Between national statistics and local observations: mapping slums using machine learning model.

Urban land is divided rather unevenly, globally. An extreme example being slums, where often large numbers of households are cramped together on a small area. A slum household is defined as a household that suffers one or more of the following deprivations: lack of access to improved water source; lack of access to improved sanitation facilities; lack of sufficient living area; lack of housing durability and lack of security of tenure. Knowing the location and density of slums is essential for monitoring Sustainable Development Goals (SDGs), for example in the context of vulnerability to natural hazards, climate change impacts, and human wellbeing.

Slums are generally detected and monitored in two different ways. First, census data, as well as large-scale surveys, provide information on the total amount of people that fulfil the criteria of slum households at the national or regional scale. They are typically the sources underlying indicators used to report progress towards the SDGs but have limited contribution to regional planning due to a lack of spatial information. Second, many studies have used satellite imagery to identify slums or informal settlements by machine models, providing rather exact information on slums distribution but covering a relatively small area, like within a city. Applying this approach in larger areas is constrained by the costs of very-high resolution imagery and computations. More importantly, results produced by these two approaches show the inconsistency as local acceptance for the model-based slum mapping varies among stakeholder groups. Therefore, a locally adapted framework is required to combine ground surveys with robust machine learning methods, and to allow the rapid extraction of consistent information on the dynamics of slums at a region scale.

Here, we propose an innovative framework to map the total population living in slums in a spatially explicit way using a Random Forest classification model to determine the likelihood that a building is occupied by a slum household. This can be achieved by six consecutive steps: 1) Extract likely residential buildings from the Google Building Footprints data. 2) Collect reference data on buildings that are known to be slums from study cases and surveys. 3) Select and process explanatory variables for slum households. 4) Build a Random Forest model to predict the probability that each building may contain slum households. 5) Allocate the total population living in slums to each building by combining the probability generated by RF model and household number from census data. 6) Aggregate the buildings with slum households to the pixel at 1km^2 resolution.

This study will generate a map indicating the number and share of slum households for Kenya in 2020, as well as the framework that can be applied to other regions, contributing to the achievement of sustainable cities and communities in regional planning.