Development Economics Research Group

Working Paper Series

17-2022

# Urban Poverty Mapping with Open Spatial Data: Evidence from Dar es Salaam

Peter Fisker

Kenneth Mdadila

December 2022

ISSN 2597-1018

University of Copenhagen
Faculty of Social Sciences
Department of Economics
www.econ.ku.dk/derg
#DERGDK

# Urban Poverty Mapping with Open Spatial Data: Evidence from Dar es Salaam

Peter Fisker* Kenneth Mdadila†

December 2021

**Abstract**

This paper combines information from a representative household survey with publicly available spatial data extracted from satellite images to produce a high-resolution poverty map of Dar es Salaam. In particular, it builds a prediction model for per capita household consumption based on characteristics of the immediate neighborhood of the household, including the density of roads and buildings, the average size of houses, distances to places of interest, and night-time lights. The resulting poverty map of Dar es Salaam dramatically improves the spatial resolution of previous examples. Extreme Gradient Boosting (XGB) performs best in predicting household consumption levels given the input data. This result demonstrates the simplicity with which policy-relevant information containing a spatial dimension can be generated.

**JEL codes:** O18, Q54, R11

**Keywords:** Poverty, small-area estimation, building footprints, prediction models

*University of Copenhagen, Development Economics Research Group, pkf@econ.ku.dk.
†University of Dar es Salaam, School of Economics

# 1 Introduction

The metropolitan region of Dar es Salaam, Tanzania, covers a vast area and encompasses five main districts with a total of 90 wards (third level administrative units), according to the population census of 2012. The region has a population of 5.4 million according to population census of 2022 and a basic needs poverty rate of 8 % according to the latest poverty assessment [World Bank, 2020]. While significantly lower than in the rural (31 %) and other urban areas of Tanzania (19 %), urban poverty not reduced between 2012 and 2018 (15.4 % to 15.8 %) and is likely to be underestimated due to continued urbanization and the Covid-19 pandemic.

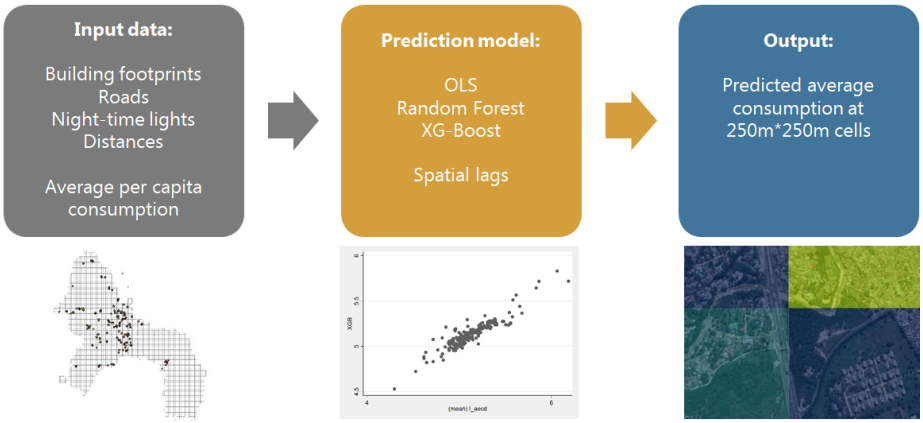
Poverty incidence is typically monitored using nationally representative household surveys, of which the latest and most commonly used in the case of Tanzania is the Household Budget Survey (HBS) 2017/18. Based on this dataset, regional or even district level poverty maps can be produced to guide poverty reduction strategies and development policy in general. However, in order to gain detailed knowledge on the distribution of poverty within a city like Dar es Salaam, which represents a special case of urbanization in Tanzania, the survey data is not suitable, since it has not been sampled to be representative at the level of wards. Of the 90 wards located within the region of Dar es Salaam, the 797 household interviews are distributed across 60 wards with either one or two enumeration areas and thus only up to 12 or 24 households interviewed in each.

It is the purpose of this paper to show that it is possible to generate high-resolution poverty maps of urban areas using readily available open spatial data. We use the types, sizes and density of buildings, distances to and length of roads of various qualities, and the intensity of nighttime lights to predict average consumption at the local level. To this end, the city is divided into grid cells, approximately the size of enumeration areas, and average household consumption aggregates estimated using the indicators mentioned above and machine learning methods. The output of the exercise is useful for various

policy targeting purposes, for instance poverty alleviation, disaster response, and infrastructure investments. Moreover, the visualizations provide a present and all-encompassing view of the urban area, including newer peri-urban settlements often not represented in official statistics.

An overview of the study is provided in figure 1. At the basic level, we combine input data from a household survey with available open source spatial data. Based on the overlaps between these, we build a prediction model and test three common methods before finally generating the outputs, which in this case can be visualized on a map.

Figure 1: Study overview




Recent years have seen a growing interest in the use of building footprints and other open spatial data for urban poverty mapping. A key advantage of using open spatial data for poverty mapping is the high spatial resolution. This allows for more precise mapping of poverty at the neighborhood level, enabling a better understanding of the spatial distribution of poverty within cities. This can be particularly useful for identifying areas of concentrated poverty, which may require targeted interventions to reduce poverty [Vijay and Patnaik, 2018].

Another advantage is availability and accessibility. Unlike traditional household surveys, which can be expensive and time-consuming to collect, open spatial data

can be easily obtained from a variety of sources, such as government agencies and online platforms [Garcia et al., 2020]. This makes it possible to generate poverty maps on a regular basis, enabling policymakers to monitor changes in poverty over time.

However, the quality of the data may vary, depending on the source and the data collection methods used [Amendola et al., 2019], which can potentially lead to errors and biases in the resulting poverty maps. Another limitation is the lack of socio-economic characteristics within households. This can make it difficult to use these data sources to identify the specific households and individuals who are living in poverty, [Vijay and Patnaik, 2018, Garcia et al., 2020].

Despite the limitations, recent research has shown that these data sources can be used to generate reliable and accurate poverty maps. For example, Amendola et al. [2019] used building footprint data and satellite imagery to map poverty in Ghana, finding that the resulting maps were highly correlated with official poverty estimates. Likewise, Sohnesen et al. [2021] and Fisker et al. [2022] have used similar techniques to map poverty in urban areas of Mozambique and São Tomé.

# 2 Data

## 2.1 Household survey

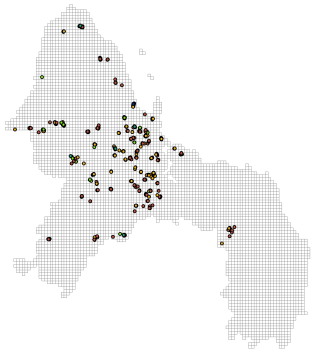
This study relies on the Household Budget Survey (HBS) 2017/18, which contains information on monthly per capita consumption along with many other household characteristics, and importantly, recordings of latitude and longitude that are accurate (i.e. non-scrambled). We use logarithm of adult equivalent per capita consumption, adjusted by spatial differences to measure and rank households by poverty status.

Dar es Salaam represents a special case of urbanization in Tanzania with a population share of 8.6% and average annual growth rate of 5.6 %, compared

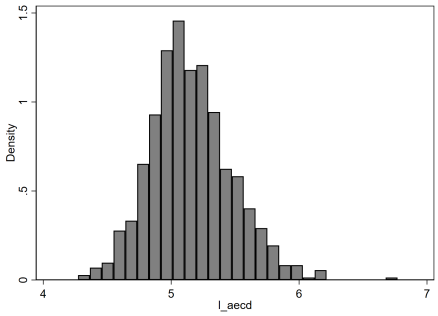
to the national average rate of 2.7%. Sampling design of the HBS takes this fact into account. Based on HBS 2017/18, there are 797 households and 3,276 individuals in the sample, distributed in 60 wards. Figure 2 shows the locations of interview points and the distribution of the outcome variable to be estimated in this study.

Figure 2: Survey data

(a) HBS 2017/18 locations



(b) Distribution of outcome variable




Source: Authors' illustration based on HBS 2017/18 and OSM.

The HBS 2017/18 is a nationally representative survey, sampled randomly based on enumeration areas drawn in preparation for the 2012 population census. Taking into account the urban development that has taken place in Dar es Salaam during the past decade, creating a poverty map solely based on these survey points would likely misrepresent the situation in areas that have seen large growth. Furthermore, the survey only covers a sub-set of census enumeration areas, which implies a risk of omission of the tails of the welfare distribution.

## 2.2 Spatial data

### 2.2.1 Building footprints

In the absence of high-quality consumption data in household surveys, economists often rely on proxies, easily identified by enumerators, for instance when en-

rolling beneficiaries to social protection programs. Such proxies include building materials of the walls, floors, and roofs, toilet facilities, and number of rooms per inhabitant. Some of these indicators of poverty can also be derived from satellite imagery; and when aggregated to units the size of a neighborhood block, idiosyncratic variation (e.g. a rich family living in a tiny house) can largely be disregarded. One of the most promising sources of large-scale, satellite-based information on living conditions are building footprints.

A building footprint is a polygon outline that defines the outline of a building as seen from above. When aggregated to cells of 250 m x 250 m, three essential pieces of information can be derived:

i Average area of the buildings (larger homes are expected in richer areas).

ii Average perimeter (more complex structures have a longer perimeter relative to the area).

iii Number of buildings (proxy for population density).

Footprints larger than 400 square meters are filtered out in order to exclude most industry facilities, offices, and public buildings. Partly, this will also mean excluding large apartment buildings commonly located near the coastline and downtown area.

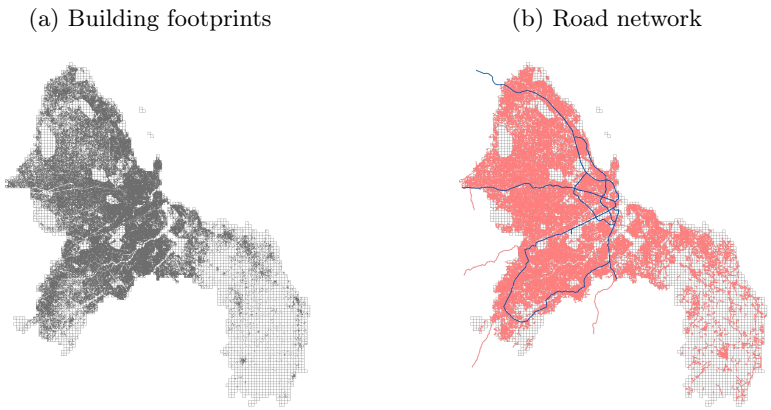
In Dar es Salaam, a complete coverage of all buildings has been added to OpenStreetMaps by a volunteer project called *Dar Ramani Huria*.[1] A total of $1,323,748$ buildings are available in Dar es Salaam region, out of which $745,989$ were mapped by around 150 university students between 2015 and 2017 in the first stage of the project, aimed at improving flood resilience in vulnerable areas.

### 2.2.2 Roads and distances

Information on the road network is obtained from OpenStreetMaps (OSM), which is a global collaborative project to create a free, editable map of the world. The

[1] see ramanihuria.org

Figure 3: Open spatial data



(a) Building footprints (b) Road network

Source: Authors' illustration based on OSM data.

road network dataset contains roads, paths, highways, and other transportation features. In this study, we focus on two road feature types: *primaryroads*, which are important for easy access to jobs and activities in other parts of the city, and *residentialroads*, which are indicators of local urban development.

For each category of road, we calculate the total road length within a cell as well as the smallest geodesic distance from the cell centroid to the nearest road. For primary roads, it is expected to be the distance measure that contributes the most to the prediction model, while for residential roads, the total length is likely to be of greater importance, due to the above-mentioned considerations.

Furthermore, for each cell centroid we calculate the direct distance to the city centre, in this case the DTV roundabout in the financial downtown of Dar es Salaam. The reasoning behind the inclusion of this variable is that the city centre is typically home to the most desirable jobs, local businesses, and amenities. As a result, those who live closer to the city centre are more likely to have higher incomes, better access to employment, and a wider selection of leisure activities. Additionally, those who live closer to the city centre tend to benefit from higher property values, due to the increased demand for housing in these areas.

### 2.2.3 Night-time lights

Additional layers of geo-spatial data can be included in the prediction model from satellite records. Here, we complement the polygon and line features with raster data on artificial lights at night. At a resolution of roughly 250 m x 250 m (similar, but not aligned to our gridded units of observation), we rely on the VIIRS (Visible Infrared Imaging Radiometer Suite) Night-Time Lights (NTL) data. The data is produced from an instrument on the Suomi NPP satellite, and is often used to monitor changes in urbanization, economic activity, population, and other environmental processes. Here, we use the most recent yearly composite from 2021, while monthly and daily images are also available.

## 2.3 Combining the data

The geographical extent of Dar es Salaam region is the area of the study. A grid of cells comprising around 250 m x 250 m is defined, generating a dataset of 5,334 units. In 165 of these cells there is an overlap with survey points from the HBS 2017/18, as shown in figure 2. These are the data points on which the prediction models are based. All cells have been intersected with all information on geo-spatial features such as building footprints, roads, distances, and NTL. The predictions based on the cells where consumption data is available are then applied to all cells to produce city-wide estimates of average per capita household consumption.

Regarding the choice of cell size, it is important to take into account factors such as usefulness, practical purposes, model performance, and computer requirements. The unit should be small enough that policy targeting becomes accurate, while also being large enough to ensure meaningful results and avoid being driven by outliers. Moreover, units of around 250 m x 250 m can reduce potential measurement errors with the position recorders for the household survey data.

# 3 Methods

The methodological approach for producing a high-resolution poverty map of Dar es Salaam using open spatial data is broadly explained in figure 1, where input data is to set up and combined, then applied to various prediction models, and finally visualized. In this section we briefly present the prediction strategies applied to the data and discuss the role of spatial dependence between our units of observation.

## 3.1 Prediction models

Following the construction of the cell-level dataset, we first predict the outcome variable using Ordinary Least Squares (OLS). This method, which is standard in economic analyses has the advantage of being relatively simple and intuitive when it comes to interpretation of the results. It can give an overview of the individual effects of variables when linearity is assumed. However, it does not take into account non-linearities and interactions between the predictors.

As an opposite approach, we also employ Extreme Gradient Boosting (XGB), which is a boosting algorithm that uses a gradient descent-based optimization technique to improve the accuracy of a prediction model [Chen and Guestrin, 2016]. While other machine learning methods such as Random Forest (RF) uses iterations of random subsets of variables and observations, essentially leaving out weak predictors, XGB goes through a similar process, but exploits excluded variables to predict residuals. This means that more *weak learners* are included in the models to focus on areas neglected by RF. XGB is also better at handling missing data than RF and is more effective in dealing with highly unbalanced datasets.

## 3.2 Spatial dependence

In both our prediction models, we include spatially lagged explanatory variables. This means that for all variables, we calculate the average value in neighboring

cells and add those values as new variables in the prediction model. This is because factors that influence living standards in one specific cell are also expected to affect surrounding cells. For instance, a unit of observation with no roads will be better off if all neighbouring units are connected to roads than if none of them are. Likewise, being surrounded by neighborhoods with high-quality infrastructure and housing will positively affect property values of a given cell, ceteris paribus. Apart from improving the predictive power of the models, this furthermore has a smoothing effect on the final predictions.

# 4 Results

## 4.1 Variable contributions

While all variables included in the models contribute to the predicted outcome values, some are more important than others. Using OLS, Table 1 lists the correlation coefficients and significance levels of the main cell-level spatial covariates. Column 1 is the baseline model while column 2 controls for spatial lags of all variables and column 3 in addition contains ward-level fixed effects.

In our preferred model (column 2), which includes spatially lagged explanatory variables, but no ward fixed effects, the most important cell-level predictors of average consumption are i) distance to the city centre, ii) distance to nearest residential road, iii) night-time lights, and iv) avg. building perimeter. As can be seen from column 3, adding ward-specific intercepts improves the resulting $R\hat{2}$; however, as this study is more concerned with the explanatory power of open spatial data, the inclusion of administrative boundaries may shroud these results.

Turning to the results of the prediction based on Extreme Gradient Boosting, figure 4 shows the importance scores of the most important variables. In this set-up, the predictor which is most likely to be included is the spatial lag of the distance to nearest residential road followed by night-time lights, length of

Table 1: Predictors of cell-level consumption using ordinary least squares

| | 1 | 2 | 3 |
|---|---|---|---|
| Avg. building area | -0.00220 | -0.00525 | 0.00550 |
| Avg. building perimeter | 0.0179 | 0.0215* | -0.0107 |
| Number of buildings | -5.63e-05 | -5.86e-05 | -3.15e-05 |
| Dist. city centre | -1.01e-05** | -1.77e-05*** | -0.000101*** |
| Dist. prim. road | -1.98e-06 | 6.85e-05 | -0.000195 |
| Dist. res. road | -0.000833* | -0.00115** | -0.000355 |
| Length of prim. roads | 3.34e-05 | -1.86e-05 | -0.000197* |
| Length of res. roads | -4.05e-06 | -1.05e-06 | -2.51e-05 |
| Night-time lights | -0.000677 | 0.0110* | 0.0111 |
| Spatial lags | NO | YES | YES |
| Ward fixed effects | NO | NO | YES |
| Constant | 4.881*** | 5.127*** | 9.215*** |
| | (0.269) | (0.512) | (1.698) |
| Observations | 165 | 165 | 165 |
| R^2 | 0.135 | 0.245 | 0.668 |

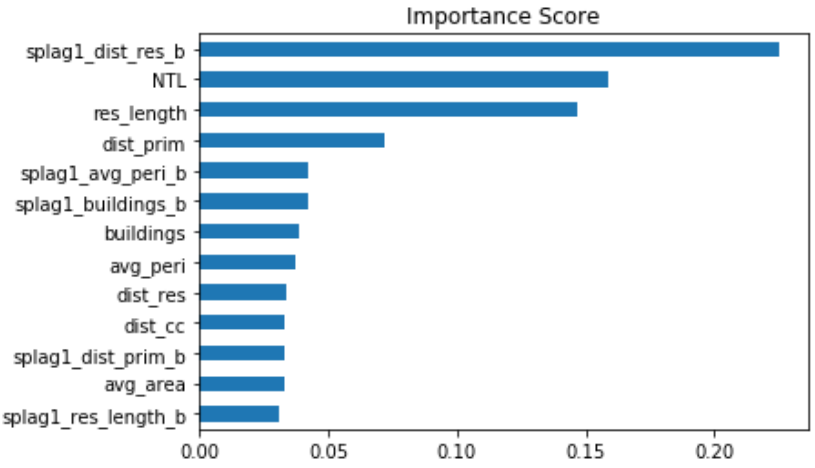
Note: *, **, *** indicate significance at 0.1, 0.05, and 0.001 level, respectively.

residential roads, and distance to nearest primary road. Only thereafter comes a number of variables related to building footprints.

### 4.2 Actual vs predicted values

Using the OLS model presented in Table 1, column 2, we found an $R\hat{2}$ of 0.24 with spatial lags and no ward fixed effects, and a within-sample Spearman rank correlation of 0.44. Extreme Gradient Boosting achieved a within-sample Spearman rank correlation of 0.87, and an out-of-sample Spearman rank correlation

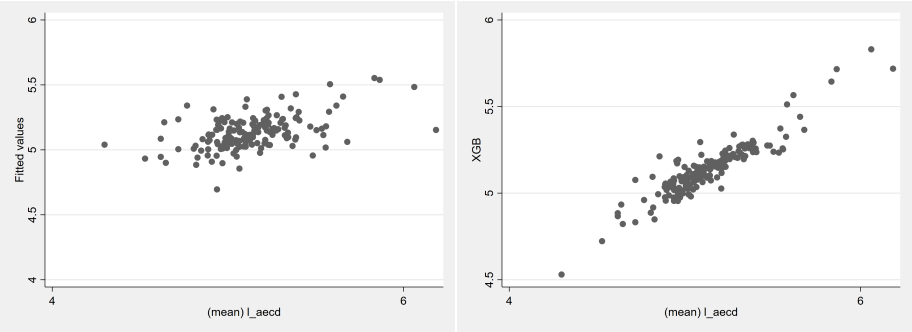
Figure 4: Variable importance, XGB




of 0.25. Our results suggest that XGB is more effective at predicting outcomes than OLS. However, the out-of-sample rank correlation shows that the predicted outcomes in cells with no survey data are not perfect; the spatial data can only be expected to be able to explain a moderate part of the variation in consumption between these cells. Figure 5 shows scatter plots of actual and predicted values for both OLS and XBG.

Figure 5: Actual vs predicted values

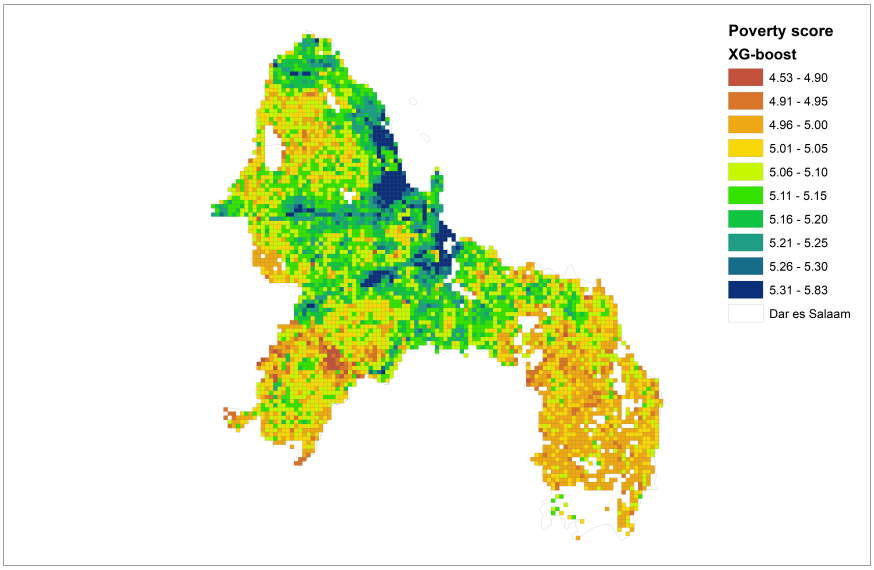
(a) OLS (b) XGB



### 4.3 Visualizations

Based on the outcomes of the XGB model, all areas of the Dar es Salaam region can be attributed a value for its predicted average (logarithm of) per capita consumption. This is illustrated on figure 6 using a color scale ranging from red (poor) over green (average) to blue (relatively well-off). Most of the poorer areas seem to be scattered on the periphery of the urban area, while a few pockets of poverty persist closer to the downtown area. Figure 7 exemplifies the level of detail contained in the map; the zoom-in reveals more organized, larger houses underneath the blue shade

Figure 6: Visualization

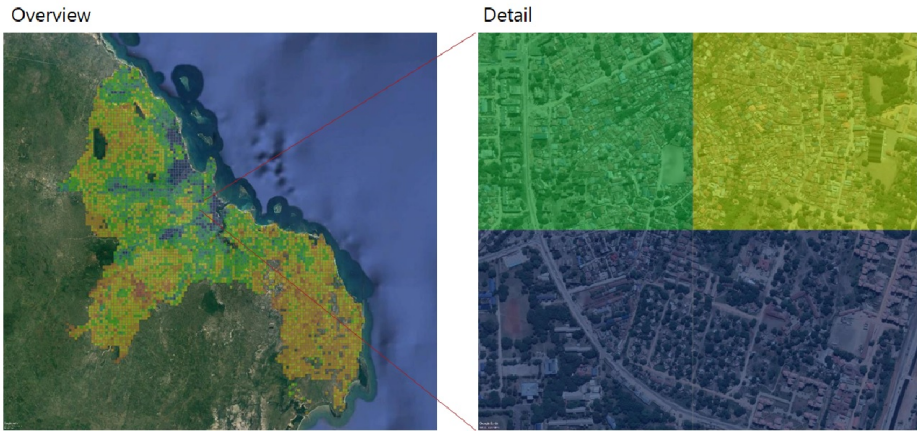



Source: Authors' illustration

## 5 Conclusion and discussion

The poverty maps provide an important tool that can be used to target policies aimed at reducing poverty and improving the quality of life for citizens in a

Figure 7: Visualization detail




Source: Authors' illustration on Google Earth background

given area. Interventions such as social protection, social pensions and disaster or pandemic response can be more effectively targeted using poverty maps. For instance, around 100,000 households in Dar es Salaam are currently enrolled in the Productive Social Safety Net (PSSN), a cash-for-work program administered by the Government of Tanzania. Currently, the beneficiaries are distributed across the 90 Wards of Dar es Salaam, and with the maps presented in this study, a potential future re-targeting exercise could improve efficiency and fairness due to better spatial resolution and coverage.

In addition, poverty maps can be used to inform urban planning decisions such as the location of new schools, hospitals, roads and other public infrastructure.

Finally, as in most other urban centres of Sub-Saharan Africa, urbanization is continuously adding new, largely unregistered peri-urban neighborhoods to the city of Dar es Salaam. These areas seem to be poorer on average than more established neighborhoods, and putting them on a map is an important first step for inclusion in policy discussions.

# References

World Bank. Tanzania mainland poverty assessment : Part 1 - path to poverty reduction and pro-poor growth. Technical report, 2020. URL https://openknowledge.worldbank.org/handle/10986/33542.

G. Vijay and P. Patnaik. Use of building footprint data in poverty mapping: A case study in india. *Applied Geography*, 96:1–12, 2018.

P. Garcia, A. Kanaan, and A. Tatem. Building footprints as a tool for poverty mapping in urban settings. *The Lancet Planetary Health*, 4(3):e135–e136, 2020.

F. Amendola, C. Rizzi, G. Gualtieri, and G. Gottardi. Poverty mapping using building footprints and satellite imagery: A case study in ghana. *Applied Geography*, 108:102217, 2019.

Thomas Pave Sohnesen, Peter Fisker, and David Malmgren-Hansen. Using satellite data to guide urban poverty reduction. *Review of Income and Wealth*, 2021. doi: https://doi.org/10.1111/roiw.12552.

Peter Fisker, Jordi Gallego-Ayala, David Malmgren Hansen, Thomas Pave Sohnesen, and Edmundo Murrugarra. Guiding social protection targeting through satellite data in são tomé and príncipe. Social Protection Jobs Discussion Paper 2212, World Bank, Washington, DC, 2022. URL https://openknowledge.worldbank.org/handle/10986/38222.

Tianqi Chen and Carlos Guestrin. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, aug 2016. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145%2F2939672.2939785.