LINKING POPULATION DATA TO HIGH RESOLUTION MAPS: A CASE STUDY IN BURKINA FASO

Basile Rousse, Sylvain Lobry, Laurent Wendling
LIPADE
Université Paris Cité
name.surname@u-paris.fr

Géraldine Duthé, Valérie Golaz
French Institute for Demographic Studies (INED)

ABSTRACT

Recent research in demography focuses on linking population data to environmental indicators. Satellite imagery can support such projects by providing data at a large scale and a high frequency. Moreover, population surveys often provide geolocations of households, yet sometimes with an offset, to guarantee data confidentiality. In such cases, the proper management of this incertitude is required, to accurately link environmental indicators such as land cover/land use maps or spectral indices to population data. In this paper, we introduce a method based on the random sampling of possible households geolocations around the coordinates provided. Then, we link a land cover map generated using semi-supervised deep learning and a Malaria Indicator Survey in Burkina Faso. After linking households to their close environment, we distinguish several types of environment conducive to high malaria rates, beyond the urban/rural dichotomy.

1 INTRODUCTION

Besides demographic data, population surveys often provide geolocations of interviewed households. For the Demographic and Health Survey (DHS) program[1] and other surveys built with a similar design, these geolocations are slightly displaced to preserve households confidentiality. However, they allow linking households data to their close environment using derived spatial data such as NDVI or more complete land classification schemes. The Local Climate Zones (LCZs) (Stewart & Oke, 2012) is a land cover/land use classification system based on the surface physical properties and human activities. LCZ classes describe different types of environment, as "Compact low-rise" cities, "Sparsely built", "Scattered trees" and "Bush/scrub" areas. This scheme can be applied globally as it is not specific to any part of the World. Recent works on LCZ mapping use machine learning techniques to train models that could be applied globally. Demuzere et al. (2022) proposes a global LCZ map using random forests algorithms on 46 spatial features. The So2Sat dataset proposed by Zhu et al. (2020) is a large-scale labeled dataset for training neural networks to classify LCZs. It is made of 32x32 Sentinel-1/2 patches over 42 cities in various part of the world. This dataset has been used to generate maps for 1642 cities, supporting environmental urban studies (Zhu et al., 2022). Although on a large scale, this strategy does not produce optimal results on sub-Saharan countries as few African cities were available in the training dataset. Recent works focus on semi-supervised domain adaptation strategy based on sub-Saharan seasonal changes to produce more accurate maps. This technique uses the So2Sat dataset and additional Sentinel-2 images to adapt training to the mapping of these countries. It allows the creation of LCZ maps with a resolution of 320m, well below DHS buffer on households' geolocations. Accurately linking such maps and DHS-like population data would benefit from lowering the error made on the selection of the close environment, due to geo-relocations. This work introduces a new method to model the environment of households in DHS studies, and applies it to the Burkina Faso Malaria Indicator Survey (MIS) of 2017-2018 (INSD, 2018). To this end, we first map the whole of Burkina Faso using the LCZ classification scheme and the semi-supervised method. Then, we characterize the households' local environment using randomly sampled areas, that are actual possibilities for their true location. Finally, we show that environmental structures can be a determining factor of malaria prevalence in Burkina Faso.

[1] https://dhsprogram.com

2 DEMOGRAPHIC DATA

DHS buffer on geolocations. To produce representative indicators at a defined study-area level, DHS surveys follow a two-stage sampling procedure: First, enumeration zones (EZs, areas canvassed by one census representative) are randomly sampled. Then, in each of the sampled EZs, households are randomly selected for interviews (Burgert et al., 2013). Each household geolocation is recorded and is grouped in its corresponding enumeration zone. Only coordinates of the centroid of each enumeration zone are provided. To preserve the respondents' confidentiality, they are randomly displaced within a circle of 2-10 kilometers radius (2km in urban areas, 5 for most of the rural enumeration areas with 1% only being displaced of up to 10 km) . The type (urban or rural) of environment is specified. To link EZs to spatial data, DHS recommends to average environmental values on this offset area (Perez-Heydrich et al., 2015). Grace et al. (2019) demonstrates that selecting environmental values of settlements near the DHS' EZs geolocations is a better estimation than averaging over the entire offset area. Other than manually, this selection can be done using gridded population data such as the Global Human Settlement Layer (Pesaresi et al., 2015). However, this method requires having reference data at the time of the study, which is not always the case are.

Malaria Indicator Survey 2017-2018, Burkina Faso. Malaria Indicator Surveys (MISs) are surveys following the general DHS population studies sampling procedure. They focus on monitoring malaria within the context of the global effort to fight this disease. MISs aim to estimate basic demographic and health indicators about malaria as well as the population knowledge about this disease. To this end, malaria rapid tests (giving results in 15 minutes) are done on all 6-59 months old children of the sampled households, with the consent of their legal representatives. The positive tests are confirmed by more reliable laboratory tests. In this paper, we use the MIS 2017-2018 in Burkina Faso. In the survey, malaria prevalence results are representative for each of the 17 study areas (administrative divisions). Among the 252 sampled EZs, 245 were visited. however, 21 of the visited EZs have corrupted geolocations and cannot be used and only 224 EZs can be finally used for analysis. We define the malaria rate R_i of the EZ i as the ratio of the number of positive 6-59 months old children by the total number of 6-59 months old children in the EZ i.

3 LOCAL CLIMATE ZONES MAPPING FOR SUB-SAHARAN AFRICA

Our objective is to extract environmental indicators from remote sensing images that can be linked to malaria rate. We use the Local Climate Zones (LCZs) taxonomy which proposes 17 classes, ranging from dense high-rise built-up areas to water areas. As mentioned in Section 1, training neural networks in a supervised way is not sufficient for accurately predicting LCZs in sub-Saharan countries. However, So2Sat can be used as a reference dataset for building semi-supervised learning to reduce the divergence between a reference domain (i.e. a labeled dataset) and a target domain (i.e. an unlabeled dataset). This type of strategy enables adding useful information to the training without requiring additional labeled data. Sub-Saharan countries experience successive dry and rainy seasons in a year which significantly change vegetation and the overall appearance of the environment, observed by optical remote sensing images. We define $D_S = (x_i, y_i)_{i \in [\![1, n_S]\!]}$ as the labeled dataset where x_i is a Sentinel-2 image from So2Sat, y_i its associated LCZ label and n_S the number of samples in D_S. Our strategy aims to make a model $F(.)$ robust to seasonal change, and learn seasonal features from countries' seasonal changes. Therefore, we define $D_T = (z_i^{s1}, z_i^{s2})_{i \in [\![1, n_T]\!]}$ made of n_T pairs of unlabeled images z^{s1}, z^{s2} from the same area at different seasons s_1 and s_2. As Burkina Faso has one dry and one rainy season, each pair in D_T uses images of both seasons. These two datasets are combined with a consistency regularization approach and contrastive learning. We employed a two tracks training, described in Algorithm 1, resulting in two losses that are combined in a second stage. The first track is a regular supervised track using D_S where Cross-Entropy L_S is computed. The second track uses contrastive learning based on D_T where z_i^{s2} is considered as a seasonal augmentation of z_i^{s2} similarly as in Mañas et al. (2021). The contrastive loss L_T is computed and combined with a weighted sum as follow:

$$L_T(z_i^{s1}, z_i^{s2}) = -\log \frac{\exp(sim(F(z_i^{s1}), F(z_i^{s2}))/\tau)}{\sum_{k=1, k \neq i}^{2N} \exp(sim(F(z_i^{s1}), F(z_k^{s2}))/\tau)} \tag{1}$$

where N is the size of the batch, $sim(\cdot, \cdot)$ is a cosine similarity measure, τ is the temperature, $(i, j) \in [\![1, N]\!]^2$, $(z_l^{s1/2})_{l \in [\![1,2N]\!]}$ samples from the batch, and (z_i^{s1}, z_i^{s2}) a positive pair. The total loss used for back-propagation is a weighted sum of the supervised and unsupervised losses, with a regularization coefficient $\alpha \in [0, 1]$: $L = \alpha \times L_S + (1 - \alpha) \times L_T$.

Algorithm 1 Semi-supervised training step

for $i = 1, 2, \ldots, N_S | j = 1, 2, \ldots, N_T$ **do**
  Predict x_i, z_j^{s1}, z_j^{s2} LCZ classes with $F(.)$
  Compute the Cross-Entropy L_S and Contrastive Loss L_T
  $L \leftarrow \alpha \times L_S + (1 - \alpha) \times L_T$
  Back-propagate L through $F(.)$
end for

Algorithm 2 EZs characterization

X $\leftarrow$ Random areas of all EZs
Separate X in clusters X_1, X_2, X_3, X_4
for $i = 1, 2, \ldots, N_{EZ}$ **do**
  Compute the mean LCZ distribution of EZ i M_i
  Predict the cluster of EZ i M_i with the fuzzy-c-means
end for

To map Burkina Faso, we selected Sentinel-2 tiles taken during the time of the study, in early January 2018. The tiles are split into 32×32 patches for classification, and the model predictions are reconstructed into the final LCZ map. Thus, the resolution of the LCZ map is 320m×320m. More information about the mapping process is available in Rousse et al. (2023).

4 Linking maps to household geolocations

Our objective is to characterize the type of environment of each EZ using the LCZs computed with the method presented in Section 3. We first model the EZs' offset areas by circles centered on their displaced centroids. As already mentioned, these circles have a radius of 2 km in urban areas or 10 km in rural areas. These circles will be referred to as C_k, k being the identifier of the EZ. For each EZ k, we semi-randomly sample n_{random} squared areas of size A (in the $320m \times 320m$ resolution LCZ map) inside C_k to model the potential true geolocations of interviewed households. These n_{random} random areas should have at least $\delta\%$ of LCZs belonging to their urban or rural type to ensure the sampling consistency. This sampling procedure results in a total of $n_{sampled}$ random areas sampled from all EZs. These $n_{sampled}$ samples give a global view on the local environment in which the household were interviewed. We propose to summarize local environments into n_E typical environments. To achieve this categorization, the $n_{sampled}$ samples are clustered into n_E clusters. Finally, we want to characterize the environment of each EZ. The mean LCZ distribution of an EZ k is computed by taking the mean proportion of each LCZ class within all the random areas selected inside C_k. Then, we separate EZs into the n_E types of environment defined by the clustering step described above, using their mean LCZ distributions. We applied this method to the Burkina Faso MIS 2017-2018 and a LCZ map generated using the method presented in 3. In this paper, we use $n_{random} = 10$ and $A = 100m$, which results in a total of 2240 random areas sampled. These areas are clustered using the fuzzy-c-means algorithm (Bezdek et al., 1984) into $n_E = 4$ typical environments. The distributions of the clustering centers are shown in Figure 1. As expected, the 4 clusters are highly dominated by one single LCZ class. Cluster 3 and 4 are rural clusters dominated by bush/scrub and scattered trees areas, cluster 1 is highly urban with compact low-rise (cities) and cluster 2 is in between rural and urban with sparely built areas. A 2D representation computed using the t-SNE algorithm (van der Maaten & Hinton, 2008) of the EZs LCZ distribution is shown in Figure 1.

5 Application: linking malaria to household environments

To explore a possible link between environmental indicators and malaria prevalence, we considered malaria rates computed using MIS data. We plot in Figure 2 the distributions of these malaria rates grouped in the $n_E = 4$ types of environment defined in Section 4 and described in Figure 1. We used $\delta = 0.9$. Malaria rates distributions differ according to the type of environment. The definition of such clusters using the LCZ classification goes beyond the urban/rural dichotomy and enables identifying different structures where the propagation of malaria is higher. The urban clusters 1 and 2, respectively focused on "compact-low-rise" and "sparsely-built", have lower malaria rates than the rural ones. Malaria rates of EZs in cluster 4 bush/scrub are not concentrated in a particular range but are distributed over the entire range of values. This class is widely represented on the Burkina

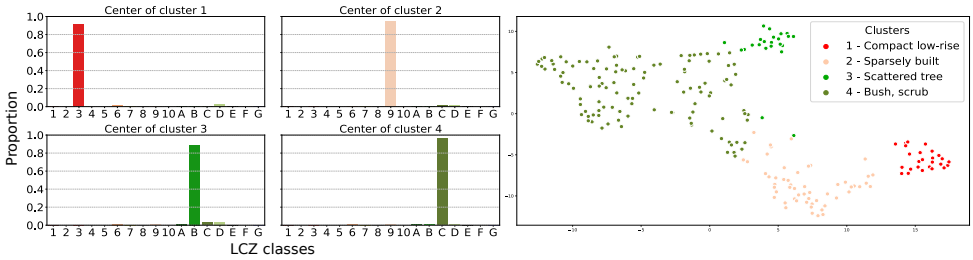

Figure 1: LCZ distributions of clustering centers (left) and visualisation of EZs, grouped by environment types as defined by the clustering step. The visualization has been done using t-SNE.

Faso LCZ map, which implies that it is very represented in our randomly selected areas. This results in about half of the available EZs being associated to this cluster. Nevertheless, as depicted in Figure 2 , the proportion of EZs belonging to cluster 3 is increasing when the malaria rates are increasing. We performed an independence t-test to look at the statistical difference between clusters, based on their malaria rate distributions. Values are shown in Table 1. All are statistically different according to this method except for the rural clusters 3 and 4. The LCZ classification enables distinguishing the types of environment where malaria propagation is lower, or higher. Further developments are required for distinguishing, if possible, scattered trees and bush-scrub areas.

Table 1: Independence t-test results

p-values	Cst1	Cst2	Cst3	Cst4
Cst1	1	0	0	0
Cst2	x	1	0.004	0.033
Cst3	x	x	1	0.160
Cst4	x	x	x	1

t-values	Cst1	Cst2	Cst3	Cst4
Cst1	0	7.706	9.223	11.899
Cst2	x	0	3.004	2.151
Cst3	x	x	0	-1.433
Cst4	x	x	x	0

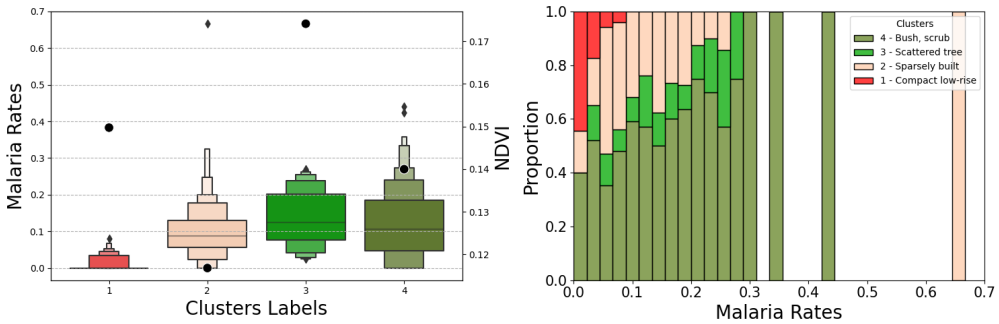

Figure 2: Distribution of EZs malaria rates grouped by cluster (left) and proportion of malaria rates, grouped by cluster and by intervals.

6 Conclusion

In this paper, we introduce a new method to model households geo-relocations on DHS surveys. We based our method on a semi-random sampling of small areas within DHS offset areas, which are possible real close environment of interviewed households. Then, we applied this method to link the LCZ environment classification scheme to the MIS survey in Burkina Faso. This method, when linked with a LCZ map, was able to distinguish different environmental structures where malaria propagation is either higher or lower. Further development on the mapping and incertitude management must be done to clarify these distinctions.

Acknowledgments

Work supported by the Data Intelligence Institute of Paris (DiiP), and IdEx Université Paris Cité (ANR-18-IDEX-0001). This work was performed using HPC resources from GENCI-IDRIS (Grant 2021-AD011013527).

REFERENCES

James C. Bezdek, Robert Ehrlich, and William Full. Fcm: The fuzzy c-means clustering algorithm. *Computers Geosciences*, 10(2):191–203, 1984. ISSN 0098-3004. doi: https://doi.org/10.1016/0098-3004(84)90020-7. URL https://www.sciencedirect.com/science/article/pii/0098300484900207.

Clara Burgert, Josh Colston, Thea Roy, and Blake Zachary. Geographic displacement procedure and georeferenced data release policy for the demographic and health surveys. 2013. doi: 10.13140/RG.2.1.4887.6563.

M. Demuzere, J. Kittner, A. Martilli, G. Mills, C. Moede, I. D. Stewart, J. van Vliet, and B. Bechtel. A global map of local climate zones to support earth system modelling and urban-scale environmental science. *Earth System Science Data*, 14(8):3835–3873, 2022. doi: 10.5194/essd-14-3835-2022. URL https://essd.copernicus.org/articles/14/3835/2022/.

Kathryn Grace, Nicholas N. Nagle, Clara R. Burgert-Brucker, Shelby Rutzick, David C. Van Riper, Trinadh Dontamsetti, and Trevor Croft. Integrating environmental context into dhs analysis while protecting participant confidentiality: A new remote sensing method. *Population and Development Review*, 45:1, 2019.

INSD. Enquête sur les indicateurs du paludisme au burkina faso. 2018. URL https://dhsprogram.com/pubs/pdf/MIS32/MIS32.pdf.

Oscar Mañas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodríguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *2021 IEEE/CVF International Conference on Computer Vision*, pp. 9394–9403, 2021. doi: 10.1109/ICCV48922.2021.00928.

Carolina Perez-Heydrich, Joshua Warren, Clara Burgert, and Michael Emch. Influence of demographic and health survey point displacements on raster-based analyses. *Spatial Demography*, 4:1–19, 06 2015. doi: 10.1007/s40980-015-0013-1.

Martino Pesaresi, Daniele Ehrlich, Aneta Florczyk, Sergio Freire, Andreea Julea, Thomas Kemper, Pierre Soille, and Vasileios Syrris. Ghs built-up grid, derived from landsat, multitemporal (1975, 1990, 2000, 2014). *European Commission, Joint Research Centre (JRC)*, 2015. URL http://data.europa.eu/89h/jrc-ghsl-ghs_built_ldsmt_globe_r2015b.

Basile Rousse, Sylvain Lobry, Géraldine Duthé, Valérie Golaz, and Laurent Wendling. Seasonal semi-supervised domain adaptation for linking population studies and Local Climate Zones. *Accepted at Joint Urban Remote Sensing Event - JURSE*, 2023.

Iain D. Stewart and Timothy R. Oke. Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society*, 93:1879–1900, 12 2012. doi: 10.1175/BAMS-D-11-00019.1.

Laurens van der Maaten and Geoffrey Hinton. Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.

Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Jian Kang, Lichao Mou, Hossein Bagheri, Matthias Haberle, Yuansheng Hua, Rong Huang, Lloyd Hughes, Hao Li, Yao Sun, Guichen Zhang, Shiyao Han, Michael Schmitt, and Yuanyuan Wang. So2sat lcz42: A benchmark data set for the classification of global local climate zones [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 8(3):76–89, 2020. doi: 10.1109/MGRS.2020.2964708.

Xiao Xiang Zhu, Chunping Qiu, Jingliang Hu, Yilei Shi, Yuanyuan Wang, Michael Schmitt, and Hannes Taubenböck. The urban morphology on our planet – global perspectives from space. *Remote Sensing of Environment*, 269:112794, 2022. ISSN 0034-4257. doi: https://doi.org/10.1016/j.rse.2021.112794. URL https://www.sciencedirect.com/science/article/pii/S0034425721005149.