Carnegie Mellon University
Tepper School of Business

Slum Mapping and Household Location Choice

Pranay Gundam

Tepper School of Business, Carnegie Mellon University

Under the guidance of Professor Laurence Ales

Introduction

The vague idea of a slum has been around since the conception of urbanized areas. There are various definitions that often differ from city to city but are typically along the lines of a dense cluster of informal housing units where living conditions are bad: these areas are usually associated with poverty, a lack of hygienic conditions, and a lack of land ownership upon which families take shelter.

At a higher macro level, Henderson (2002)[1] indicates that often times a mismatch in the rate of urbanization and investment capital available to a city leads to characteristics of a city that incentives slum growth. However, even though every income class of country on average experiences urban population growth rather than a decline, slum populations on average as a proportion of urban populations seem to be paradoxically decreasing. This project is in two parts:

1. Addressing the problem of little low cost reliable data on mapping slums
2. Exploring the discrepancy between slum growth and urban growth by creating a household choice model

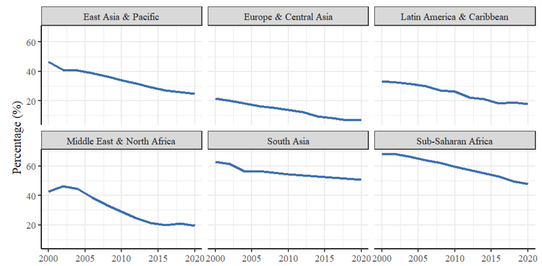

Figure 1:Average Percent of urban households living in slums has been on the decline in every global region

Relevant Data Sources

This project uses a considerable amount of data, there are five main sources listed below

- **Version 4 DMSP-OLS Nighttime Lights Time Series**: This paper uses nighttime light images from 2000 to 2013 stored in a TIF format - which are dense images which contain additional information at each pixel to a resolution of 1 sq km. In this case, each raw file download from NOAA contains an image of the world as below and and must be processed first using some spatial data processing software.

Figure 2:Averaged light values in 2008

- **World Pop Hub: Population Density Data**: A very encompassing database that includes a granular count of population density with a resolution of 1 kilometer.
- **Slum Dwellers International**: Training dataset that includes already identified slums, their coordinates, and additional characteristics of the slum such as access to sanitation or public transportation. This data does not exist in a consolidated format and must be scraped from SDI.org. This information was collected by consulting locals to survey areas from 2013 to 2018 in cities that are widely considered by locals to be slums.
- **World Bank: World Development Indicators**: This dataset is used to gague urban and slum populations.
- **Numbeo: Rent Characteristics of Countries**: This source collects information on rent in countries around the world through surveys and is used for the household location choice model.

Question 1: Slum Mapping

The primary purpose of the mapping models is twofold:

- Identify if there is some level of relationship between nighttime light values, population density, and the location of slums
- Create a model that is able to provide data on slum locations without the use of any intensive visual algorithms on hard to come by high resolution satellite imagery

I hypothesize that slums tend to be located in areas with high population density and low nighttime light levels. I develop four classification models that are fed a discretized value of both nighttime lights and population density.

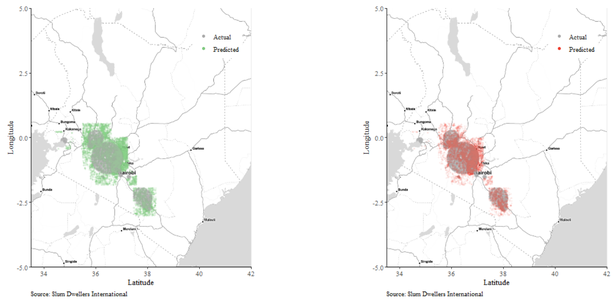

Figure 3:Weighted Logit and Random Forest Predicted Spatial Distribution of Slums in Kenya

Question 2: Household Location Choice

One way to model household behavior is a simplistic two choice model where households can choose whether or not they want to live in a slum or in an area in the city that is not a slum. To formalize, consider the problem setup; a household must choose whether to live in a slum S_{slum} or in the main non-slum city S_{city}. The household's rent r and wage w is determined by where they live, and we can define the utility of living in either location as

$$\begin{cases} U_{city} & = w_{city} - r_{city} + \epsilon_{city} \\ U_{slum} & = w_{slum} - r_{slum} + \epsilon_{slum} \end{cases} \tag{1}$$

For binary choice models, one commonly used tool is a logistic regression which would then define the probability of a household to live in either location as

$$\begin{cases} p(S_{slum}) & = \frac{\exp\left(u(S_{slum})\right)}{\exp\left(u(S_{slum})\right)+\exp\left(u(S_{city})\right)} \\ p(S_{city}) & = \frac{\exp\left(u(S_{city})\right)}{\exp\left(u(S_{slum})\right)+\exp\left(u(S_{city})\right)} \end{cases} \tag{2}$$

Since people living in slums often times still have jobs in other areas in the city, we can assume that $w_{slums} = w_{city}$ and then continue to conclude that

$$\ln\left(\frac{p(S_{slum})}{p(S_{city})}\right) = r_{city} - r_{slum}. \tag{3}$$

The calibrated binomial regression model is as follows

Table 1:Household Location Choice Calibration

	Dependent variable:
	Slum population as percent of urban population
Difference between slum and non-slum rent	0.0001473
	(0.0002)
Constant	−0.7524405**
	(0.320)
Note:	*p<0.1; **p<0.05; ***p<0.01

Key Takeaways

Slum Mapping

- Weighted logistic regression performs the best in terms of error metrics.
- Low light level values and high population density values *do* make the likelihood of a certain location being a slum higher.
- High light level values regardless of population density makes the likelihood of a certain location being a slum lower.

Table 2:Model validation error metrics

Model	Error Metric	Slum Predictions	Total Slum	Error Rate
Vanilla Logit	Total Dataset	2696	10604	0.4048472
(5 bins)	Positive Slum Observations	1525	4647	0.6718313
Bins Logit	Total Dataset	2759	10604	0.3904187
(10 bins)	Positive Slum Points	1651	4638	0.6474482
Weighted Logit	Total Dataset	9055	10604	0.3564527
(5 bins)	Positive Slum Points	2997	4657	0.3564526
Random Forest	Total Dataset	3399	10604	0.3564527
(continuous values)	Positive Slum Points	2085	4530	0.5397351

Household Location Choice

- Not enough data to come to a statistically significant statement.
- Conclusions are severely held back by the availiability of data.
- Model predicts that a widening of the gap increases the log-likelihood that a household will live in a slum.

Future Work

This project primarily suffers from a lack of clean and precise data. This is to be expected when working with developing countries as data is sparse and often times not as specific as desired. Future work after this project will primarily revolve around fixing these issues.

- The slum dataset is not as precise as can be. There are a lot of assumptions being made currently in the spatial organization of slums. Future work should get more precise spatial boundaries of slums to better match with the nighttime lights and population density data.
- This project was computationally restricted and stronger computers would be able to more accurately match the nighttime light and population density data. Currently, I sacrifice a bit of the granularity to speed up the processing process.
- The current project uses only 5 and 10 bins just to retain interpretability but future implementations of the mapping models can increase the bin counts to maximize predictive power.
- The household location choice model faces drastic issues of data availability. Future work can either increase the quality of data or delve into other computational methods like simulations or machine learning processes to create synthetic data.

References

[1] Vernon Henderson.
Urbanization in developing countries.
The World Bank Research Observer, 17(1):89--112, 2002.

Acknowledgements

In addition to my advisor I want to thank Slum Dwellers International and Anouska Siva assisting with my data collection process.