OpenStreetMap and Sentinel-2 data for the automatic production of environmental indices for demographic studies

Master internship

General information

- Keywords: Remote sensing, Deep learning, Sentinel 2, Local climate zones, Africa, OpenStreetMap, Population
- Duration of the internship: 6 months (standard stipend). To start between February and April 2023.
- Institutes: Université de Paris, Laboratoire d'Informatique Paris Descartes (LIPADE), équipe Systèmes Intelligents de Perception et Institut national d'études démographiques (INED), unité Démographie des pays du Sud (Demosud)
- Location: 45 rue des Saints-Pères, 75006 Paris (LIPADE) and 9, cours des Humanités, 93322 Aubervilliers (INED)
- Supervision: Basile Rousse, Sylvain Lobry, Laurent Wendling - (first.lastname@u-paris.fr), Valérie Golaz, Géraldine Duthé (first.lastname@ined.fr)
- **Application: Please send a cover letter and a CV to stage-diip@listes.ined.fr. You will receive a confirmation by email. The position is open until filled.**

Proposed topic

Motivation

In a globalized context increasingly impacted by climate change, demographic studies would gain from taking environmental data into account and be carried out at the transnational level. However, this is not always possible in sub-Saharan Africa, as matching harmonized demographic and environmental data are seldom available. The large amount of data regularly acquired since 2015 (in 2019 only, Sentinel satellites from the European Space Agency produced 7.54 PiB of open-access data) provides an opportunity to produce relevant standardized indicators at the global scale. In particular, Sentinel-2 satellites produce multi-spectral images at a high temporal frequency (maximum 5 days) and at medium resolutions (10 to 60 meters depending on the spectral band). These images are both freely available and enable land-cover/land-use mapping. However, Sentinel-2 images alone may not be sufficient for all remote sensing tasks. For instance, fine-grained land cover mapping requires data with higher resolution, or other sources which can add to Sentinel-2 images. Several indicators and legends have been developed to help understanding geographical realities in a consistent (i.e. not location dependent) manner. Among them, Local Climate Zones (LCZs) have been initiated by WUDAPT (World Urban Database and Access Portal Tools) to systematically label urban areas [1]. LCZs consist in 17 urban and rural classes built on surfaces properties defined without any cultural considerations. They enable describing both cities and rural areas according to their vegetation type, densities, heights and materials. Houses with light materials may not be detected by deep learning

models when trained on Sentinel-2 data because of their resolution. These misdetections will alter LCZ classification results, especially for urban and semi-urban areas. Other data sources are required to increase mapping accuracy. For instance, very-high resolution orthophotos are often used for such tasks. However, as they are expensive to acquire, they are not available globally. Furthermore, they do not allow for the frequent updating of land-cover maps. Collaborative data from OpenStreetMap (OSM) could add useful information. This internship will focus on the combination of OSM data and Sentinel-2 images for generating LCZ maps of sub-Saharan countries.

Background and state of the art

The LCZ classification is meant to provide a way of mapping the world, in open-access, that can later be used by researchers for a wide range of studies. LCZ data have been used to understand energy usage [2], climate [3] or geoscience modeling [4]. An important amount of work has been dedicated in the recent years to the automatic generation of such data, from sensors such as Landsat 8 or Sentinel 2. In a research competition organized by the IEEE IADF, several methods have been tried out to map LCZ from Landsat, Sentinel 2 and OpenStreetMap data [5]. Recent studies focus on the use of deep learning models to tackle the task of automatically mapping LCZs [6, 7]. However, none of them focuses on sub-Saharan Africa where data are scarce. For instance, the So2Sat dataset of [7] is made of labels of 42 cities among which only two are located in sub-Saharan Africa. This is problematic, as spatial variations within different sub-Saharan countries can be large, and that spatial generalization of machine learning based methods is a challenge [8]. To overcome this issue, recent studies are based on supervised and contrastive learning to extract useful global features, using the So2Sat dataset, and more specific local features, using unlabeled data. This technique produces LCZ maps that provide enough information to contextualise localised demographic information.

OSM is a collaborative platform providing a geo-referenced database of the world. Contributors and users can create and use up-to-date maps for geographic applications. It has been widely used for mapping tasks [9] or LCZ classification [10] with a positive impact when data quality is sufficient. Whereas OSM is well developed and regularly updated in urban areas in Western countries, it is not necessarily up-to-date in Sub-Saharan countries because of the very fast pace of change in land use and settlement related to population growth and development. For instance, [11] states common errors in OSM dataset for rural areas :

- Annotations are geometrically misaligned or not well geolocalized,
- buildings exist but are not annotated as shown in figure 1,
- buildings are annotated but do not exist (misannotation or building destroyed).

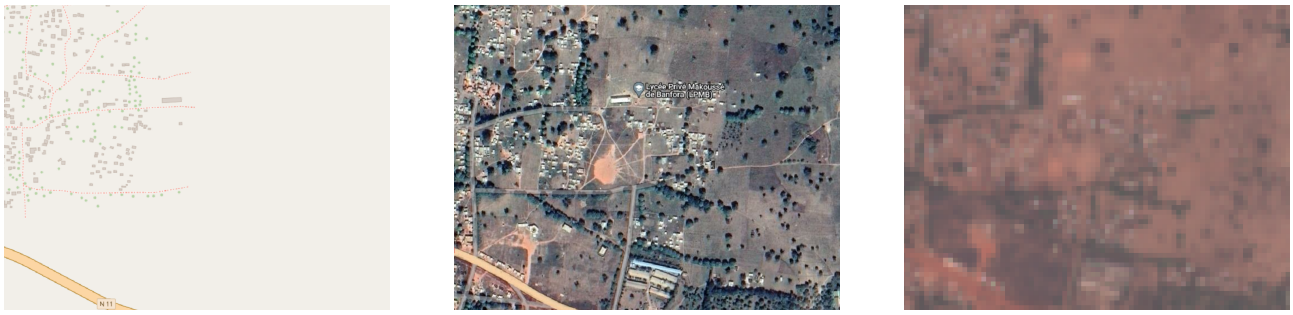

Figure 1: OSM data (left), google maps image (center), and Sentinel-2 image (right) of Banfora in Burkina Faso. Some buildings are annotated, but some buildings far from the center are not notified.

Work to be done

The work to be conducted during the proposed M2 internship will lead to the following three contributions:

- Contribution A: **Development of a model to classify LCZs using high quality OSM data.**
  This rule-based model will allow to better understand the LCZ classification scheme. Furthermore, it will provide a baseline to the multi-modal methods to be developed during the internship.

- Contribution B: **Multi-modal models for LCZ classification**
  Using an already trained deep-learning based method to classify LCZs, we will study different fusion mechanisms (including late fusion, rule based fusion) to integrate the information from the rule-based model. Furthermore, we will develop an end-to-end deep learning based model taking rasterized OSM and Sentinel-2 data as an input. These methods will be compared and evaluated in Ouagadougou, Burkina Faso and Antananarivo, Madagascar.

- Contribution C: **Link with demographic studies and writing of the master thesis**
  The obtained results will be linked to demographic data in the two previously mentioned regions to better understand the underlying geo-spatial components in population studies. These results will be compared with a baseline developed during the PhD of Basile Rousse.

Desired background

We are looking for a student in Master 2 or final year of MSc, or engineering school in computer science. The ideal candidate would have knowledge in image processing, computer vision, machine learning, geo-information sciences and Python programming and an interest in handling large amount of data, remote sensing and demography. An experience in statistical data analysis would be a plus.

Bibliography

[1] Benjamin Bechtel et al. "Mapping local climate zones for a worldwide database of the form and function of cities". In: *ISPRS International Journal of Geo-Information* 4.1 (2015), pp. 199–219.

[2] Paul John Alexander, Gerald Mills, and Rowan Fealy. "Using LCZ data to run an urban energy balance model". In: *Urban Climate* 13 (2015), pp. 14–37.

[3] Jan Geletič et al. "Spatial modelling of summer climate indices based on local climate zones: expected changes in the future climate of Brno, Czech Republic". In: *Climatic Change* 152.3-4 (2019), pp. 487–502.

[4] Hendrik Wouters et al. "The efficient urban canopy dependency parametrization (SURY) v1. 0 for atmospheric modelling: description and application with the COSMO-CLM model for a Belgian summer". In: *Geoscientific Model Development* 9.9 (2016), pp. 3027–3054.

[5] Naoto Yokoya et al. "Open data for global multimodal land use classification: Outcome of the 2017 IEEE GRSS Data Fusion Contest". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11.5 (2018), pp. 1363–1377.

[6] Chunping Qiu et al. "Feature importance analysis for local climate zone classification using a residual convolutional neural network with multi-source datasets". In: *Remote Sensing* 10.10 (2018), p. 1572.

[7] Xiao Xiang Zhu et al. "So2Sat LCZ42: A Benchmark Data Set for the Classification of Global Local Climate Zones [Software and Data Sets]". In: *IEEE Geoscience and Remote Sensing Magazine* 8.3 (2020), pp. 76–89. DOI: 10.1109/MGRS.2020.2964708.

[8] Emmanuel Maggiori et al. "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark". In: *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. 2017, pp. 3226–3229.

[9] John E Vargas-Munoz et al. "OpenStreetMap: Challenges and opportunities in machine learning and remote sensing". In: *IEEE Geoscience and Remote Sensing Magazine* 9.1 (2020), pp. 184–199.

[10] Patricia Lopes et al. "Using OpenStreetMap data to assist in the creation of LCZ maps". In: *JURSE*. Mar. 2017.

[11] John E Vargas-Muñoz et al. "Correcting rural building annotations in OpenStreetMap using convolutional neural networks". In: *ISPRS journal of photogrammetry and remote sensing* 147 (2019), pp. 283–293.