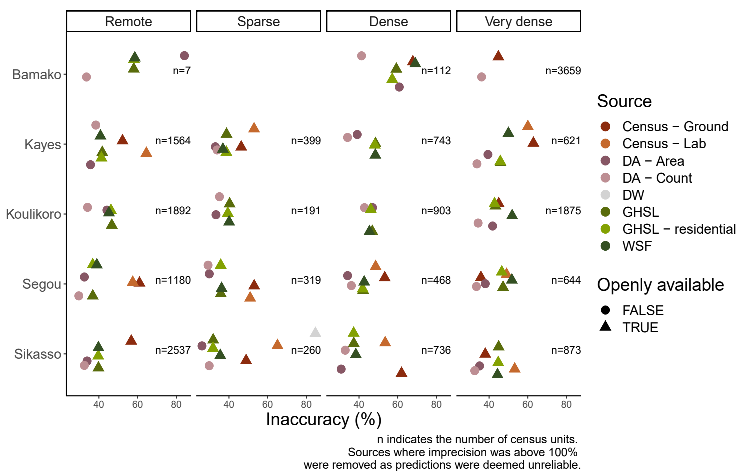
Acknowledgements

We thank Heather Chamberlain, Attila Làzàr, Chris Nnanatu, and Andrew J. Tatem for their fruitful comments and suggestions.

Comparison results

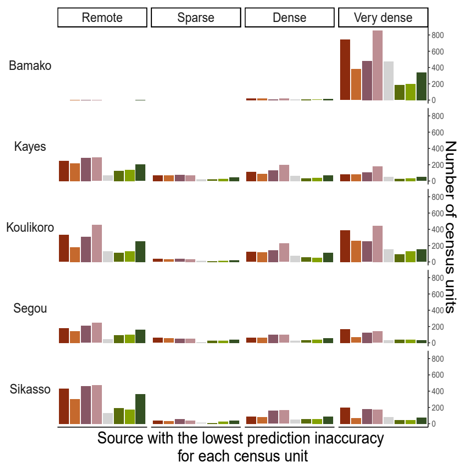
To assess the impact of built-up maps on the population model, we extract the 8 mean predictions for each census unit and compute an *inaccuracy* and an *imprecision* metrics corresponding respectively to the average and the standard deviation of the absolute residuals in percentage of the observed population totals.

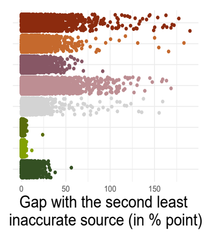
A clear winner at the global scale ...



n indicates the number of census units.
Sources where imprecision was above 100% were removed as predictions were deemed unreliable.

.... masking constrasts at the individual scale





And now, we want to pool the various built-up maps together in an unique model: **Thoughts?**