Slum Mapping

Pranay Gundam, Carnegie Mellon University

March 28, 2023

Abstract

Please provide an abstract of no more than 300 words. Your abstract should explain the main contributions of your article, and should not contain any material that is not included in the main text.

Contents

1 Slum Formation

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. How slums tend to emerge is an issue that is currently being studied and will be addressed throughout this paper as well.

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 overcrowding, excessive traffic, and significant health costs due to air and water pollution; characteristics of a city that not only incentives slum growth but also make slums more dangerous for the people living in them. This is a key issue for developing countries right now who are urbanizing at a rate much faster than developed countries.

India for example has seen rapid urbanization in recent years and 29.4% of its population lives in slums, while the percentage of slum dwellers reached 42% in Mumbai, the largest city and financial capital of the country [2]. Further, in Rio de Janeiro, although the city's total population grew only 3.4% over the last decade, the favela population has grown by 27.7% [3].The case of China is a clear example of the opposite side of the story where there are cities whose infrastructure expenditure has outpaced the rate of urbanization, and as a result, experience a relative lack of slums compared to the megacities in its developing country counterparts. Lui and Zhang [3] explain, however, that China's dual track urbanization of first government driven spatial expansion of cities and related infrastructure and second farmer driven urbanization through the construction of housing after their crop land is lost to city expansion keeps the supply of housing high enough to accommodate migrant workers who may have otherwise been living in slums.

Urbanization - which has been a consistent trend since the creation of cities - seems to trend according to the stage of development that nations are in. Despite this, the fact remains that 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, however, seem to be paradoxically decreasing. Although this an aggregate trend and may not be the case with every country, it still raises questions about the factors that influence slum growth. And although infrastructure spending seems to have on averaged experienced an upwards trend across regions relative to past decades, this trend doesn't seem to correlate well with the relative magnitudes of slum decline and some regions that don't experience any increase in infrastructure spending to match growing urban areas also witness a decline in slum proportions amongst urban populations.

The largest proportional population living in slums is in the Asia region, specifically South and South East Asia, which happen to be the regions that are urbanizing at the fastest rate. The conventional belief is that these urban areas overwhelm the ability of government agencies to coordinate on the city and national level and create sustainable low-income housing. Even beyond an inability to do so, however, is the adoption of the idea that economic growth will create

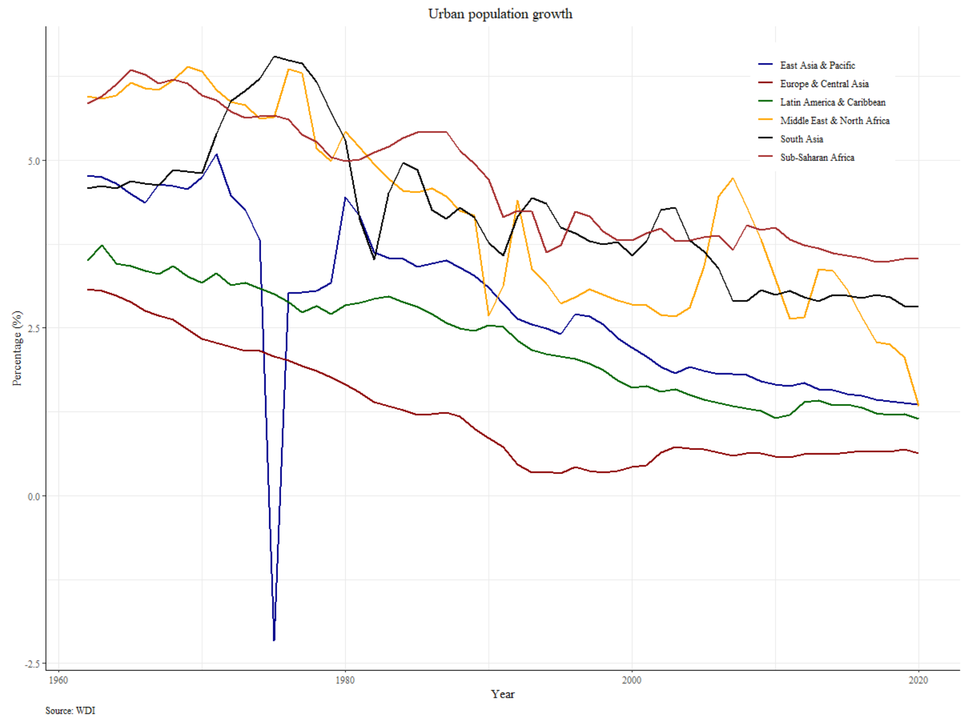

Figure 1: Relative to other classes of countries, urban population growth in low income countries has been increasing for the past 6 years

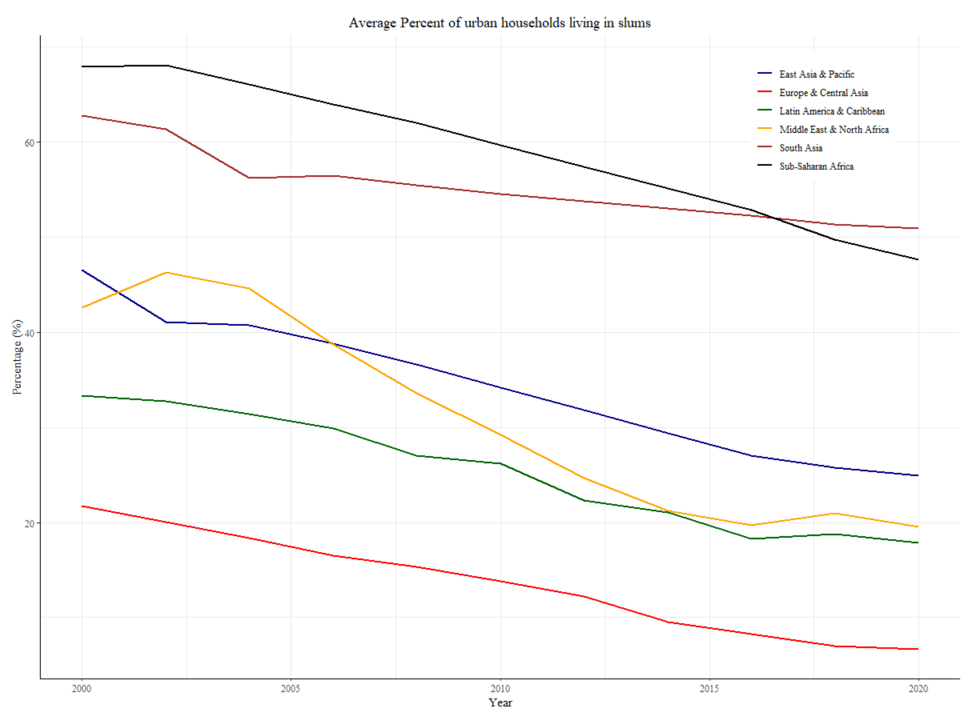

Figure 2: Caption

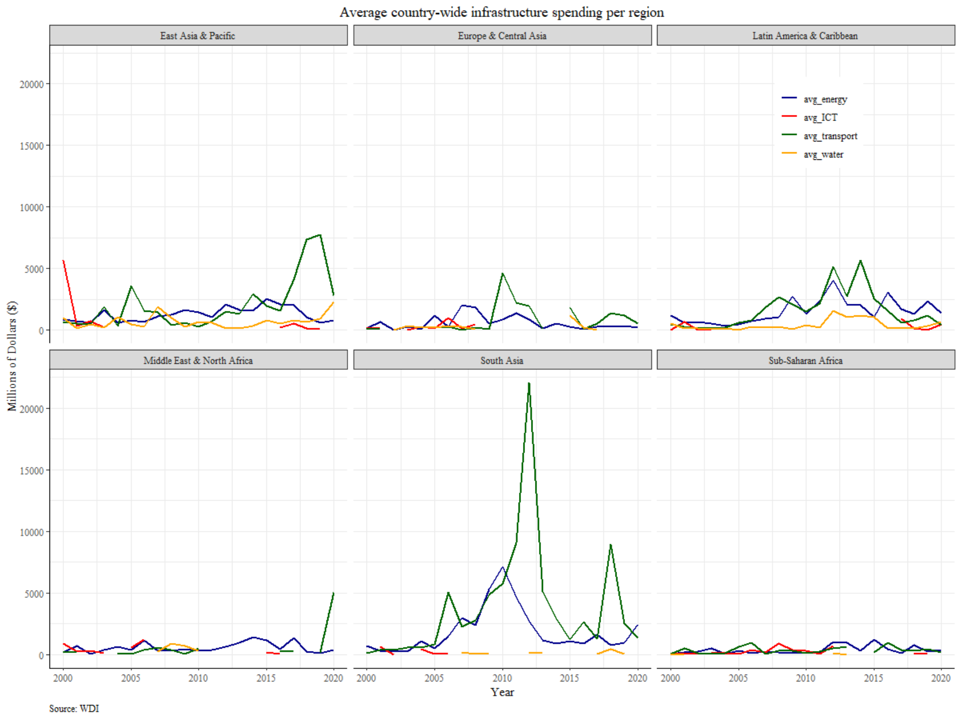

Figure 3: Caption

opportunities for and take care of the need for housing and health infrastructure. Despite this, the private housing industry in countries in Asia that are on the higher income end of the spectrum, such as Malaysia, often times focus on building homes for middle income markets [4]. Despite its prevalence across the world, slums all have varying characteristics from region to region. This project, however, is focused on Sub-Saharan Africa since that is where most of the training data has been collected from and is the region with the second highest average proportion of urban populations living in slums. Urban movement in Africa is often miss-classified as a mass exodus from Africa to Western Nations, in fact Africa has the lowest intercontinental out migration rates of all world regions [5]. And although intercontinental migration, especially to Europe, has been on the rise in Africa itself it, this has is primarily focused from North Africa and still pales relative to intracontinental migration within Africa itself which has also been increasing.

Table 1: Estimated total stocks of migration from, to, and within Africa [6]

	Africa to the rest of the world	Rest of the world to Africa	Within Africa
1960	1 830 776	2 811 930	6 176 385
1980	5 418 096	1 872 502	7 966 359
2000	8 734 478	1 532 746	10 500 000

These migration patterns have a significant effect on the urban landscape in Africa. As is generally the case around the world, migration mobility in Africa has been seen to be linked to the wealth of a given country and it's habitants. The wealthier an individual, the more facility

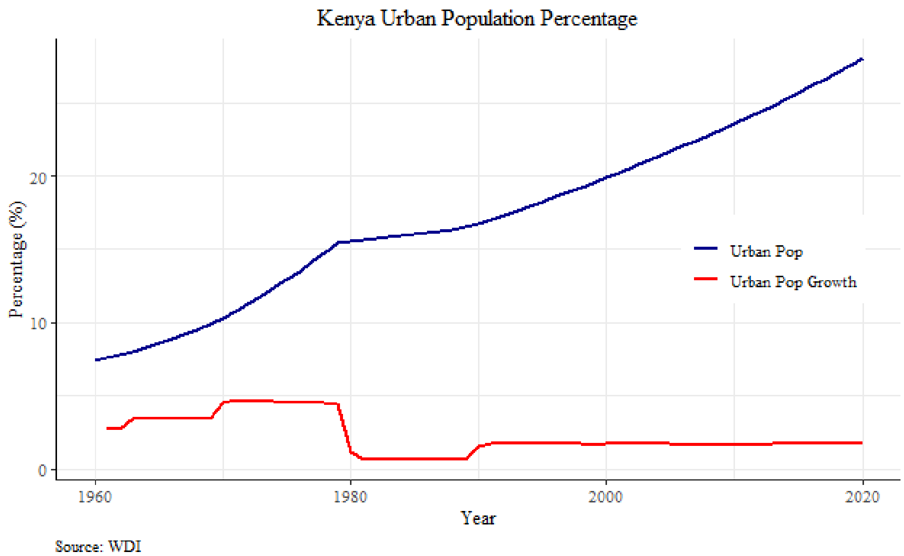

Figure 4: Urban population growth as a percentage of the total population of Kenya has stayed fairly consistent over the past 20 years.

they have to relocate and relocate to further distances as well. This is why a significant proportion of urban movement lies within rural relocation to urban areas in search of greater opportunity. Teye explains that "although a significant proportion of the rapid increase in urban population is caused by the high rate of natural increase in towns and re-classification of settlements into urban areas, migration accounts for a significant proportion urbanisation in Africa. In some countries, rural-urban migration accounts for about 60% of urban growth because rural-urban inequalities of development force people to move from rural areas to urban areas in search of jobs" [5]. The packing of poor groups of people into already overburdered urban areas due to the migratory movement of rural peoples into urban areas is what this paper studies.

A significant amount of this paper will focus on Kenya so it is important to take a cursory look at the migratory characteristics of Kenya as well. Figure 2 shows how urban population growth in Kenya has stayed fairly consistently positive over the past 30 years but that slum proportions have been on the decline since 2000. Infrastructure spending data in Kenya is sparse and is not enough to conclude some sort of trend.

2 Relevance

Although studying how slums comes to form is an interesting question in itself there are a few reasons as to why we should be concerned with the formation and characterization of them beyond trying to find an explanation for the decline of slums despite urban populations increasing and an unmatched spending on infrastructure. Even if one argues that some may prefer urban poverty to rural poverty there are several unique characteristics that slums pose that make them unique relative to the general case of lower income neighborhoods. Marx, Stoker, and Suri [7]

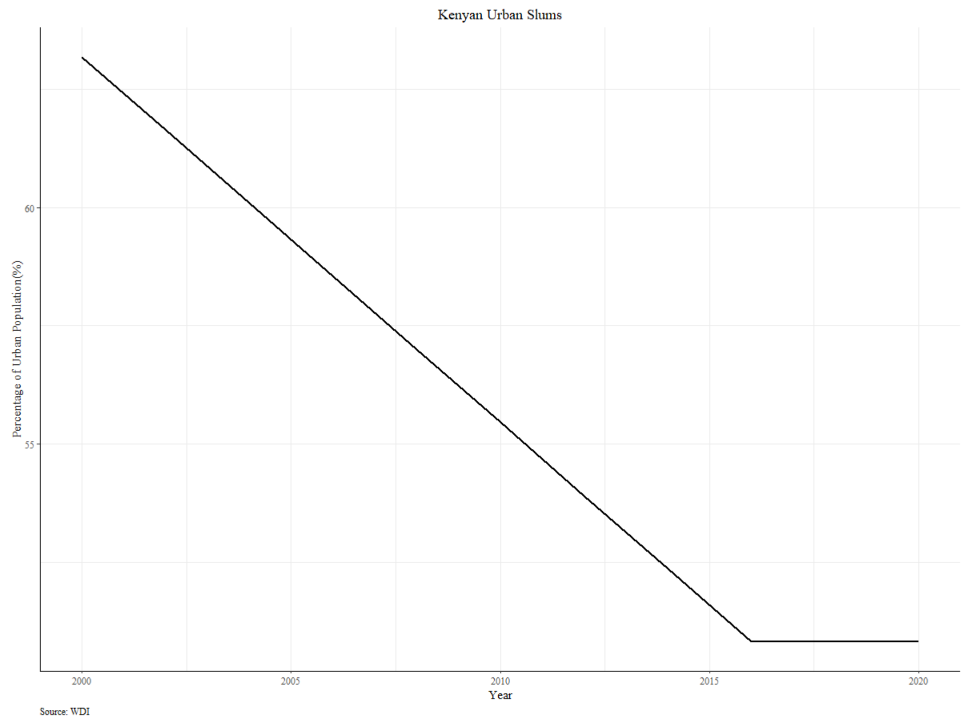

Figure 5: Despite urban population growth, the slum proportion of urban populations has been on the decline

explain that slums act as a poverty trap for the majority of their residents. They "document how human capital threshold effects, investment inertia, and a policy trap may prevent slum dwellers from seizing economic opportunities offered by geographic proximity to the city".

Slum mapping and prediction has been a relatively new area of interest. Some papers like Friesen, Rausch, Pelz, and Furnkranz [8] try to analyze national indicators that are most correlated with urban slum populations using data from the World Bank and modeling methods such as Random Forest, JRIP (rule learning algorithm), and J48 (decision tree learner). Their dataset, however, is not minute enough to make any meaningful claims. Using country-level effects to predict city-level indicators confounds many variables and can paint a wrong picture especially in countries in which the disparity between cities is large. On the other hand, new satellite imagery is allowing researchers to look at very small details that allow for a more micro level analysis. Image recognition software is being used to characterize photos with slums in order to have a more clear global pictures. Issues arise, however, in the accessibility and quantity of these high resolution satellite images.

Mapping and identification is critical to ensuring that potential future policy that target slums is accessible to all the areas that need assistance. Marx, Stoker, and Suri [7] for example recommend a paradigm for policies to address the growth of slums but in order to begin any sort of action one must have proper documentation as a foundation for any concrete plans. In addition, although slums may be the symptom of greater forces at play rather than the disease itself, learning more about how they form and why an individual chooses to live in a slum is crucial to understanding what of those greater forces need attention.

The following work will be in two parts. The first to explore new avenues for mapping slums, particularly using a novel data set, specifically the Defense Meteorological Satellite Program Operational Line Scanner (DMSP/OLS) Nighttime Light Image Timeseries from NOAA, and nuanced classification techniques. Since existing databases on slum identification is scarce, the second part of this paper uses the mapping model created in the first part to calibrate a discrete choice model that takes a micro economic glance at what factors drive households to live in slums. While we have a clue about the general correlation at a macro level of what motivates slum growth, identifying the systematic decision bias that is the threshold between choosing to live in a slum or not can help stimulate targeted policy action and give more insight as to what is driving down slums proportions despite an increase of urban populations.

3 Data Processing

The first portion of this paper uses a considerable amount of data and is primarily restricted by both computational power and availability of a database of slums. There are four primary categories that are listed below.

- **Version 4 DMSP-OLS Nighttime Lights Time Series**

  This dataset is the most novel out of all the other datasets and as such it also requires special treatment on the processing end as well. These files are stored in a TIF format which are dense images which contain additional information at each pixel. 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.

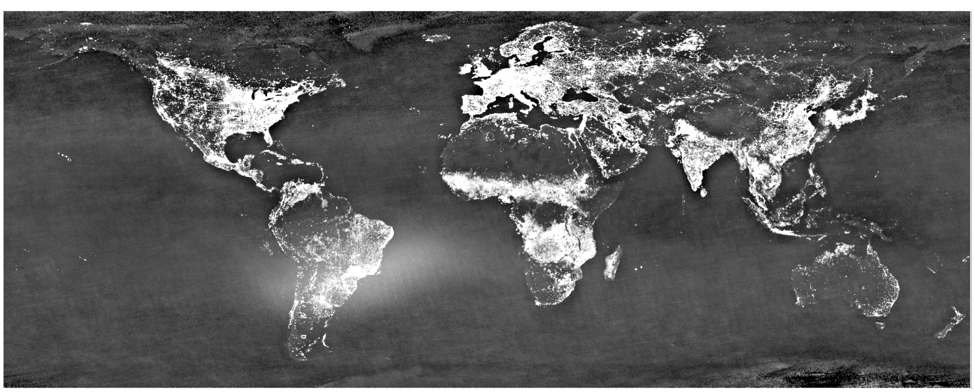

Figure 6: Averaged light values in 2008

  This paper uses nighttime light images from 2000 to 2013 and to reduce computational expenditure whilst also improving model relevance since most of the training dataset of identified slums is from African countries in 2013. Each pixel has a resolution of one square kilometer.

- **World Pop Hub: Population Density Data**

  The second dataset that this paper is contingent on is population density. WorldPop.org is a very encompassing database that includes a granular count of population density with a resolution of 1 kilometer.

- **Slum Dwellers International**

  This dataset acts as the 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.

Several decisions had to be made about the scope of the lights and population density data to consider since although data was plenty in these two areas, the slums dataset was very limited. Making assumptions that the slums dataset had profiled all the slums in a given country would drastically skew models as there might be many areas that are slums in real life and whose light and population density characteristics indicate to be so but are not captured in the slums dataset. As such, lights and population density observations were taken from grid areas around cities we know were surveyed for slums. In the future, given a more precise slums dataset, this paper can be much better defined.

Once all the data sets are individually processed we still must merge them all by coordinate points. This introduces another set of assumptions. The first issue is the matching of coordinate points between the nighttime lights and population density datasets, there are several methods of conducting this matching process that vary in time complexity and accuracy. After conducting a time complexity and accuracy tradeoff analysis this paper conducts the matching process by creating larger pixel grids, calculating the average population density of each larger grid space and then assigning that new aggregate population density value to each of the smaller pixels within that larger grid unit. In addition, to classify coordinate points as slums or not, we have to do a similar process of first making a simplifying assumption that slums are circular in nature and translate the square meter area into distance difference in geo-coordinates. This area to geo-coordinate translation also includes another simplifying assumption in that we disregard the curvature of the earth; this assumption, however, is fairly minute and often taken as a given when working with physical areas as small as the ones in the slums data set. Once the radius of each slum is identified, calculating whether or not a geo-coordinate pair in the lights data set is regarded to be a slum is a simple issue of calculating if distance to the center of any of the documented slums is less than that given slum's radius.

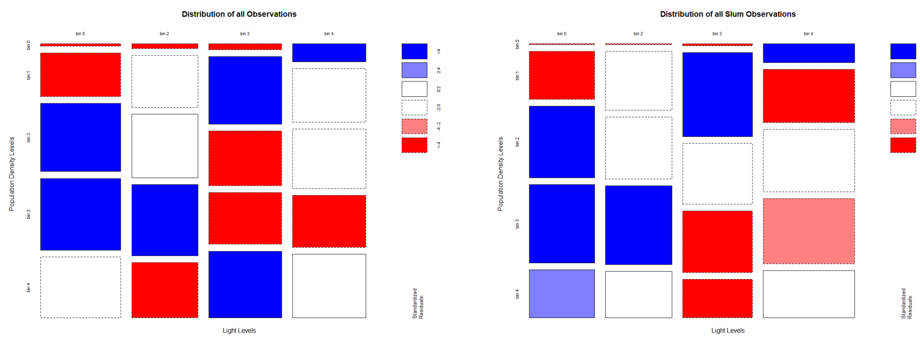

Figure 7: Distribution of lights and pop density values among slums is statistically biased

4 Exploratory Data Analysis

4.1 Compiled Dataset Summary Statistics

Each data set has a number of unique qualities and there is quite a bit of data to cover in order to have a thorough discussion of each nuance. Regardless there are still quite a few trends to consider before engaging in the model making. Note that unless otherwise specified, data here is from 2013, which is what we will use to create the slum mapping models as the training slums dataset is from 2013 as well.

Table 2: 2013 Kenya Cleaned Data Summary Statistics

	habitation	total_count	avg_light	avg_pop	sd_light	sd_pop
1	Not a Slum	28986	0.67	189.60	2.50	301.45
2	Slum	22381	2.39	137.63	8.03	430.09
3	Uninhabited	1650	8.89	0.00	16.68	

We see that our dataset is slightly imbalanced. The ratio of slum to no slum observations in the training data is approximately 1:1.295 which is relatively much better than proportions if we were to process data from the entire country which would imbalance the dataset to a ratio of about 1:16. Next, to discern the mixed effects of lights and population density, we engineer two new categorical variables by discretizing the range of both of these values into five quantiles; each of which is represented by a bin number, where the lower the bin number, the lower the quantile value.

Before we delve into any of the model building, Figure 7 is a quick sanity check to check homogeneity of these categorical variables amongst both positive slum observations and the entire dataset. We can see some promising initial evidence that the distribution of slums does have a statistical bias towards low light levels and high population density, although there is also a few clusters at high light levels and low population density. At a qualitative level, the marginal distributions of the continuous population density and light values conditioned on a location being a slum or not seem to be about similar. But the sheer magnitude of the sample size reveals a

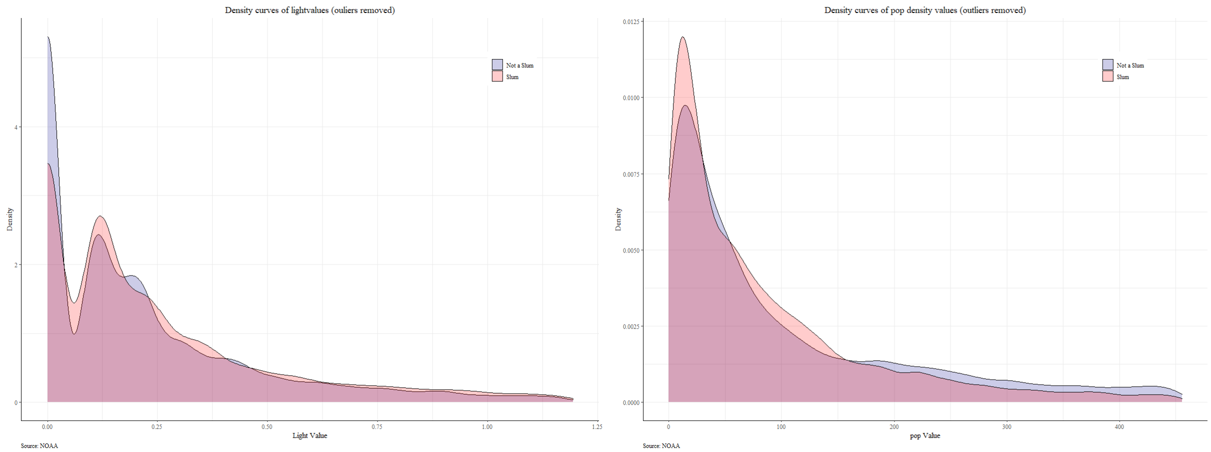

Figure 8: Caption

different story when looking at statistical tests and show that there are indeed statistically significant differences in the means of the distribution and the distribution as a whole as well at a 95% confidence level.

Table 3: Distribution Tests

	Test	p-value	df	t-score
1	Two-sample Kolmogorov Smirnov test	$< 2.2e - 16$		
2	Welch Two Sample T-test	$< 2.2e - 16$	25752	30.849

In the context of modeling, however, we discretize the continuous domain of light and population density values to capture the hypothesized non-linear affects of the interaction between population density and nighttime lights. For the basic model we break the light values into five quantiles and assign each observation and categorical light and population density a "bin". Figure 9 shows the distribution when we discretize the two features into ten bins whose distribution we can observe in figure 9: although population density seems to be approximately normally distributed with a high variance, the lights bins do not seem to be as well behaved as most light values tend to be very low and the bottom three quantiles are all the same.

The benefit of increasing the number of bins is that we can capture a bit more non-linearity the models than with lower bins, but increasing the number of bins also has an added affect of potentially over-fitting and reducing the interpretability of the model: it no longer becomes a question of "low" and "high" light values but more so model effects contingent on individual bins and little insight into trends in the behaviour of the data. It is important to note that the range of values, especially on the highest bin, may vary a lot due to outliers but cropping outliers can often times remove very key pieces of information since outliers in the context of this data often refers to city centers or very densely populated areas of a city which we have hypothesized may be home to slums.

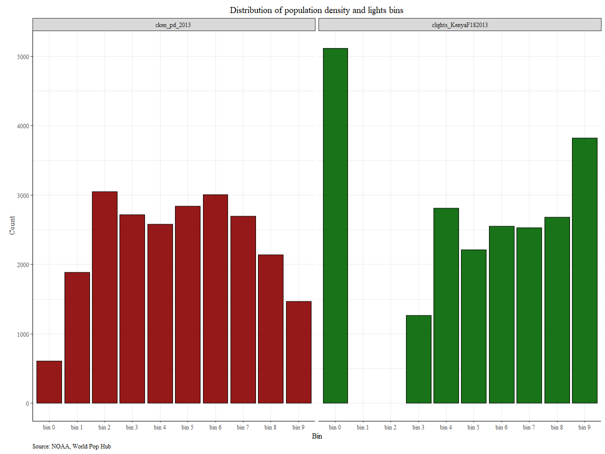

Figure 9: Caption

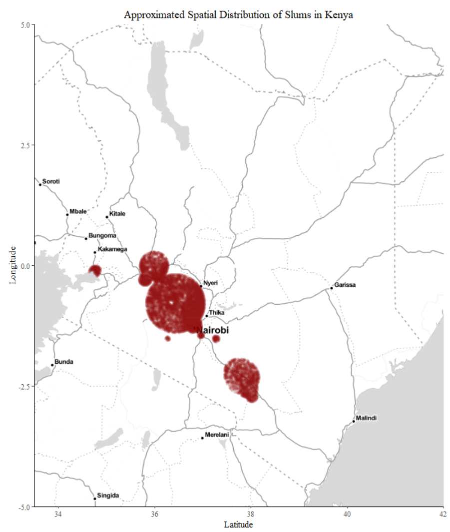

Figure 10: Caption

Finally, the data this paper works with is primarily spatial in nature: each observation is a latitude, longitude geo-coordinate pair. From an initial glance, we can clearly see assumptions about the shape of slums at play. Future iterations of the paper will embelish upon the slums dataset and define more precise slum boundaries to better capture potential light and population density trends.

4.2 Growth Accounting

It is important to note that in this paper, we treat light values and population density as a potential indicator of slums rather than there being some causal relationship. The main drivers of slum formation is likely the state of the economy and key characteristics such as the ratio between wages and rents which may in turn be connected to lights. We explore some of these macro relationships as well to gain more insight about any indirect effects. To begin, as the title of this

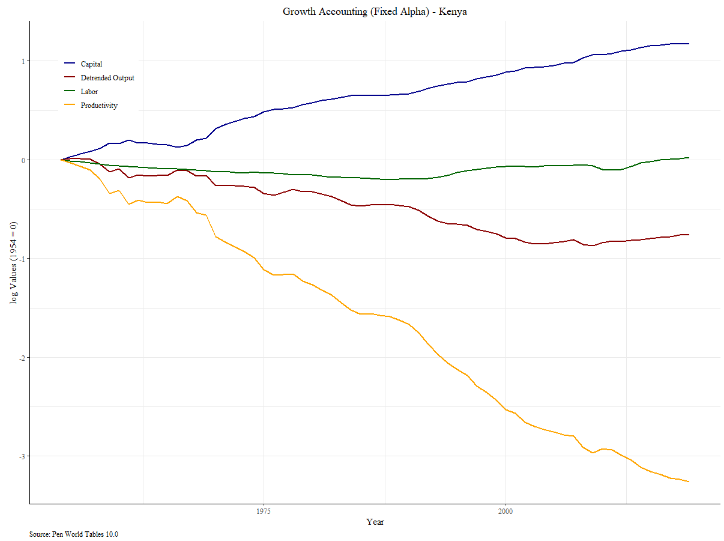

Figure 11

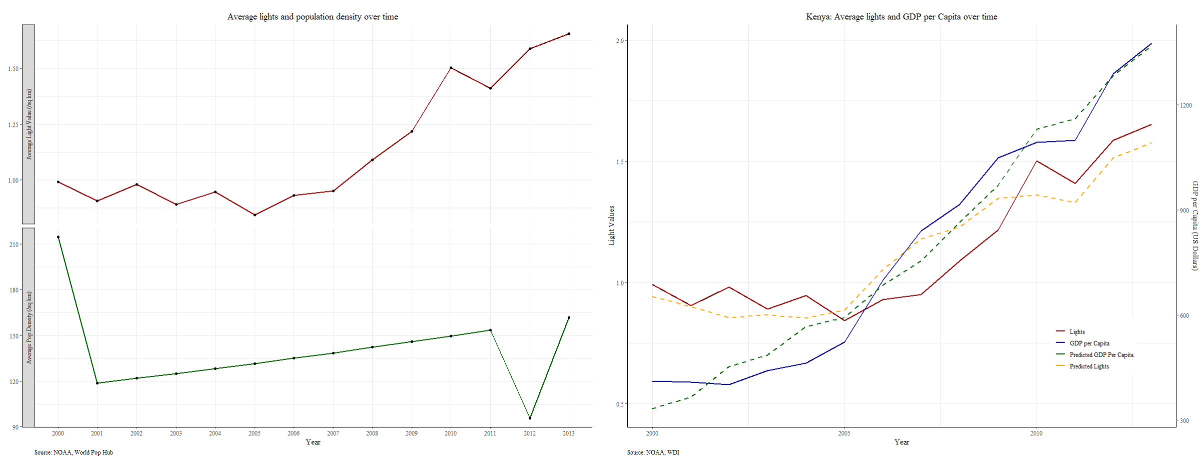

Figure 12

subsection entails, Figure 10 shows the solow macro model growth dynamics for Kenya. We see that the main reason why GDP growth is on the decline is a drop in the productivity term.

To further investigate, figure 11 shows both how aggregate average lights per square kilometer and population density per square kilometer in Kenya has fared from 2000 to 2013. As further insight into relationships between lights, population density and how a potential indirect relationship may be, the second graphic below displays the simple linear regressions of both lights and year on GDP and GDP and year on lights. In both cases, the impact of light on GDP and GDP on light have statistically significant positive coefficients which speaks to some potential confounders in terms of establishing a causal relationship between and population density and light. Since we are initially only trying to estimate some indicative relationship to use as a tool to map, this confounder in the causal relationship is good because it establishes at least some link between slums and lights.

Table 4: Correlation Plot (GDP is per capita)

	Year	GDP	Pop Density	Lights	Detrended GDP
Year	1.00	0.97	-0.12	0.85	0.98
GDP	0.97	1.00	-0.01	0.91	0.99
Pop Density	-0.12	-0.01	1.00	0.05	-0.02
Lights	0.85	0.91	0.05	1.00	0.85
Detrended GDP	0.98	0.99	-0.02	0.85	1.00

5 Modeling

5.1 Slum Mapping

The primary purpose of the mapping models is twofold; to identify if there is some level of relationship between nighttime light values, population density, and the location of slums; and second to create a model that is able to provide data on slum locations to use later in the calibration of the household decision model. Due to restrictions on available computational power, this paper uses data only from surveyed slums in Kenya: as such, there is no need for any country level control variables. One key factor to remember is that since the following models only train on data from Kenya, any model interpretation will reveal only characteristics about Kenyan slums or slums in countries and cities that are very similar to Kenya.

5.1.1 Model 1: Vanilla Logistic Regression

We begin with a bare-bones logistic regression model with specification as follows

$$p(x) = \frac{1}{1 + e^{-\beta^T x}}.$$

To discern the mixed effects of lights and population density, remember that we discretize the range of both of these values into five quantiles. We then interact these two new engineered categorical variables in the model. The initial logistic regression model is shown in table 5.

These results bode well for the initial hypothesis that slums tend to appear in areas with low light level value, but also seems to indicate that in general, low light level values are a good indicator of slums regardless of population density, and the biggest factor that contributes towards areas not being considered as slums is the highest bin of light values. Results for population density are a bit more inconclusive in terms of determining a general trend. Ultimately, the biggest takeaway is that low light level values drive the biggest increase in the log odds of a specific geo-coordinate observation being considered a slum and the highest light level values are the biggest drivers of decreasing the log odds of a specific geo-coordinate observation being considered a slum.

Although there seems to be promising statistical significance results it is important to assess the predictive ability of this mode as well. Logistic regression prediction often considers a

Table 5

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.0369	0.3544	-5.75	0.0000
clights_KenyaF182013bin 2	-0.7435	0.5083	-1.46	0.1435
clights_KenyaF182013bin 3	0.3378	0.4032	0.84	0.4021
clights_KenyaF182013bin 4	2.2856	0.3639	6.28	0.0000
cken_pd_2013bin 1	1.5698	0.3582	4.38	0.0000
cken_pd_2013bin 2	1.5032	0.3570	4.21	0.0000
cken_pd_2013bin 3	1.5363	0.3567	4.31	0.0000
cken_pd_2013bin 4	1.1077	0.3577	3.10	0.0020
clights_KenyaF182013bin 2:cken_pd_2013bin 1	1.2158	0.5135	2.37	0.0179
clights_KenyaF182013bin 3:cken_pd_2013bin 1	0.1776	0.4088	0.43	0.6640
clights_KenyaF182013bin 4:cken_pd_2013bin 1	-1.6974	0.3707	-4.58	0.0000
clights_KenyaF182013bin 2:cken_pd_2013bin 2	0.9302	0.5123	1.82	0.0694
clights_KenyaF182013bin 3:cken_pd_2013bin 2	0.0456	0.4082	0.11	0.9110
clights_KenyaF182013bin 4:cken_pd_2013bin 2	-1.4757	0.3693	-4.00	0.0001
clights_KenyaF182013bin 2:cken_pd_2013bin 3	1.1900	0.5118	2.33	0.0201
clights_KenyaF182013bin 3:cken_pd_2013bin 3	0.1197	0.4083	0.29	0.7694
clights_KenyaF182013bin 4:cken_pd_2013bin 3	-1.0234	0.3697	-2.77	0.0056
clights_KenyaF182013bin 2:cken_pd_2013bin 4	1.1204	0.5132	2.18	0.0290
clights_KenyaF182013bin 3:cken_pd_2013bin 4	-0.5722	0.4092	-1.40	0.1620
clights_KenyaF182013bin 4:cken_pd_2013bin 4	-1.7319	0.3698	-4.68	0.0000

prediction probability threshold of 0.5, this is, however, one hyperparmeter than can be tuned and optimized for the sake of arriving at a better predictive model. At a normal 0.5 prediction threshold.

Table 6: Logit Model 1 Validation error metrics

Error Metric	Slum Predictions	Slum Observations	Error Rate
Total Dataset	2696	10604	0.4048472
Positive Slum Observations	1525	4647	0.6718313

After conducting a gridsearch through potential threshold values we see that we can drastically increase slum predictive accuracy while only sacrificing a little accuracy in the context of the entire dataset. Finally, Figure 14 provides a spatial view of how the model seems to be performing using a prediction threshold of 0.4.

5.1.2 Model 2: Logit with Increased Bins

Having a dataset with only five bins creates a lot of potential for categorizing values that are in reality drastically different from each other into the as the same bin value. One way to address this issue is to increase the number of bins.

Again, the mosaic plots indicate a pretty similar story as with the data with just five bins. There is, however, additional nuance with the extreme high light and population density values.

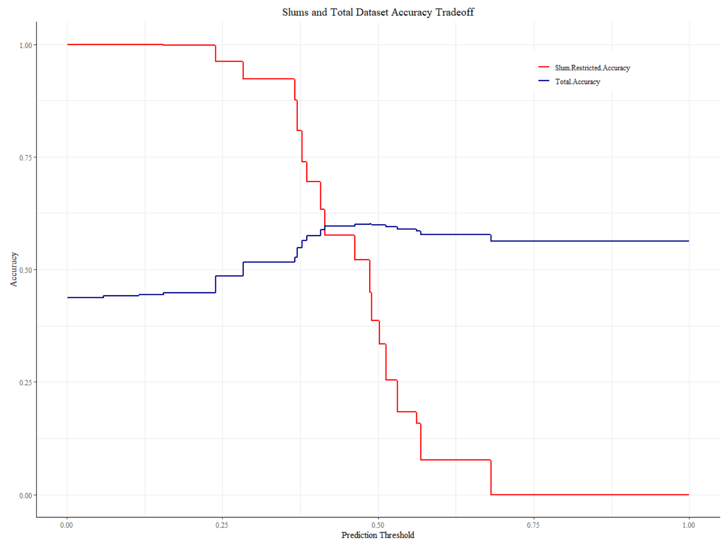

Figure 13: Caption

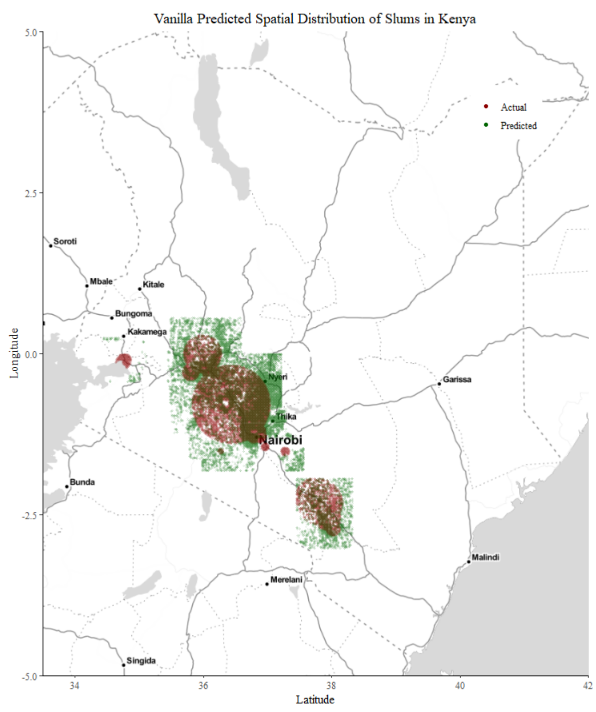

Figure 14: Caption

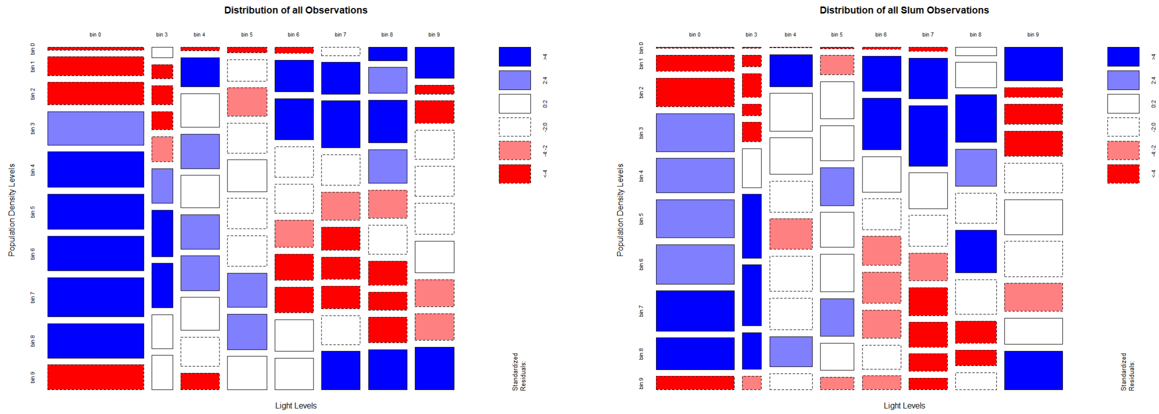

Figure 15: Caption

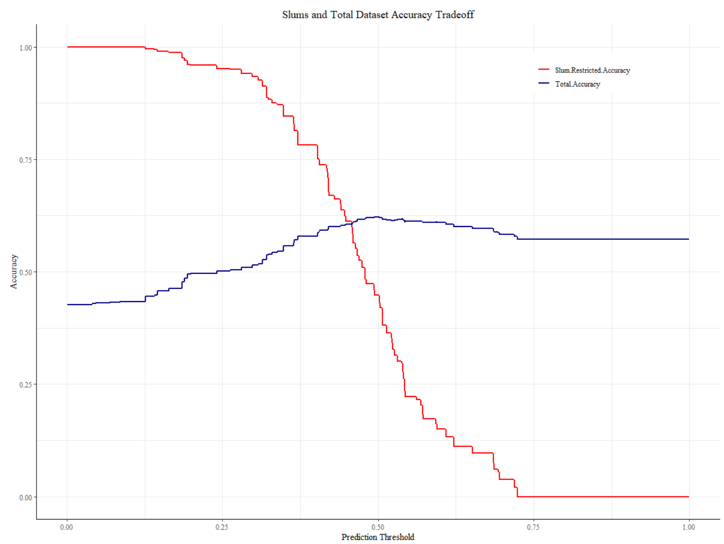

Figure 16: Caption

Once more, the regression results relate a very similar story as with the logistic regression model with only five bins. As a whole, the predictive capabilities of this model seem to be a bit worse to that of the normal logistic regression with 5 bins with respect to general error metrics at a predictive threshold of 0.5. Again after conducting a gridsearch through potential threshold values, however, we see that we can drastically increase slum predictive accuracy while only sacrificing a little accuracy in the context of the entire dataset a bit more than the previous regression model with only five bins. Finally, Figure 17 provides a spatial view of how the model seems to be performing using a prediction threshold of 0.4 and shows how it is able to mark one more slum area near Kakamega which the previous model was not able to do.

5.1.3 Model 3: Weighted Logit

As a whole, the weighted model seems to perform a lot better than the previous two models.

Title

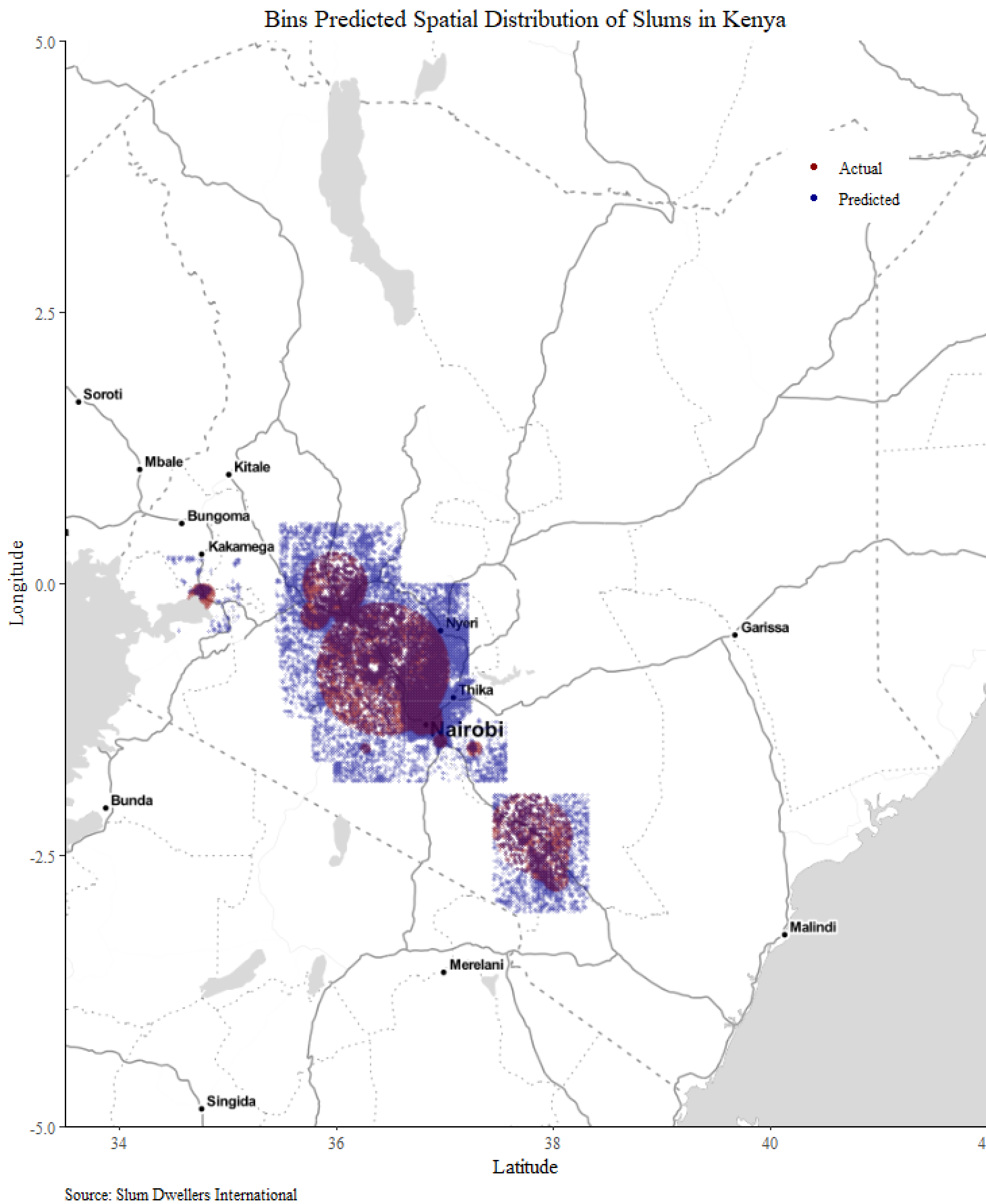

Figure 17: Caption

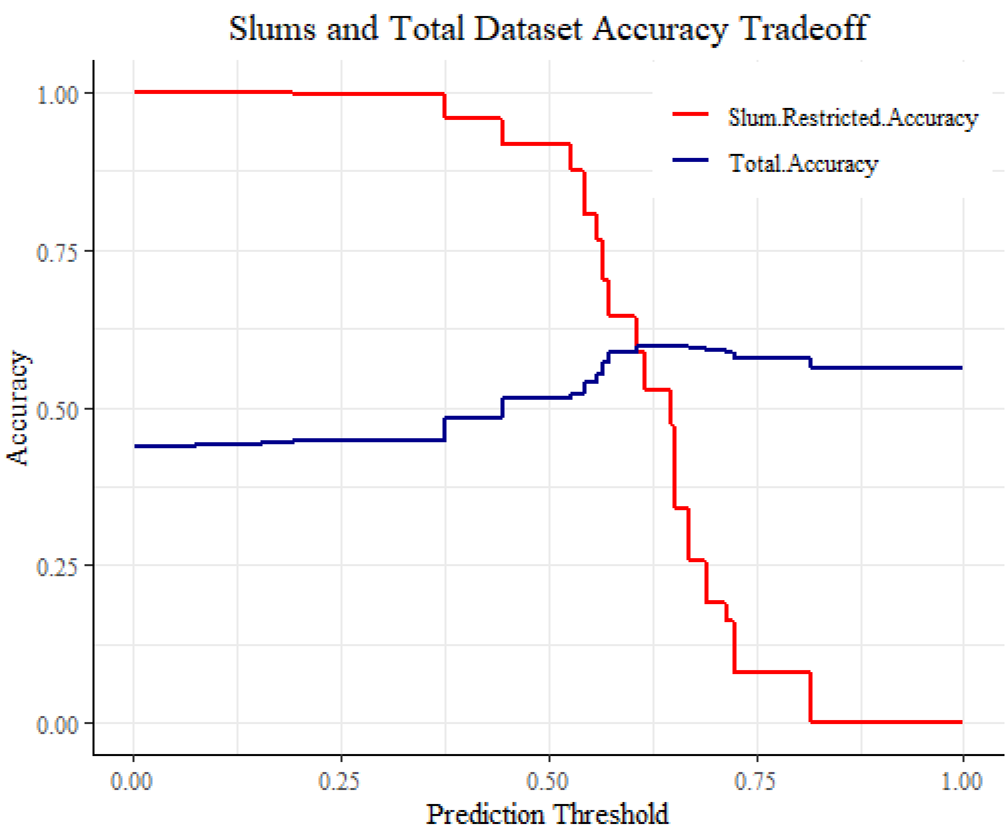

Figure 18: Weighted logistic regression threshold cuttoff graphs

Title

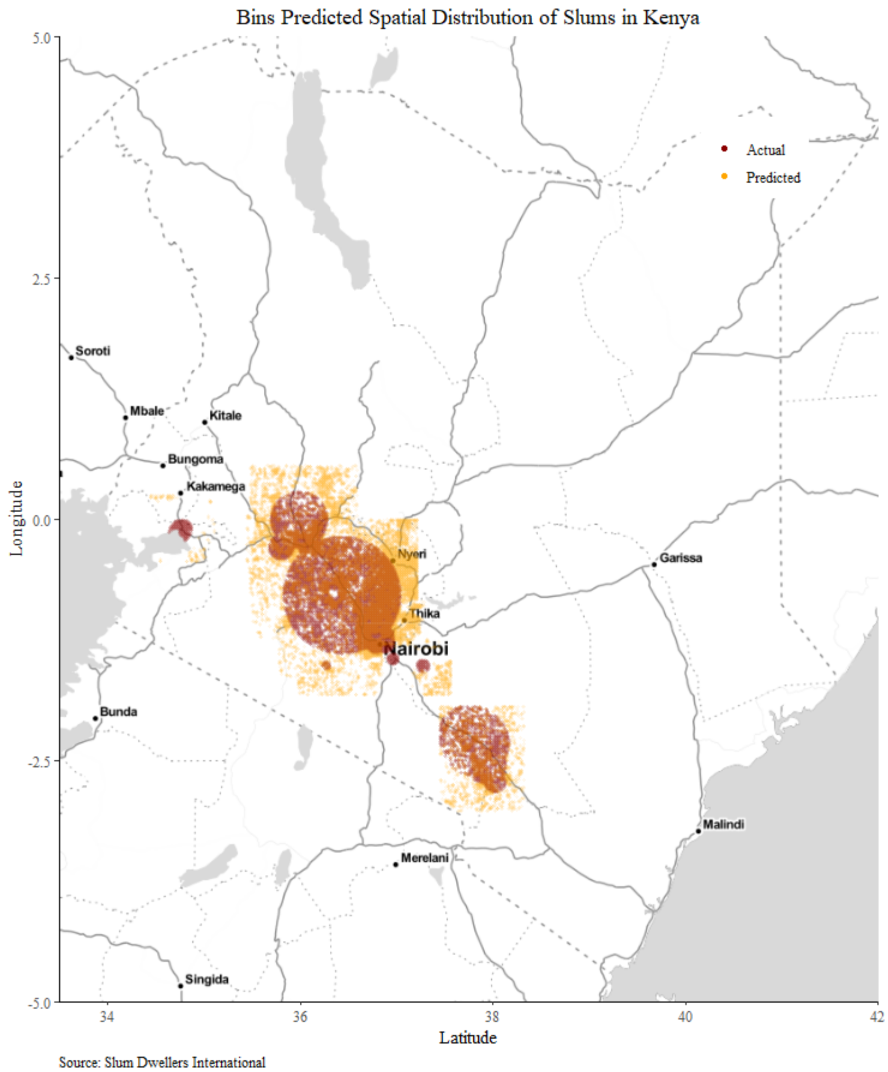

Figure 19: Caption

5.1.4 Model 4: Random Forest

5.2 Household Spatial Choice Models

As of yet, this paper has been a statistics question regarding how one can build a model to identify where slums are located based on nighttime light and population density data. The next parts of this paper will build and calibrate a decision model for households that face the choice between living in different areas. There are two ways to model the decision making process that households face.

5.2.1 Binary Choice Model

The first, more simplistic setup is a two choice model where households can choose whether or not they want to live in a slum or in the city that is not a slum. 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

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

We can then define the probability of a household to live in either location as

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

and further that the log of the ratios of the probabilities as

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

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 write

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

5.2.2 Multi-Choice Model

Consider the problem setup; a household H_j has the option to choose from a finite set of N locations $\{S_i\}_{i\in[N]}$. Each S_i for some $i \in [N]$ has rent $r_i > 0$ and exogenous characteristics z_i. We would then define the utility experienced by a household h making wage w and living in location

S_i as

$$u(h, S_i) = w - r_i + z_i. \tag{1}$$

We can define the probability of a household to live in location S_i as

$$p(h, S_i | S_1, \ldots, S_N) = \frac{\exp\left(u(S_i)\right)}{\sum_{j=1}^{N} \exp\left(u(S_j)\right)}. \tag{2}$$

Now, to further analyze the model, we shall assume that our set of locations $\{S_i\}_{i=1,\ldots,N}$ are around some modal city where there is a finite set of indices $I \subseteq [N]$ with cardinality M (this set need not be non-empty) $\{S_k\}_{k \in I} \subseteq \{S_i\}_{i \in [N]}$ where each S_k can be considered a slum. We denote a slum as some location S_k where the rent is less than some constant threshold $c \in \mathbb{R}^+$; this is a reasonable assumption since people who cannot afford higher rent are often pushed into living in slums, but note that there is an additional exogenous cost of living in a slum as well. We could then write the probability of a household living in a slum as

$$\sum_{k \in I} p(h, S_k | S_1, \ldots, S_N) = \frac{\sum_{k \in I} \exp(u(S_k))}{\sum_{i \in [N]} \exp(u(S_i))}. \tag{3}$$

6 Conclusion

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.3848	0.3950	-6.04	0.0000
clights_KenyaF182013bin 3	-0.7932	0.7093	-1.12	0.2634
clights_KenyaF182013bin 4	-0.6597	0.8244	-0.80	0.4236
clights_KenyaF182013bin 5	-0.2252	0.6089	-0.37	0.7114
clights_KenyaF182013bin 6	0.5765	0.5219	1.10	0.2693
clights_KenyaF182013bin 7	0.9861	0.4673	2.11	0.0348
clights_KenyaF182013bin 8	1.6464	0.4282	3.84	0.0001
clights_KenyaF182013bin 9	3.1593	0.4104	7.70	0.0000
cken_pd_2013bin 1	1.6017	0.4057	3.95	0.0001
cken_pd_2013bin 2	2.2234	0.4020	5.53	0.0000
cken_pd_2013bin 3	1.9907	0.4001	4.98	0.0000
cken_pd_2013bin 4	1.7567	0.4001	4.39	0.0000
cken_pd_2013bin 5	1.8526	0.3999	4.63	0.0000
cken_pd_2013bin 6	2.0590	0.3997	5.15	0.0000
cken_pd_2013bin 7	1.8271	0.3994	4.57	0.0000
cken_pd_2013bin 8	1.6319	0.4004	4.08	0.0000
cken_pd_2013bin 9	0.8984	0.4059	2.21	0.0269
clights_KenyaF182013bin 3:cken_pd_2013bin 1	1.1907	0.7490	1.59	0.1119
clights_KenyaF182013bin 4:cken_pd_2013bin 1	1.5381	0.8360	1.84	0.0658
clights_KenyaF182013bin 5:cken_pd_2013bin 1	0.1974	0.6297	0.31	0.7539
clights_KenyaF182013bin 6:cken_pd_2013bin 1	0.2161	0.5390	0.40	0.6885
clights_KenyaF182013bin 7:cken_pd_2013bin 1	-0.0405	0.4866	-0.08	0.9337
clights_KenyaF182013bin 8:cken_pd_2013bin 1	-0.9978	0.4533	-2.20	0.0277
clights_KenyaF182013bin 9:cken_pd_2013bin 1	-1.5621	0.4671	-3.34	0.0008
clights_KenyaF182013bin 3:cken_pd_2013bin 2	1.1037	0.7361	1.50	0.1338
clights_KenyaF182013bin 4:cken_pd_2013bin 2	0.8484	0.8333	1.02	0.3086
clights_KenyaF182013bin 5:cken_pd_2013bin 2	0.3645	0.6224	0.59	0.5582
clights_KenyaF182013bin 6:cken_pd_2013bin 2	-0.2586	0.5343	-0.48	0.6284
clights_KenyaF182013bin 7:cken_pd_2013bin 2	-0.6544	0.4806	-1.36	0.1733
clights_KenyaF182013bin 8:cken_pd_2013bin 2	-1.4583	0.4434	-3.29	0.0010
clights_KenyaF182013bin 9:cken_pd_2013bin 2	-2.6165	0.4348	-6.02	0.0000
clights_KenyaF182013bin 3:cken_pd_2013bin 3	0.1513	0.7404	0.20	0.8381
clights_KenyaF182013bin 4:cken_pd_2013bin 3	1.1094	0.8325	1.33	0.1826
clights_KenyaF182013bin 5:cken_pd_2013bin 3	0.2885	0.6213	0.46	0.6424
clights_KenyaF182013bin 6:cken_pd_2013bin 3	-0.0607	0.5362	-0.11	0.9099
clights_KenyaF182013bin 7:cken_pd_2013bin 3	-0.4192	0.4826	-0.87	0.3850
clights_KenyaF182013bin 8:cken_pd_2013bin 3	-1.2791	0.4431	-2.89	0.0039
clights_KenyaF182013bin 9:cken_pd_2013bin 3	-2.4714	0.4279	-5.78	0.0000
clights_KenyaF182013bin 3:cken_pd_2013bin 4	0.7456	0.7301	1.02	0.3072
clights_KenyaF182013bin 4:cken_pd_2013bin 4	1.1216	0.8332	1.35	0.1782
clights_KenyaF182013bin 5:cken_pd_2013bin 4	0.6146	0.6203	0.99	0.3218
clights_KenyaF182013bin 6:cken_pd_2013bin 4	-0.1678	0.5364	-0.31	0.7544
clights_KenyaF182013bin 7:cken_pd_2013bin 4	-0.5035	0.4833	-1.04	0.2975
clights_KenyaF182013bin 8:cken_pd_2013bin 4	-1.1064	0.4463	-2.48	0.0132
clights_KenyaF182013bin 9:cken_pd_2013bin 4	-2.0367	0.4293	-4.74	0.0000
clights_KenyaF182013bin 3:cken_pd_2013bin 5	1.1930	0.7236	1.65	0.0992
clights_KenyaF182013bin 4:cken_pd_2013bin 5	0.8571	0.8327	1.03	0.3033
clights_KenyaF182013bin 5:cken_pd_2013bin 5	0.5155	0.6202	0.83	0.4058
clights_KenyaF182013bin 6:cken_pd_2013bin 5	-0.1215	0.5368	-0.23	0.8209
clights_KenyaF182013bin 7:cken_pd_2013bin 5	-0.3626	0.4861	-0.75	0.4557
clights_KenyaF182013bin 8:cken_pd_2013bin 5	-0.2951	0.4479	-0.66	0.5100
clights_KenyaF182013bin 9:cken_pd_2013bin 5	-1.6643	0.4306	-3.86	0.0001
clights_KenyaF182013bin 3:cken_pd_2013bin 6	1.3958	0.7208	1.94	0.0528

Table 7: Logit Model 2 Validation error metrics

Error Metric	Slum Predictions	Slum Observations	Error Rate
Total Dataset	2759	10604	0.3904187
Positive Slum Observations	1651	4638	0.6474482

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.7047	0.3138	-5.43	0.0000
clights_KenyaF182013bin 2	-0.8210	0.4543	-1.81	0.0707
clights_KenyaF182013bin 3	0.2697	0.3598	0.75	0.4535
clights_KenyaF182013bin 4	2.6155	0.3225	8.11	0.0000
cken_pd_2013bin 1	1.9388	0.3169	6.12	0.0000
cken_pd_2013bin 2	1.8749	0.3158	5.94	0.0000
cken_pd_2013bin 3	1.9673	0.3156	6.23	0.0000
cken_pd_2013bin 4	1.4771	0.3161	4.67	0.0000
clights_KenyaF182013bin 2:cken_pd_2013bin 1	1.1909	0.4585	2.60	0.0094
clights_KenyaF182013bin 3:cken_pd_2013bin 1	0.1996	0.3644	0.55	0.5838
clights_KenyaF182013bin 4:cken_pd_2013bin 1	-2.0506	0.3283	-6.25	0.0000
clights_KenyaF182013bin 2:cken_pd_2013bin 2	1.0786	0.4573	2.36	0.0183
clights_KenyaF182013bin 3:cken_pd_2013bin 2	0.0345	0.3638	0.09	0.9245
clights_KenyaF182013bin 4:cken_pd_2013bin 2	-1.8261	0.3270	-5.58	0.0000
clights_KenyaF182013bin 2:cken_pd_2013bin 3	1.1809	0.4571	2.58	0.0098
clights_KenyaF182013bin 3:cken_pd_2013bin 3	0.0917	0.3638	0.25	0.8010
clights_KenyaF182013bin 4:cken_pd_2013bin 3	-1.3927	0.3278	-4.25	0.0000
clights_KenyaF182013bin 2:cken_pd_2013bin 4	1.1560	0.4578	2.53	0.0116
clights_KenyaF182013bin 3:cken_pd_2013bin 4	-0.5558	0.3640	-1.53	0.1267
clights_KenyaF182013bin 4:cken_pd_2013bin 4	-2.0961	0.3269	-6.41	0.0000

Table 8: Logit Model 3 Validation error metrics

Error Metric	Slum Predictions	Slum Observations	Error Rate
Total Dataset	9055	5723	0.3564527
Positive Slum Observations	2997	4657	0.4136175

References

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

[2] Mumbai Transformation Support Unit (MTSU). Study on appropriate institutional models for slum free mumbai & urban renewal. *MTSU*, 2015.

[3] Shouying Liu. Yue Zhang. Cities without slums? china's land regime and dual-track urbanization. *Elsevier*, 2020.

[4] A. Kofi Amegah. Slum decay in sub-saharan africa. *J Urban Health*, 84:27–34, 2007.

[5] Joseph Teye. Urbanization and migration in africa. *United Nations Expert Group Meeting For The Review And Appraisal Of The Programme Of Action Of The International Conference On Population And Development And Its Contribution To The Follow-Up And Review Of The 2030 Agenda For Sustainable Development*, 2018.

[6] The World Bank. Global bilateral migration, 2011. data retrieved from Global Bilateral Migration Database, https://databank.worldbank.org/source/global-bilateral-migration.

[7] Marx. Stoker. Suri. The economics of slums in the developing world. *Journal of Economics Perspectives*, 27(4):187–210, 2013.

[8] Friesen. Rausch. Pelz. Furnkranz. Determining factors for slum growth with predictive data mining methods. *Urban Science*, 2018.