A Machine Learning Approach for The Identification of Informal Settlements by Remote Sensing Data and Socio-Economic Linkages, Evidence From 68 Cities Around the Globe

by

Zhiang Chen

August 2023

A Dissertation submitted in part fulfilment of the

Degree of Master of Science in Urban Spatial Science

Dissertation Supervised by Dr. Ollie Ballinger

The Centre for Advanced Spatial Analysis

Bartlett Faculty of the Built Environment

University College London

CONTENTS

Abstract **2**

Declaration **3**

Acknowledgement **4**

List of Figures **5**

List of Tables **6**

List of Acronyms and Abbreviations **7**

1 Introduction **8**

2 Literature Review **10**
1 Informal Settlement Identification In Urban Area..... 10
2 Architectural and Social Indicators Linkages and Predicting Potential 11

3 Data **12**
1 DHS Survey Data..... 12
2 Open Buildings 15

4 Methodology **18**
1 Informal Settlements Detection 18
2 Building Indicators..... 21
3 Socio-Economic Linkages 24

5 Results **26**
1 Economics..... 27
2 Education..... 29
3 Living Conditions..... 31

6 Discussion **36**

7 Conclusion **39**

References **42**

Appendix **43**

ABSTRACT

In this research endeavor, we investigate the building data of 68 of the most populous cities across 27 developing and least developed countries (LCDs) globally provided by Open Buildings extracted by deep learning algorithms from high-resolution satellite imagery. Leveraging metrics of building area indicators obtained from building footprint and Gaussian kernel-processed small building density scores (SBDS), we discern informal settlements and slums within the urban areas. Moreover, we use DHS data to establish a comprehensive framework encompassing multiple linear regression, logistic regression, and decision tree models. This framework spans the spectrum of individual and survey cluster levels, allowing for an exploration of the interplay between architectural indicators and socioeconomic data from the vantage points of economics, education, and residential conditions. We derive quantitatively substantiated conclusions through the exploration and introduce an innovative alternative approach for potential mass socioeconomic data collection with the fusion of remote sensing data and machine learning techniques. This study contributes not only to a quantitative comprehension of the complex relationship between architectural dynamics and societal economic aspects but also outlines a novel pathway for the acquisition of macrosocioeconomic data.

Keywords: Informal settlement detection, Urban building footprints, Developing countries, Least developed countries, Multiple linear regression, Logistic regression, Decision tree, Machine learning, Gaussian kernel, Statistical modeling, Socioeconomic data, Economics, Education, Residential conditions, Open Buildings data, Remote sensing data, DHS data, Quantitative analysis, GIS

DECLARATION

I, Zhiang Chen, hereby declare that this dissertation is all my own original work and that all sources have been acknowledged. It is **10136** words in length.

Zhiang Chen, 25th of August 2023

ACKNOWLEDGEMENTS

First and foremost, I would like to extend my heartfelt gratitude to my supervisor, Dr. Ollie Ballinger, for his invaluable guidance throughout the course of this paper. His generous support has been of paramount importance to me. I would also like to express my appreciation to all the faculty members of CASA for their enthusiastic teaching over the past year, from which I have derived immense benefit. Lastly, I am deeply thankful for the unwavering support of my partner Ruiwen Zhang, family, and friends. Their presence and encouragement have been a constant source of strength and comfort on this journey.

LIST OF FIGURES

3.1 Maps of 68 Cities Around the World 12
3.2 Distribution of Wealth Index in Different Countries 13
3.3 R300 Freeway between Philippi and Mandalay, Johannesburg, South Africa 16
3.4 500m Grid Map of Makoko, Lagos, Nigeria 16

4.1 3D Plot for Gaussian Kernel Weight 18
4.2 Satellite Image of Dharavi, Mumbai, India 20
4.3 Open Buildings Polygon of Dharavi, Mumbai, India 20
4.4 SBDS Layer of Dharavi, Mumbai, India 20
4.5 Distribution of Building Area Indicators at Individual Level 21
4.6 Distribution of Building Area Indicators at Individual Level after Log Transformation 21
4.7 Distribution of SBDS Indicators at Individual Level 22
4.8 Distribution of SBDS Median at Individual Level 22
4.9 Distribution of Building Area Indicators at the Cluster Level 23
4.10 Distribution of SBDS Indicators at the Cluster Level 23
4.11 Correlation Matrix for Building Indicators 24

5.1 Tree Model Output at Individual Level Using Building Area Indicators 30

6.1 Regression Plots for Model 15 37

7.1 Flow Chart for Acquiring Socio-Economic Data Using Remote Sensing Data 40

LIST OF TABLES

3.1 Comparison of EIP and Wealth Index in Some Countries 14

5.1 Table of All 17 Models in This Study 27
5.2 Estimation of EIP by Building Indicators 27
5.3 Estimation of the Wealth Index by Building Indicators in Different Countries 29
5.4 Estimation of Years of Education by Building Indicators..... 31
5.5 Estimation of Building Indicators by Living Conditions at Individual Level 32
5.6 Estimation of Building Indicators by Living Conditions at Cluster Level..... 33
5.7 Estimation of Water Sources by Building Indicators 34
5.8 Estimation of Toilet Facility by Building Indicators 34
5.9 Estimation of Roof Material by Building Indicators 35
5.10 Estimation of Electricity and Transportation by Building Indicators..... 35

LIST OF ACRONYMS AND ABBREVIATIONS

LDCs Least developed countries

GDP Gross domestic product

DHS Demographic and Health Surveys

UNICEF United Nations International Children's Emergency Fund

VIF Variance Inflation Factor

SD Standard Deviation

BA Building Area

SBDS Small Building Density Scores

EIP Estimated Individual Product

POI Point of Interest

GIS Geographic Information System

1. INTRODUCTION

Urbanization stands as a pivotal keyword characterizing the development of this century. According to the 2022 World Cities Report by the United Nations Human Settlements Programme (UN-Habitat, 2022), rapid urbanization has led to half of the population residing in urban areas around the globe. This trend is projected to escalate, with over two-thirds of the world's inhabitants expected to live in cities by 2050. Consequently, cities are poised to become the epicenter of human activity today and in the foreseeable future. Contrasting rural areas, where agriculture predominantly drives economic activities, cities emerge as the luminous gems of modern industrial society. Within these urban confines, individuals, often strangers to one another, can engage in specialized roles within a comprehensive industrial chain. This heightened division of labor not only bolsters production efficiency but also generates greater value, leading to higher wages and increased employment opportunities. Such prospects have invariably drawn a significant influx of population migration to urban areas.

In the midst of such accelerated urbanization, an inevitable consequence has been the emergence of slums of varying scales in the majority of large cities within developing nations. This rapid influx of population, coupled with a lack of infrastructure and improper management, has given rise to these areas characterized by high population density, unauthorized and non-compliant constructions, and a dearth of essential amenities such as waste management or potable water infrastructure. Concurrently, the absence of order in these regions often leads to heightened instances of violence, drug dealing, and other criminal activities (Bird et al., n.d.). The delineation of boundaries for these informal urban settlements often remains ambiguous, with administrative boundaries being overly broad, potentially encompassing parts of formal construction zones. Thus, the first research question of this study is: Can building footprints be utilized to accurately identify and demarcate the boundaries of informal residential areas or slums within the major cities of developing countries?

Furthermore, architectural differences across different communities often serve as a reflection of their inherent socio-economic and cultural fabric. Distinct communities, whether demarcated by socio-economic status or other factors, invariably manifest discernible differences in their architectural forms. For instance, the architectural footprints of low-density residential areas, characterized by sprawling layouts and expansive spaces, stand in stark contrast to the compact and vertically oriented structures in high-density zones. These variations in residential zones, beyond their physical attributes, are indicative of underlying socio-economic disparities. Such disparities might manifest in the diverse living conditions, economic opportunities, and educational backgrounds of the residents. A deeper examination of the building footprints within a region can provide invaluable insights into its socio-economic conditions and the prevailing quality of life. Consequently, the second research question of this study becomes even more pertinent: Is there a tangible correlation between architectural indicators and socio-economic data? If such a correlation exists, what might be the nature and magnitude of its effect on social economic well-being?

Nowadays, the collection of socio-economic data is often a complex and costly endeavor, primarily facilitated through surveys and interviews conducted via various channels (Bakibinga

et al., 2019). This methodology implies significant financial outlays, a high technical threshold, and extended timeframes for data acquisition. In many developing and least developed nations, where poverty remains pervasive and limited public resources are predominantly allocated to essential needs such as basic healthcare or food provision, there is an acute scarcity of resources dedicated to comprehensive socio-economic surveys. This results in substantial data gaps, posing challenges for numerous research endeavors. Consequently, the third research question of this study emerges: If a correlation exists between socio-economic data and architectural indicators, can we solely rely on remote sensing data, processed through statistical and machine learning models, to derive socio-economic metrics? Such an approach could circumvent the high costs associated with traditional surveys, offering a more cost-effective and unrestricted method for socio-economic data collection, ensuring relative accuracy in the process.

In this study, we embarked in-depth exploration of methodologies that leverage architectural footprints as a tool for discerning informal settlement within urban environments. Drawing from the repository of building footprint data, we meticulously devised a convolution technique using Gaussian kernels based on the previous study to establish the Small Building Density Scores, aiming to accurately pinpoint informal residential sectors within cities. Subsequently, we constructed to investigate the correlation between socioeconomic metrics and architectural indicators, focusing on economics, education, living conditions, culminating in a quantified analysis of these relationships. In the study's denouement, we introduced a pioneering methodology that exclusively employs remote sensing data, synergized with advanced machine learning algorithms, to forecast socio-economic indices. This discourse was enriched with a contemplative evaluation of the inherent merits and potential pitfalls of our approach, supplemented by a discussion on prospective refinements to enhance its efficacy.

2. LITERATURE REVIEW

1. INFORMAL SETTLEMENT IDENTIFICATION IN URBAN AREA

Slums, as areas with the highest concentration of exploited populations in modern cities, especially in developing countries, are garnering increasing attention. One of the primary challenges scholars face is distinguishing these area from the broader urban landscape. Various methodologies have been considered for this purpose, primarily bifurcating into two main approaches: one based on social surveys utilizing questionnaires and interviews to delineate slums, and the other leveraging remote sensing data combined with statistical methods to demarcate these areas.

Research on slums is inherently human-centric. Our focus extends beyond the mere observation of narrow, grimy streets or dilapidated temporary shelters; it is centered on the individuals residing within these confines. Engaging in direct communication with the inhabitants facilitates a swift determination of which urban sectors should be classified as slums. In his research, Pauline introduced a systematic sampling survey method specifically tailored for slum environments (Bakibinga et al., 2019). By establishing a robust sampling framework, he aimed to procure accurate data pertaining to these marginalized areas.

While face-to-face social surveys remain invaluable, they are often costly and challenging to execute. Recent years, we have witnessed a growing emphasis on leveraging machine learning techniques to analyze remote sensing data and employ Geographic Information Systems (GIS) methods for distinguishing informal settlements. In his research, Juan proposed a method that extracts features from high spatial resolution images and employs machine learning algorithms to discern whether urban areas qualify as slums (Duque et al., 2017). Michael, in his study, utilized Sentinel-2A data for training and applied the Random Forest algorithm to classify the outputs from kernels, achieving a high accuracy rate in identifying extensive slum areas (Wurm et al., 2017). Prima, on the other hand, harnessed WorldView-2 satellite data to analyze urban living patterns, road conditions, textures, vegetation, and proximity to rivers on Java Island, Indonesia, aiming to identify slum regions (Widayani, n.d.). These studies have directly employed satellite imagery for model training, culminating in the successful identification of slum areas.

Beyond the direct use of satellite imagery, another innovative method involves the utilization of building footprint data, wherein conclusions are drawn from an analysis of these footprints. In his research, Robert successfully automated the categorization of building types within footprint data through machine learning techniques (Hecht et al., 2015). Concurrently, in the Informal Settlement Mapper project, Ollie constructed a convolutional layer based on building footprints. By calculating the density of small-area structures, he achieved precise identification of informal residential areas in Dar es Salaam (Ballinger, n.d.), which significantly informed and inspired our study. We amalgamated both approaches in our study by utilized building boundaries outputted from high-resolution satellite imagery, and employed the kernel method based on building footprint data. After meticulous adjustments and testing, we achieved accurate identification of slum areas. Compared to previous studies, our research boasts a broader scope, encompassing

68 cities worldwide, thereby offering a paradigm in developing countries on global scale.

2. ARCHITECTURAL AND SOCIAL INDICATORS LINKAGES AND PREDICTING POTENTIAL

Several studies have illuminated the potential correlation between building area and socio-economic indicators. In his research, Gebhard employed a Random Forest classifier to associate household income, assets, and educational levels with building types, based on interview data from cities in Honduras (Warth et al., 2020). While this approach can establish more precise connections, the model's methodology predominantly relies on an ensemble of decision trees, where the majority rules. Consequently, even though the model boasts superior performance, its interpretability is somewhat limited, hindering the direct extraction of quantifiable conclusions. Chen, in her research, employed a multivariate linear regression model to investigate the relationship between urban vitality and socio-economic indicators (Jia et al., 2021), including economic metrics and Points of Interest (POI). While, Luca utilized a spatial regression model in his study, illustrating that urban complexity augments with increases in population and per capita income (Salvati and Carlucci, n.d.). Christos, in his research, harnessed street view images processed through machine learning algorithms (Diou et al., 2018). He then established a regression model using the model's output to estimate socio-economic status data within specific regions.

These models collectively underscore a discernible correlation between architecture and socio-economic activities. While their performance might not rival the accuracy of decision tree models or neural network models, the results they yield are more interpretable, and valuable insights can be gleaned. Compared to some models that employ econometric methods to determine causal relationships, even though these models might not conclusively ascertain the direction of the effect, the mere existence of such correlations holds intrinsic significance. It offers the potential for a reasonably accurate quantitative estimation of the socio-economic conditions within a region, given an understanding of its architectural landscape.

Furthermore, obtaining socio-economic data through remote sensing has long been an aspiration in social surveys. With the rapid iteration and enhancement of statistical models, improved model performance and reliability, coupled with an increasing abundance of rich and cost-effective remote sensing data, this aspiration is progressively nearing realization. In his research, Bin discovered a correlation between nighttime illumination data and socio-economic indicators in Chinese cities (Wu et al., 2019). Urban tiers can be determined using city lighting data, with the model's predictive outcomes aligning almost precisely with actual city rankings. Conversely, Faisal, in his study, explored the relationship between the industrial built-up areas and GDP across nine cities in Canada (Faisal and Shaker, 2014). A high R-squared value substantiated a robust association between socio-economic and remote sensing data.

These studies collectively validate the potential of harnessing remote sensing data to procure socio-economic information. In our research, we will revisit methods of obtaining socio-economic data through GIS via remote sensing and introduce an approach grounded in our quantitative machine learning models.

3. DATA

1. DHS SURVEY DATA

According to 2023 World Economic Situation and Prospects published by the United Nations (United Nations, 2023), the developing countries of the world today are mainly located in Africa, Asia, Latin America, and the Caribbean, and the least developed countries (LDCs) are mainly located in Africa. Therefore, as shown in Figure 3.1, based on the 2023 World City Populations rankings provided by World Populations Review (World Population Review, 2023), we selected 68 cities with populations of more than 2 million in 27 developing or LDCs in 6 regions, West Africa, South and Central Africa, North and East Africa, South Asia, South East Asia, and Latin America. The 68 cities have a total population of approximately 467 million, 5.8% of the global population, and a complete list of cities by country is attached in Appendix 3.

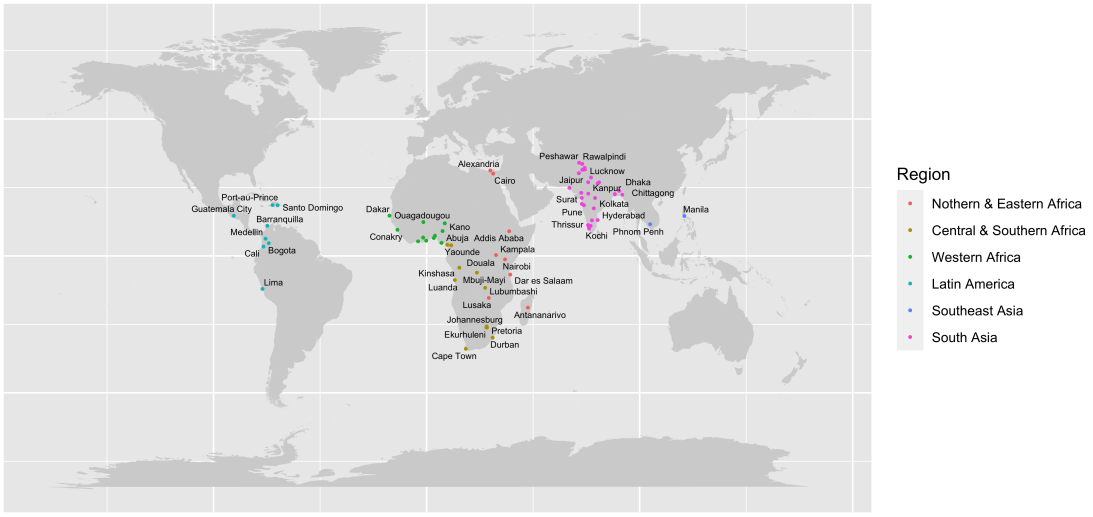

Figure 3.1: *M*aps of 68 Cities Around the World

Among them, India, the world's most densely populated country, has 30 cities with populations of more than 2 million people. In order to avoid the over-representation of data from India, where data imbalances may cause analytical bias leading to a lack of global generalisability of the results, we chose 19 cities in India with populations of more than 3 million people as our sample. India is the largest number of cities in a single country in the sample, accounting for 28% of the total number of all cities in this study.

The main datasets we used were from three sources, DHS program data, Open Buildings data, and World Bank Open Data. The Demographic and Health Surveys (DHS) data are USAID-funded sociodemographic data collection programs in countries around the world (DHS, 2023). Surveys are conducted through questionnaires and interviews every few years to ensure continuity and include a range of socio-economic indicators, education, population health, and

disease for different countries. We selected the most recent Standard DHS data for 27 countries for which both Survey Datasets and GPS Datasets are available, with two exceptions; Peru uses Continuous DHS data from 2014 since the most recent Standard DHS was completed in 2000; and Senegal whose most recent Standard DHS was completed in 2010 and therefore uses Continuous DHS data from 2019 instead. In addition to that the remaining 25 countries' Standard DHS surveys ranging from 2010 to 2022 were used respectively, with Burkina Faso, Colombia, Dominican Republic, Congo Democratic Republic, Egypt, Ghana, and Cote d'Ivoire 7 countries surveyed between 2010 and 2014, and the remaining 18 countries were completed in 2015 and after, a detailed list of DHS data is in Appendix 4.

For Survey Datasets we used individual-level survey data, preprocessed the raw data using the SAS code in Appendix 2 and performed data cleaning using R. The data contains a total of approximately 73,000 respondents from 68 cities in 27 countries, with the largest number being India with around 21,000 respondents, followed by Colombia with around 8,000 respondents, and the remaining 21 countries with 1,000-3,000 respondents respectively.

The core variables include the V190 wealth index, which categorizes respondents on a level from 1 to 5 by their economic and living conditions, and which we use to measure the socio-economic status of the respondents. In particular, we focus on the most populated cities in developing and LDCs, most of which are the economic centers, capitals, or largest cities of the country. Therefore, as shown in Figure 3.2, it is reasonable that a high proportion of respondents are classified as level 5 in the wealth index since the upper class in developing world countries is overwhelmingly clustered in the country's largest cities, while residents of rural areas or other small and medium-sized urban centers are socio-economically disadvantaged compared to their counterparts in the country.

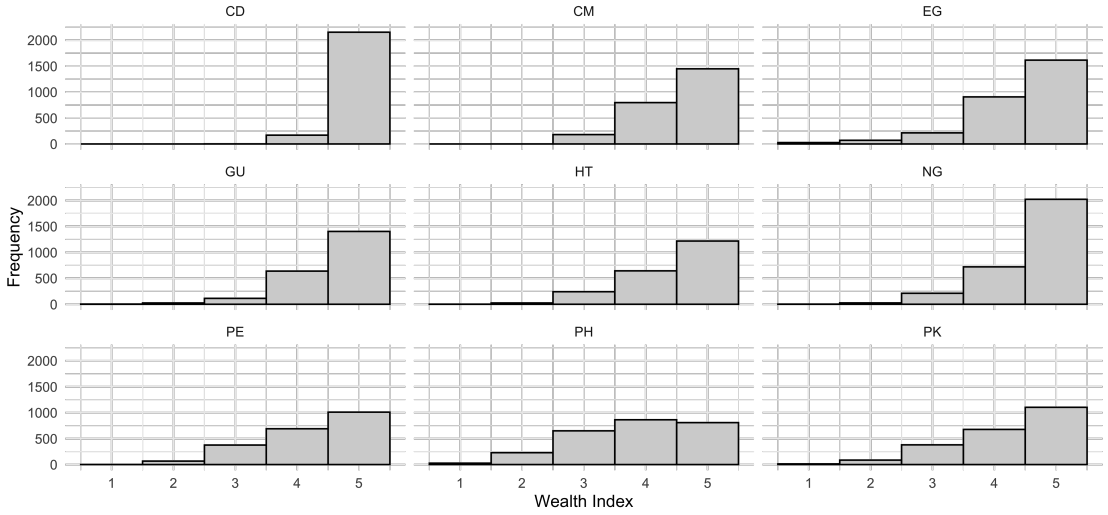

Figure 3.2: *D*istribution of Wealth Index in Different Countries

It is important to note that the index is only valid within countries and represents the socio-economic level of the individual respondent within his or her country. The identical wealth index may suggest vastly different economic levels for different countries. In the case of Tanzania and Colombia, due to the nearly six-fold difference in GDP per capita between Tanzania and Colombia, for example, there is clearly a significant difference in the actual economic level of a Tanzanian with the wealth index of 3 compared to a Colombian with the wealth index of 3. Hence, we cannot make direct comparisons between countries if we only consider the wealth index. In order to quantitatively compare the economic differences between

respondents in different countries around the world, it is necessary to simultaneously consider the economic level of the country and the socio-economic status of the respondent in that country in order to obtain a globally generalized economic indicator for each respondent. Meanwhile, the Gini coefficient measures the equality of distribution within a country, and we use this indicator to measure the differences between wealth and poverty within the country. Therefore, we designed the Estimated Individual Product (EIP), an indicator based on the country's GDP per capita, the country's Gini coefficient, and the DHS wealth index for each respondent.

We used the country's GDP per capita in current US Dollars for the year of the DHS Survey provided by the World Bank (World Bank, 2023). We define the country's GDP per capita as the country's average individual output, which corresponds to level 3 in the wealth index, and the individual output of those in the country's advantaged or disadvantaged position will be adjusted according to this baseline. For the Gini coefficient, we adopted the Gini coefficient for the year closest to the country's DHS Survey in cases where data from the World Bank are incomplete and therefore partially missing (World Bank, 2022). The Gini coefficient is an indicator of the polarisation of wealth within a country derived from the Lorenz curve, so we consider that the larger the Gini coefficient, the greater the difference between the advantaged or disadvantaged in the country and the baseline.

Subsequently, we adjust the difference between different wealth indexes according to the country's GDP per capita level. Obviously the difference between rich and poor in countries with 5000 GDP per capita will be larger than countries with 1000 GDP per capita. We then introduced a constant to rescale this value and finally, we get the Estimated Individual Product as shown in the equation.

$$EIP = (Wealth\ Index - 3) * Gini\ Index * GDP\ Per\ Capita/150 + GDP\ Per\ Capita$$

Table 3.1 presents the EIP corresponding to the wealth index for selected countries. Colombia has a similar level of GDP per capita as India, but Colombia has a much larger Gini coefficient than India, so we find that the variation across wealth index levels is much larger in Colombia than in India. Bangladesh and Burkina Faso have similar Gini coefficients, but Bangladesh has a higher level of GDP per capita and therefore has greater variation across wealth index levels. Also, EIP can be compared between countries. We can assume that an Angolan with a wealth index of 2 has a similar EIP to a Bangladeshi with a wealth index of 3, and therefore has a similar standard of economy and living.

Table 3.1: Comparison of EIP and Wealth Index in Some Countries

Country	GDP Per Capita	Gini Index	Estimated Individual Product (EIP)				
			Wealth Index 1	Wealth Index 2	Wealth Index 3	Wealth Index 4	Wealth Index 5
Angola	3110	51.3	1063.9	2086.9	3110.0	4133.1	5156.1
Bangladesh	2049	32.4	1511.3	1780.1	2049.0	2317.9	2586.7
Burkina Faso	627	35.3	431.7	529.3	627.0	724.7	822.3
Colombia	6393	54.6	1738.9	4065.9	6393.0	8720.1	11047.1
India	5377	35.7	2817.5	4097.3	5377.0	6656.7	7936.5

Other columns of the Survey Datasets that we focused on include, V107, which records information about the respondent's education, the respondent's highest year of education completed, and V149. Respondents' educational attainment, which includes 6 categories of no

education, whether they completed primary school, secondary school, and higher education.

V113, V116, and V129 recorded information about the housing respondents lived in, including source of drinking water, type of toilet facility, and roof top material, respectively. V119, V120, V121, and V122 are a series of dummy variables that record whether the respondent has electricity and some electrical appliances. About 94% of the respondents in our data indicate that they have electricity in their homes, 88% have a TV, and 69% have a refrigerator, which may suggest that the vast majority of the population in large cities in both developing countries and LDCs have access to modern life. V123, V124, and V125 record whether the respondent owns transport, such as a bicycle or a car.

In addition to the individual level, we also investigated at the cluster level, where we aggregated the individual data in each cluster, where EIP, years of schooling, and level of schooling calculated the average of the non-null values in each cluster, and for categorical variables such as roofing material, type of toilet, etc., we created numeric variables to calculate the proportion of each category, such as 71% of the residents in cluster 10 of Angola use water piped to the yard as their source of drinking water and only 25% is water piped into the dwelling. For dummy variables, we also calculate the average of the non-null values, which indicates the proportion of the inhabitants of the cluster who own a refrigerator or a TV.

Besides the DHS Survey Datasets, we also used the DHS GPS Datasets, which record information about the coordinates of each respondent cluster in the survey. In order to protect the privacy of the respondents, the coordinates of each respondent cluster in the city were offset by 2 km in random directions (Mayala et al., 2018), which means that the actual location of respondents in the city could be anywhere within a 2km radius of the coordinates.

The GPS Datasets also contain a range of data obtained through remote sensing, as many variables are updated every 5 years, to ensure data alignment we used the most recent value that each variable had at the time of that DHS Survey. The data includes the population within a 5 x 5 km pixel provided by WorldPop (WorldPop, 2023), and UN Population Count records the population within a 1 x 1 km grid cell (Center for International Earth Science Information Network, 2018). The Global Human Footprint, provided by NASA, a value ranging from 0 to 100 within the 1 x 1 km grid cell records the degree of urbanization within the grid cell (Wildlife Conservation Society and Center for International Earth Science Information Network, 2005). The Global Nighttime Lights provided by NOAA records the nighttime brightness at the location of the cluster (NCEI, 2019), as well as local remote sensing data such as average temperature, diurnal temperature difference, and others, which we use in the model.

2. OPEN BUILDINGS

Open Buildings was built by the Google team based on the U-Net structure Convolutional Neural Network model (Quinn, 2021), using high-resolution satellite images to train the deep learning model, the output accurately identified building footprints (Sirko et al., 2021). The data provides building data covering most areas in Africa, South Asia, Southeast Asia, and Latin America, including the centroid latitude and longitude coordinates, building polygon boundaries, building area, and confidence level of the building detected. As shown in Figure 3.3, the model can accurately identify the regular boundaries of formal buildings and the irregular boundaries of informal buildings.

Figure 3.3: *R*300 Freeway between Philippi and Mandalay, Johannesburg, South Africa

In this study we use Google Earth Engine to process Open Buildings 3rd version data released in May 2023 (Sirko et al., 2021), filter building with confidence level higher than 0.65. We found that the model's confidence level for small or informal buildings is significantly lower than large or formal buildings, therefore, if we increase the confidence level threshold, it may lead to a large number of informal buildings being neglected, after testing 0.65 is the the optimal choice.

Firstly, to cover 68 cities and visualize the differences in living conditions, we used 500m * 500m grids to cover all cities and calculated the average building area within each grid. To improve the data accuracy, we filtered the output grid data with buildings that covered more than 5% of the grid area and had more than 100 buildings within the grid. We used Mapbox and Javascript to build dynamic interactive maps to demonstrate the differences in the web page. Figure 3.4 shows the map of the average area of buildings on the grid for Makoko, Largos, which is known for being the largest water slum in the world, and we can see the dark red water slum at the bottom of the figure. However, the boundaries of the grid did not match the boundaries of the actual building densities, therefore this method is only used as reference and visualization.

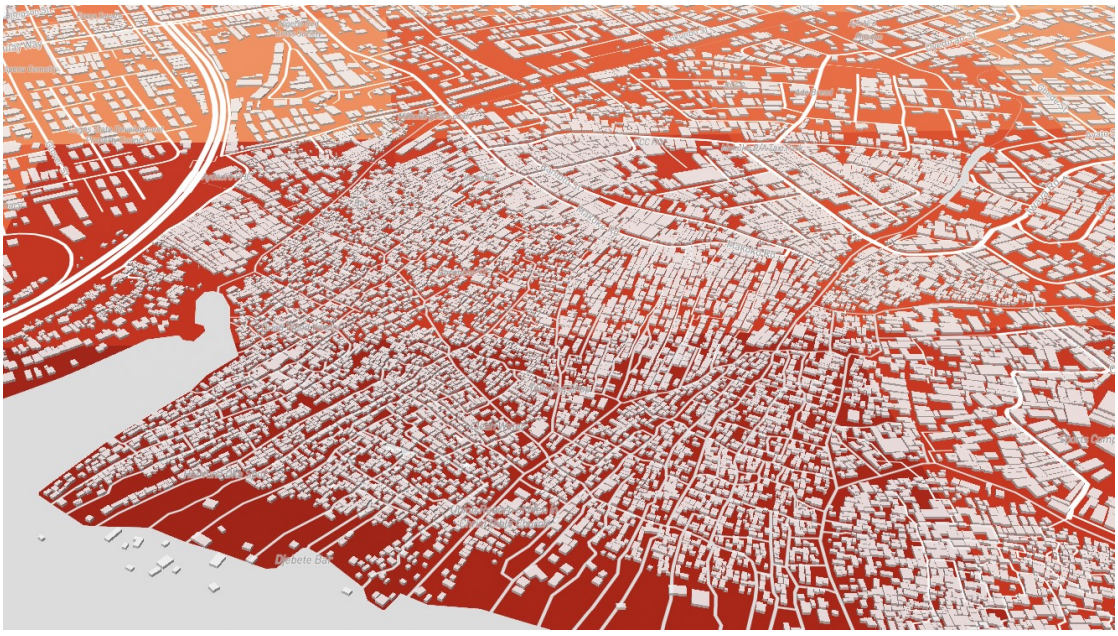

Figure 3.4: 500m Grid Map of Makoko, Lagos, Nigeria

Secondly, since the coordinate of DHS clusters in the city makes random displacements within 2 kilometers in any direction of the coordinate point, respondents may physically be located anywhere in the circle within 2 kilometers radius circle around the coordinate point, therefore we created a buffer with a radius of 2 kilometers for each point and computed the average, median and standard deviation of building area for all buildings in the circle as well as the total building area.

4. METHODOLOGY

1. INFORMAL SETTLEMENTS DETECTION

We found that the average, median, and standard deviation of building area within the 2 km radius circle capture information about informal settlements in the city, for instance, if there are numerous informal settlements in the circular area, the area will have a lower average and median building area compared to others. However, to determine whether an area is an informal settlement, it is important to consider not only the size of each building, but also the size of the neighbouring buildings. For example, in a villa area, besides the main building, there are also ancillary buildings such as garages and security kiosks, which also contribute to a lower median or average building area. Therefore, in order to accurately identify informal building areas in cities, we refer to Ollie Ballinger's methodology in Dar es Salaam Informal Settlement Mapper, which calculates Small Building Density Scores (SBDS) within a 2km radius centred on 3,775 clusters in 68 cities. Density Scores (SBDS).

Firstly based on Open Building's polygon layer we create building area reciprocal raster layer and we take the reciprocal for all the points where the building exists. For large buildings the area will have a smaller value after the derivation, for small buildings the area will have a very high value after the derivation, and for areas where no building exists 0 is taken. Then we define the Gaussian Kernel as shown in Figure 4.1 to convolve the building area reciprocal layer in order to obtain the Small Building Density layer after Gaussian Blur.

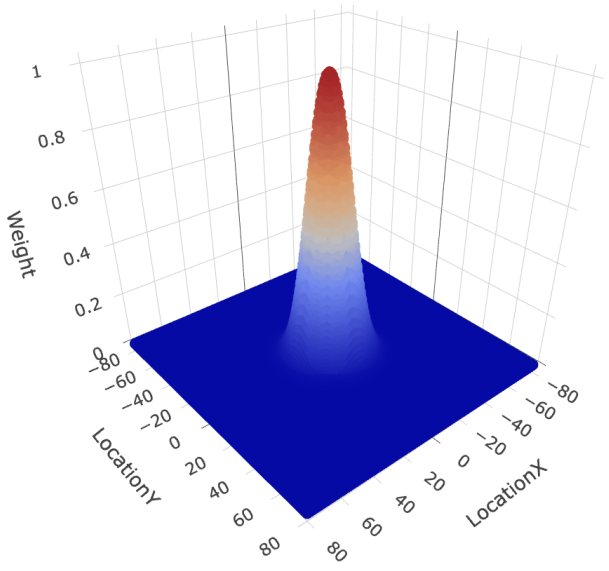

Figure 4.1: *3*D Plot for Gaussian Kernel Weight

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x)^2}{2\sigma^2}}$$

where x is the input value, which is the value in the building area reciprocal layer. σ is the standard deviation of the Gaussian distribution, and we adjust this value to control the breadth of the Gaussian distribution, while we use the radius to define the range of influence of the Gaussian kernel. $F(x)$ is the output value after convolution, which is Small Building Density Scores, where a higher output value means that the point has a higher small building density and is therefore considered more likely to be an informal settlement.

After testing we chose the radius of 80 and $\sigma = 10$ where the kernel has the best performance, accurately identifying the areas of the city that are informal settlements, whilst having clear boundaries with the regular areas of the city.

Compared to directly calculating the average or median building area in the buffer area, this approach better captures to the relationship between the building's area itself and the surrounding building area within the 80-metre radius. The algorithm calculates the inverse of the building area of the point itself, and the sum of the inverse of the building areas within the radius that are progressively less weighted according to normal distribution based on the distance, as shown in Figure 4.1, where the closer the distance within 80 metres the higher the weighting, and the further the distance the lower the weighting.

This suggests that individual low-area structures will not be emphasized, while clusters of highly concentrated low-area buildings will be recognized. This is consistent with the definition by UN-HABITAT of slums or informal settlements as densely populated urban areas where housing within the region is substandard and unsanitary (UN-Habitat, 2003). It mandates that such informal housing clusters achieve a certain magnitude. As a result, our model demonstrates exceptional performance in detecting these informal settlements.

Dharavi, Mumbai, famous for the Academy Award-winning film Slumdog Millionaire, was once the largest slum in Mumbai, with about one million inhabitants living in less than 3 square kilometres of land, and it is still one of the most densely populated areas in the world today. Figures 4.2 to 4.4 show the performance of the model in Dharavi. Figure 4.2 shows the satellite image of Dharavi, where we can see that the majority of the area is covered by continuous small buildings, except a few large buildings. Figure 4.3 shows the building polygon identified by the Open Buildings deep learning model, which identifies the buildings in the area accurately, and the building boundaries match the satellite image. Figure 4.4 shows the output of Small Building Density Layer after Gaussian Kernel processing, we can see that the areas with a large number of small buildings are highlighted in red, while the regular urban areas are coloured in blue. We can see that the boundaries of the areas are clear, while the informal settlement annotations are consistent with the information in the satellite images.

Figure 4.2: *S*atellite Image of Dharavi, Mumbai, India

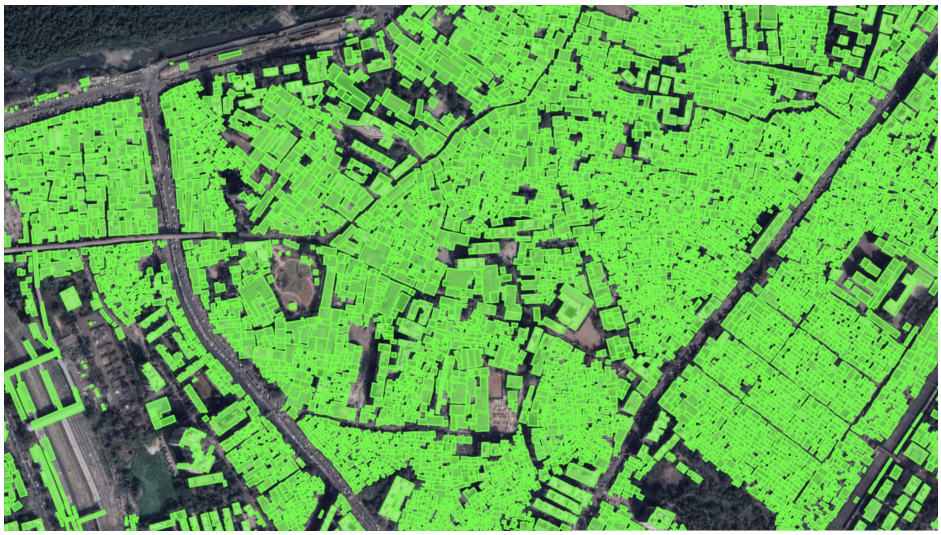

Figure 4.3: *O*pen Buildings Polygon of Dharavi, Mumbai, India

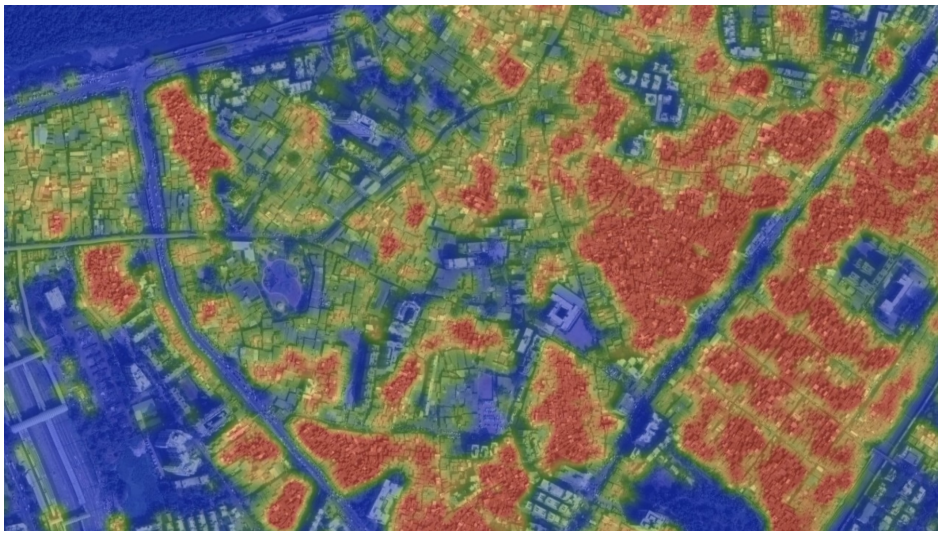

Figure 4.4: *S*BDS Layer of Dharavi, Mumbai, India

2. BUILDING INDICATORS

At individual level, we use data from approximately 73,000 respondents, focusing on 7 values within a 2km radius of each co-ordinate point, including the average, median, standard deviation, and total of the building area, as well as the average, median and standard deviation of the Small Building Density Scores (SBDS).

Figure 4.5 shows the average, median, total and standard deviation of the building area within the 2-kilometre radius of each respondent's co-ordinate point, and we can see that all three values are right skewed, except for the total building area, which is more like a normal distribution.The standard deviation possesses a very high frequency of small values, but this is consistent with the assumption that these values are non-negative and uncapped.

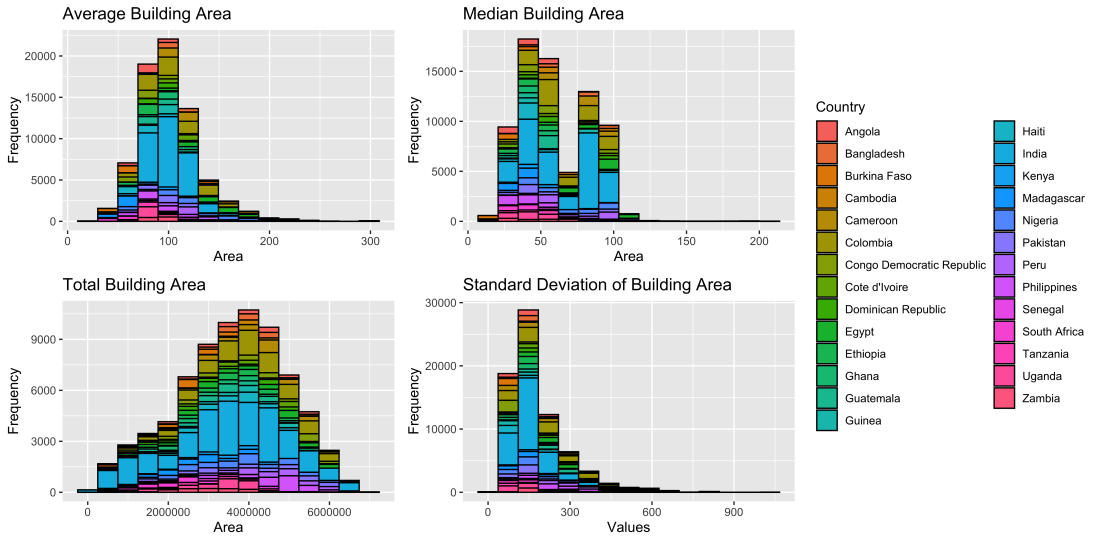

Figure 4.5: *D*istribution of Building Area Indicators at Individual Level

In order to comply with the multiple linear regression assumption, we take logarithms for the mean, median, and standard deviation of the building area, and the distribution of the transformed values is shown in Figure 4.6.

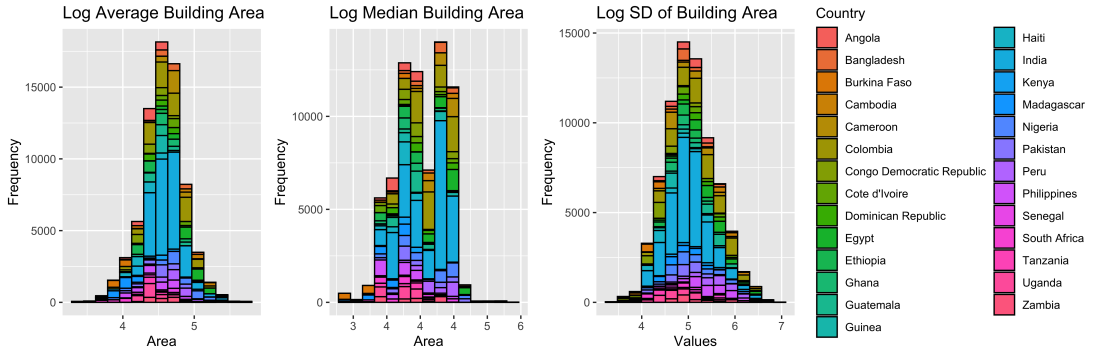

Figure 4.6: *D*istribution of Building Area Indicators at Individual Level after Log Transformation

For the Small Building Density Scores (SBDS), Figure 4.7 shows the mean and standard deviation of the SBDS for the 2-kilometre area around the coordinates of all respondents, and

we find that their distribution approximates to a normal distribution.

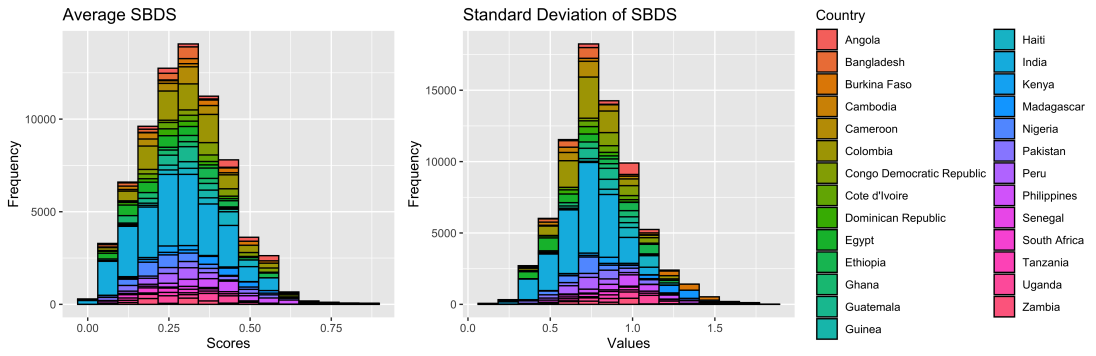

Figure 4.7: *D*istribution of SBDS Indicators at Individual Level

As for the SBDS median, as shown on the left side of Figure 4.8, there exists very strong right skewed, the figure only shows the part of the SBDS median less than 0.005, about 96% of the data. Part of the large value have orders of magnitude differences with the vast majority of the data, so we take the logarithm of the DBDS median, and the converted distribution is shown on the right side of Figure 4.8.

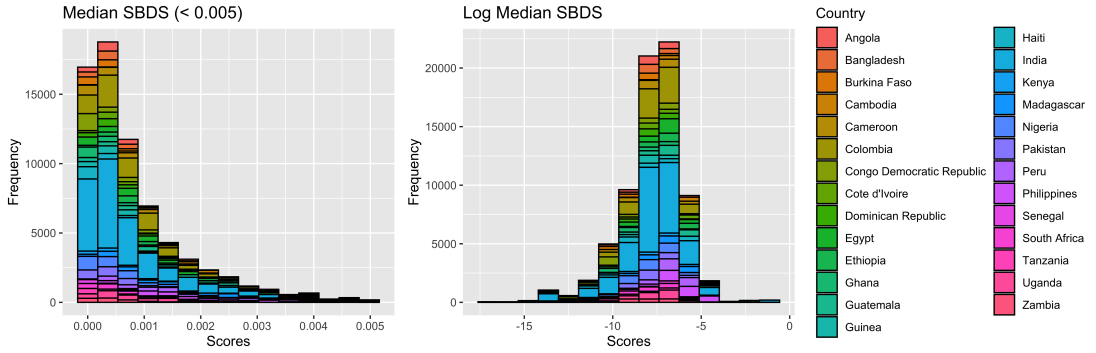

Figure 4.8: *D*istribution of SBDS Median at Individual Level

At cluster level, we take the logarithm of the median and standard deviation of the building area, and the distribution of the building area indicator after processing is shown in Figure 4.9.

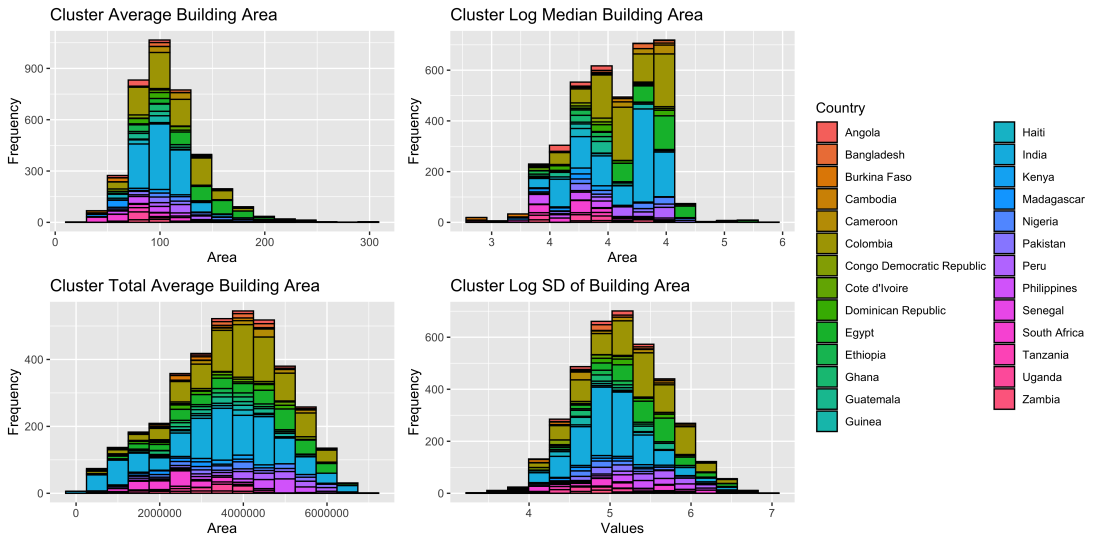

Figure 4.9: *D*istribution of Building Area Indicators at the Cluster Level

For reasons similar to the individual level SBDS median, we took the logarithm of the median, and the distribution of the processed SBDS metrics is shown in Figure 4.10.

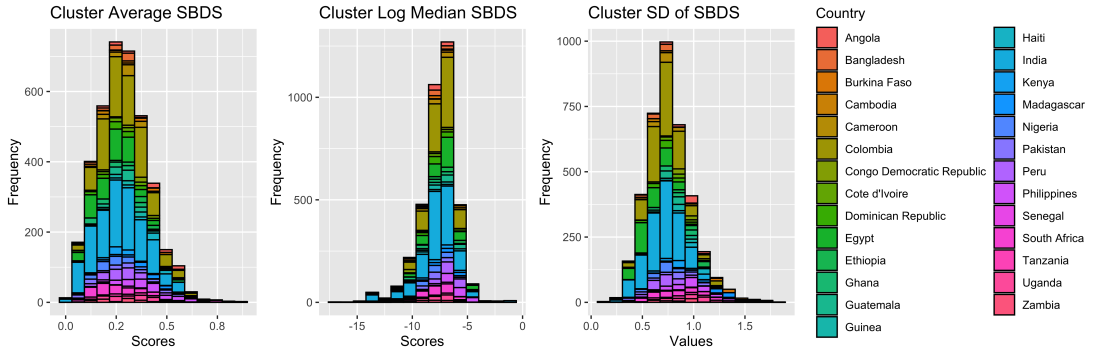

Figure 4.10: *D*istribution of SBDS Indicators at the Cluster Level

Figure 4.11 shows the correlation between the 4 variables based on building area and the 3 variables based on SBDS, we found a strong correlation between the two groups and therefore we will model them separately. Within SBDS group, the correlation between the mean and standard deviation of the SBDS is 0.85, therefore in the modelling of the SBDS we will use the indicator that has been tested to perform better among the mean and the standard deviation to avoid multicollinearity.

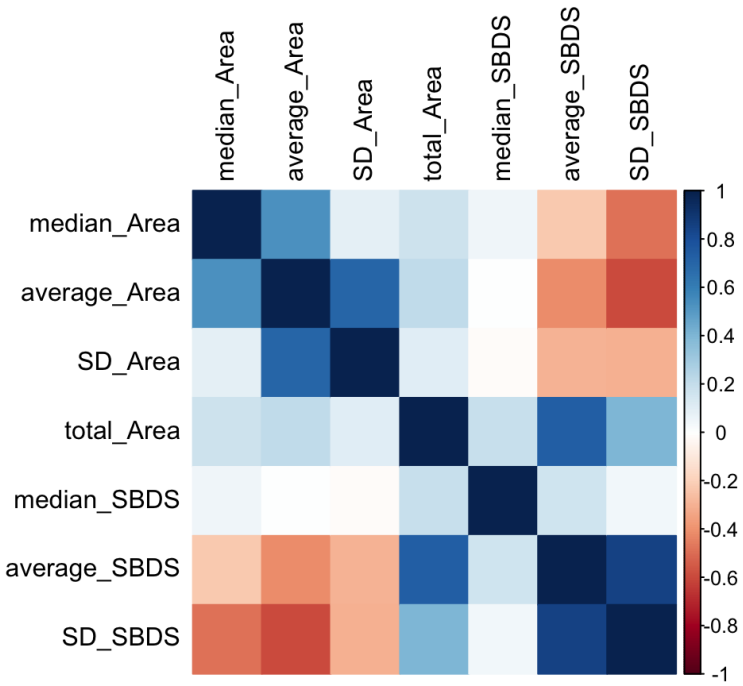

Figure 4.11: *C*orrelation Matrix for Building Indicators

3. SOCIO-ECONOMIC LINKAGES

In order to estimate the socio-economic data of various districts in different cities through the building data from high-resolution satellite imagery, a series of multiple linear regression models were developed to correlate the data obtained through remotely sensed imagery with the DHS survey data at both the individual and cluster levels.

We use the mean, median, standard deviation, and total building area to fit the dependent variable from the DHS Survey, and since the age of the respondents highly correlated with the EIP, we control the age of the respondents, as well as a range of data obtained through remote sensing data including 5 x 5 km of the estimated population, population under 5 years of age, Global human footprint index representing the degree of urbanisation, nighttime light levels in the region, and Population Count and Density provided by the United Nations, where the same data was provided for different years we chose the most recent value according to the timing of the DHS Survey.

We use Variance Inflation Factor to avoid multicollinearity caused by high correlation between control variables to the extent that we cannot accurately estimate the regression coefficient.For different models we select different control variables based on VIF and drop the for different models, we choose different control variables based on VIF and discard non-essential variables with higher VIF.

For each multiple linear regression model, we test the appropriateness of using linear model fit with a residual VS fitted plot, as well as testing the constant variance assumption to ensure that the residual approximates a horizontal line and does not exhibit a distinct pattern. We test for the presence of an outlier with the residual vs leverage plot to avoid the presence of an outlier. Normal Q-Q plot to test whether the data approximately follows a normal distribution, the Scale-Location plot to test the assumption of equal variance. All the models were tested and the potential violation of assumptions will be discussed in the discussions section.

Using model 1 as example, we use the building area data to build a multiple linear regression model on Estimated Individual Product (EIP), and based on the results of the VIF, we keep only the United Nations population data within 5 x 5 km grid. We find that the R-square of the model is 0.454, which means that the mean, median, standard deviation, and total building area, along with the variation in control variables, explain 45.4% of the variation in EIP, and we obtain the regression function:

$$ln(EIP) = \beta_0 + \beta_1 \times ln(Median\ Building\ Area) + \beta_2 \times ln(Average\ Building\ Area) + \beta_3 \times ln(SD\ of\ Building\ Area) + \beta_4 \times Total\ Building\ Area + \sum \gamma_i C_i + \epsilon$$

Where EIP represented the Estimated Individual Product, C_i represents the control variable i,γ_i represents the estimated effect of control variable i, and ϵ is the error term.

We can conclude that when the median building area in the 2km area around the respondent rises by 1% correlated to β_1% increase in EIP on average. And when the average building area in the area rises by 1% correlated to EIP rises by β_2 % on average. When the standard deviation of buildings in the region rises by 1% correlated to EIP rises by β_3 % on average. Based on the model, we can use the remote sensing data and the output of the regression model to estimate the EIP of the respondents. Using this method, we model a series of economic, educational, and living environment indicators at the individual and cluster level.

As we found that the R-square was low in EIP model in some countries, which means that the variation in building area indicators in some countries does not explain the fluctuations in years of education very comprehensively, therefore in model 4, we used four building area indicators, as well as age and country to build a tree model of respondents' education levels (primary, secondary, high school, etc.) to determine the key distinguishing respondents' education level indicators and key values. Since there are about 9,600 respondents in the data who are minors (around 13.26%), while about 15,000 respondents are younger than or equal to 20 years of age (around 19.96%), to avoid ignoring the effect of age on the highest possible education level, we added age to the tree model as well. After testing when Complexity Parameter is set to 0.001 we can get the decision tree with tree height of 8. The results of the decision tree will be discussed in the results section.

We focus on the proportion that calculated based on categorical variables or dummy variables in the individual level in model 12, such as the proportion of drinking water in the cluster that comes primarily from piped into dwelling or the the proportion of electricity in the cluster. Since the dependent variable is a probability that falls between 0 and 1, we first use linear regression to model using the dependent variables of interest and use the sigma function to map the results of the linear model to obtain a probability that its estimate is between 0 and 1, as shown in the following equation:

$$P = \frac{1}{1+e^{-Z}}$$
$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \sum \gamma_i C_i + \epsilon$$

Since our predicted value is the probability of having a specific residence condition in that cluster, which is a continuous value between 0 and 1, and not as a binary classifier, the commonly used ROC curves or AUC values for models that have been logistically transformed are not applicable, instead we use the AIC (Akaike Information Criterion) and the BIC (Bayesian Information Criterion) to determine the degree of model fit and compare the results, and use RMSE (Root Mean Squared Error) to observe the difference between the model predicted value and the actual value to evaluate the prediction. R-squared is used to determine the magnitude of the explanation of variation of dependent variable by variation in independent variables in the model.

5. RESULTS

We obtained our quantitative results in a series of models. In Model 1 we focus on the relationship between the building area indicator and Estimated Individual Product (EIP). In Model 2, we focus on the relationship between the building area indicator and the wealth index, and since the wealth index represents the socio-economic status of the individual within that country and cannot be directly compared between countries, we build separate multivariate linear regressions for each country. In Model 3, our focus is on the respondents' year of education. The educational duration of urban populations in developing and LDCs is significantly influenced by national educational policies. For instance, during Nasser's era in Egypt, the implementation of a national free education policy led to a rapid proliferation of education in a short span. Consequently, to account for the varying impacts of different countries on educational duration, we incorporated the country as a control variable in our model. We then established a panel regression model using building area metrics, controlling for country-specific fixed effects. In model 5 we focus on the correlation of a range of respondents' living condition-related variables with the building area indicator, using four multiple linear regression models using variables including the material of the respondent's roof, the source of drinking water, the type of latrine, and the availability of electricity in the household.

We repeat the steps in model 1 to model 3 to build model 6 to model 8 using the Small Building Density Scores (SBDS) metrics, and repeat the steps in model 5 to build model 9. The results of the model and the comparison of the SBDS metrics with the building area metrics will be discussed in the results section.

In model 10, we built a multiple linear regression model to predict the average EIP of each cluster through the building area indicator, and selected remotely sensed data as control variables using the VIF method. Vegetation Index, average annual land surface temperature, average number of days of rain per month have been added compared to the individual level. In model 11, a multiple linear regression model was built to predict the average year of education in each cluster using the building area indicator. In model 13, similar to model 5, we build a multiple linear regression model using a range of building-related variables including roof type, toilet type, and other indicators of building area at cluster level. In contrast to model 5, which contains dummy variables created based on categorical data, in model 13 we calculate the proportion of each type in each cluster, thus the dependent variables are a series of continuous variables.

In model 14 to model 17, we repeat the steps of model 10 to model 13 to build the model using the SBDS indicator. In model 16, since we are concerned with the degree of model fit and predictive accuracy and do not consider the effects of the individual variables, we include both the SBDS average and standard deviation in the model.

All the 17 models developed in this study are shown in the Table 5.1 below. We discuss the results of the 17 models separately in three sections: economy, education, and living conditions respectively.

Table 5.1: Table of All 17 Models in This Study

	Economics		Education		Living Conditions	
	EIP	Wealth Index	Year of Education	Education Level	Housing	Building Indicators
Individual BA	Model 1	Model 2	Model 3	Model 4		Model 5
Individual SBDS	Model 6	Model 7	Model 8			Model 9
Cluster BA	Model 10		Model 11		Model 12	Model 13
Cluster SBDS	Model 14		Model 15		Model 16	Model 17

1. ECONOMICS

We focus on the economy from the EIP perspective, where we consider that higher EIP implies higher productivity and better economic conditions. We build multiple linear regression models to explore the correlation between EIP using building area indicators and SBDS with Gaussian Kernel convolution, respectively. The results of the models are presented in Table 5.2.

Table 5.2: Estimation of EIP by Building Indicators

	Individual EIP		Cluster EIP	
	Building Area Model 1	SBDS Model 6	Building Area Model 10	SBDS Model 14
$medianArea\_log$	0.2178^{***}		0.3262^{***}	
	(0.0081)		(0.0325)	
$avgArea\_log$	0.3002^{***}		-0.0016^{**}	
	(0.0145)		(0.0005)	
$stdDevArea\_log$	0.1034^{***}		0.2141^{***}	
	(0.0066)		(0.0281)	
$totalArea$	-0.0000^{***}		-0.0000^{***}	
	(0.0000)		(0.0000)	
$medianSBDS\_log$		0.0244^{***}		0.0136^{*}
		(0.0015)		(0.0067)
$stdDevSBDS$		-1.2114^{***}		-1.0662^{***}
		(0.0101)		(0.0460)
age	✓	✓		
$control\_group\_1$	✓	✓	✓	✓
$control\_group\_2$			✓	✓
$constant$	5.5061^{***}	9.0133^{***}	6.7548^{***}	9.5180^{***}
	(0.0376)	(0.0250)	(0.2011)	(0.1303)
RSE	0.5836	0.5593	0.5315	0.5113
R^2	0.4541	0.4987	0.4298	0.4720
$AdjustedR^2$	0.4540	0.4986	0.4277	0.4704
$Observations$	71887	71887	3299	3299

$^{*}\ p < 0.05$; $^{**}\ p < 0.01$; $^{***}\ p < 0.001$

On the individual level, from model 1, we find that a 10% increase in median building area within 2km of individual is correlated to 2.2% increase in individual EIP on average, and 10% increase in average building area is correlated to 3% increase in individual EIP. Also, from model 2 we find that the average SBDS shows a stronger correlation when the SBDS metric is used, with 0.1 increase in the standard deviation of SBDS correlating with 11.41% decrease in EIP on average, calculated as below.

$$(e^{0.1\times-1.2114}-1)\cdot 100\% = -11.41\%$$

In cluster level, we find that the effects of median and standard deviation of building area are reinforced and the effect of mean building area almost disappears. 10% increase in median building area is correlated to 3.3% increase in cluster EIP on average, and 10% in crease in standard deviation of building area is correlated to 2.1% in crease in cluster EIP. The use of SBDS metrics at the cluster level did not change significantly from the individual level. Also we find that the model using the SBDS indicator has better performance, in terms of smaller RSE and larger R-squared, which means that the model fits the data better and can also explain more of the variation in EIP.

We use the same approach to model the individual wealth index within countries, and the results for Colombia, Ethiopia and India are presented in Table 5.3. We find that the same index may have completely opposite significant effects in different countries. For example, 10% increase in median building area is correlated to 1.4% decrease in the wealth index in Colombia, while it is correlated to more than 2% increase in Ethiopia and India. At the same time, the effects of certain indicators may vary by orders of magnitude across countries, such that an increase of 5% in the standard deviation of building area is correlated to 0.91 wealth index decrease in Colombia, 0.25 decrease in Ethiopia, and 3.3 decrease in India on average, which also illustrates the wide variation in architectural differences within countries themselves. For the SBDS indicator, in order to avoid the high correlation between the average and the standard deviation, after testing the model with the average having better performance than the standard deviation, therefore we use the SBDS average. We find significance in the average SBDS in Colombia, the median SBDS in Ethiopia, and both in India.

Table 5.3: Estimation of the Wealth Index by Building Indicators in Different Countries

	Colombia Wealth Index		Ethiopia Wealth Index		India Wealth Index	
	Building Area Model 2	SBDS Model 7	Building Area Model 2	SBDS Model 7	Building Area Model 2	SBDS Model 7
$medianArea\_log$	-0.1414^{***}		0.2970^{***}		0.2170^{*}	
	(0.0301)		(0.0546)		(0.1075)	
$avgArea\_log$	0.7280^{***}		0.4704^{***}		2.1250^{***}	
	(0.0823)		(0.0965)		(0.1954)	
$stdDevArea\_log$	-0.1823^{***}		-0.0504		-0.6589^{***}	
	(0.0255)		(0.0525)		(0.0460)	
$totalArea$	0.0000		0.0000^{***}		0.0000^{***}	
	(0.0000)		(0.0000)		(0.0000)	
$medianSBDS\_log$		-0.0025		0.0834^{***}		-0.0623^{***}
		(0.0035)		(0.0097)		(0.0125)
$avgSBDS$		0.3705^{***}		-0.2853		2.2553^{***}
		(0.0633)		(0.1799)		(0.2344)
age	✓		✓		✓	
$control\_group\_1$	✓	✓	✓	✓	✓	✓
$control\_group\_2$		✓		✓		✓
$constant$	2.9102^{***}	4.4066^{***}	-0.4105	4.2228^{***}	-3.7461^{***}	1.9845^{***}
	(0.1936)	(0.0556)	(0.3252)	(0.1930)	(0.7692)	(0.3047)
RSE	0.2275	0.2309	0.6635	0.6756	0.6471	0.6942
R^2	0.2742	0.2515	0.3342	0.3091	0.2482	0.1339
$AdjustedR^2$	0.2711	0.2489	0.3318	0.3071	0.2447	0.1306
$Observations$	2325	2325	2834	2834	2131	2131

$^{*}\ p < 0.05$; $^{**}\ p < 0.01$; $^{***}\ p < 0.001$

2. EDUCATION

Education is another important issue we focus on, and Table 5.4 presents the results of the 4 models on years of education. We find that the R-squared of the models at the individual level is low at only 8%, which means that changes in building indicators and control variables at the individual level explain only a small fraction of the variation in the years of education. This is also consistent with the common sense that there are many factors that affect the length of education of each individual, and the building in which they live explains only a small fraction of them. Therefore, at the individual level we build tree model to identify important variables, and the results of model 4 are presented in Figure 5.1, where the first few levels of categorization are mainly determined by the age of the respondent and the county where the respondent is located. Age determines the highest level of education currently possible, while different education policies in different countries create notable differences between countries. The building indicator, on the other hand, only functions in the bifurcation at the end of the tree, for example, among adult respondents in Pakistan, an average building area of less than 72 square meters in the vicinity of 2 kilometers of the respondent has a 56% probability that he or she is not educated.

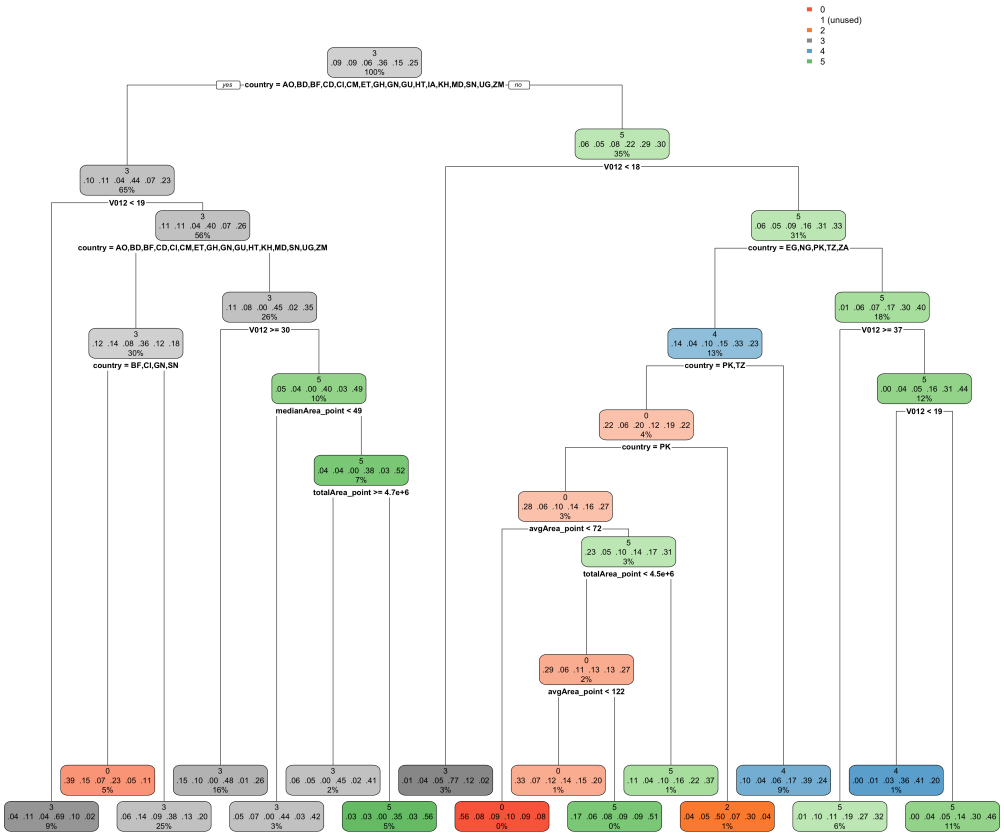

Figure 5.1: *T*ree Model Output at Individual Level Using Building Area Indicators

For the cluster level model, which has R-squared of approximately 39%, where the average years of education in the cluster averaging out the large variation that exists between individuals, thus the building metrics carry greater explanatory power. We find significance in the standard deviations of the indicators that capture equality within cluster, SD of both building area and SBDS have negative effects, implying that an increase in inequality might be correlated with decrease in the years of education. Model 15 shows that 1 increase in the standard deviation of the SBDS is correlated to 0.22 decrease in years of education within cluster.

Table 5.4: Estimation of Years of Education by Building Indicators

	Individual Year of Education		Cluster Year of Education	
	Building Area Model 3	SBDS Model 8	Building Area Model 11	SBDS Model 15
$medianArea\_log$	0.0464		0.0645	
	(0.0281)		(0.0389)	
$avgArea\_log$	0.2392^{***}		0.0020^{**}	
	(0.0581)		(0.0006)	
$stdDevArea\_log$	-0.0868^{***}		-0.0932^{**}	
	(0.0257)		(0.0346)	
$totalArea$	0.0000^{*}		0.0000	
	(0.0000)		(0.0000)	
$medianSBDS\_log$		0.0082		0.0049
		(0.0056)		(0.0083)
$stdDevSBDS$		-0.1581^{***}		-0.2203^{**}
		(0.0468)		(0.0674)
age	✓	✓		
$country$	✓	✓	✓	✓
$control\_group\_1$	✓	✓	✓	✓
$control\_group\_2$			✓	✓
$constant$	2.4650^{***}	3.4722^{***}	3.7935^{***}	3.9790^{***}
	(0.1621)	(0.1035)	(0.3398)	(0.2902)
RSE	1.8325	1.8330	0.5936	0.5944
R^2	0.0823	0.0819	0.3898	0.3877
$AdjustedR^2$	0.0818	0.0814	0.3830	0.3814
$Observations$	65651	65651	3293	3293

$^{*}\ p < 0.05$; $^{**}\ p < 0.01$; $^{***}\ p < 0.001$

3. LIVING CONDITIONS

We use series of variables related to residential environment to estimate building indictors, and Table 5.5 shows the results of the model at the individual level, and Table 5.6 presents the cluster level.

On the individual level, we find that the R-squared of the model ranges between 12% to 24%, implying that different living environments explain part of the fluctuations in the building indicators. Building area and SBDS show different preferences across indicators, but the differences are insignificant.

Table 5.5: Estimation of Building Indicators by Living Conditions at Individual Level

	Building Area Model 5			SBDS Model 9		
	log_median	log_average	log_SD	log_median	average	SD
water_source_piped_yard	-0.0394^{***}	-0.0443^{***}	-0.0585^{***}	-0.2314^{***}	0.0012	0.0068^{*}
	(0.0048)	(0.0035)	(0.0066)	(0.0223)	(0.0017)	(0.0028)
water_source_public_tap	-0.0630^{***}	-0.0997^{***}	-0.0431^{***}	-0.0706^{**}	0.0074^{***}	0.0452^{***}
	(0.0058)	(0.0042)	(0.0079)	(0.0268)	(0.0020)	(0.0033)
water_source_tube_well	0.0616^{***}	-0.0110^{*}	-0.1260^{***}	-0.8817^{***}	-0.0412^{***}	-0.0754^{***}
	(0.0060)	(0.0043)	(0.0081)	(0.0276)	(0.0021)	(0.0034)
water_source_bottled_water	0.0092	0.0115^{**}	0.0717^{***}	-0.0691^{**}	-0.0226^{***}	-0.0370^{***}
	(0.0056)	(0.0041)	(0.0077)	(0.0260)	(0.0020)	(0.0032)
toilet_flush_septic_tank	-0.0145^{***}	-0.0678^{***}	-0.1869^{***}	-0.5860^{***}	-0.0070^{***}	-0.0013
	(0.0038)	(0.0028)	(0.0053)	(0.0179)	(0.0013)	(0.0022)
toilet_flush_pit_latrine	-0.0053	-0.0320^{***}	-0.1689^{***}	-0.7296^{***}	-0.0304^{***}	-0.0291^{***}
	(0.0073)	(0.0053)	(0.0099)	(0.0336)	(0.0025)	(0.0041)
toilet_pit_latrine	-0.0861^{***}	-0.1605^{***}	-0.3632^{***}	-0.6946^{***}	0.0230^{***}	0.0841^{***}
	(0.0054)	(0.0039)	(0.0073)	(0.0250)	(0.0019)	(0.0031)
toilet_open_pit	-0.0446^{***}	-0.1608^{***}	-0.3193^{***}	-0.8993^{***}	0.0088^{**}	0.0540^{***}
	(0.0088)	(0.0064)	(0.0119)	(0.0404)	(0.0030)	(0.0050)
roof_metal	-0.1038	-0.0354	0.0931	0.7987^{*}	0.0449	0.0959^{*}
	(0.0708)	(0.0511)	(0.0955)	(0.3247)	(0.0245)	(0.0401)
roof_roofing_shingles	0.1753^{*}	0.1756^{***}	0.1532	0.8112^{*}	0.0033	-0.0145
	(0.0724)	(0.0523)	(0.0977)	(0.3323)	(0.0251)	(0.0410)
electricity	0.2198^{***}	0.2240^{***}	0.3100^{***}	0.7856^{***}	-0.0065^{*}	-0.0822^{***}
	(0.0079)	(0.0057)	(0.0107)	(0.0364)	(0.0027)	(0.0045)
radio	-0.0630^{***}	-0.0463^{***}	-0.0283^{***}	-0.1268^{***}	-0.0026^{*}	0.0382^{***}
	(0.0034)	(0.0024)	(0.0045)	(0.0154)	(0.0012)	(0.0019)
television	0.0203^{***}	0.0159^{***}	0.0247^{***}	0.1488^{***}	0.0082^{***}	0.0079^{**}
	(0.0051)	(0.0037)	(0.0069)	(0.0233)	(0.0018)	(0.0029)
fridge	0.0340^{***}	0.0164^{***}	0.0033	-0.0287	-0.0048^{***}	-0.0213^{***}
	(0.0036)	(0.0026)	(0.0048)	(0.0164)	(0.0012)	(0.0020)
constant	3.8795^{***}	4.4331^{***}	4.8484^{***}	-8.5081^{***}	0.2862^{***}	0.8310^{***}
	(0.0713)	(0.0515)	(0.0962)	(0.3271)	(0.0247)	(0.0404)
RSE	0.3533	0.2552	0.4767	1.6215	0.1223	0.2001
R^2	0.1410	0.1991	0.1569	0.1736	0.1206	0.2402
$AdjustedR^2$	0.1401	0.1982	0.1559	0.1727	0.1196	0.2394
Observations	60719	60719	60719	60719	60719	60719

$^{*}\ p < 0.05$; $^{**}\ p < 0.01$; $^{***}\ p < 0.001$

In terms of source of drinking water, respondents had an 3.9% lower median building area, 4.3% lower average building area, and 0.0012 higher in average SBDS if their water source was from water piped into yard or plot compared to water piped into dwelling. For water source is from public tap or standpipe it has 6.1% lower median building area, 9. 5% lower mean building area, and 0.0074 higher average SBDS compare to the baseline.

For toilets, if the respondent has a pit latrine toilet, there is 8.2% lower median building area,

14.8% lower average area, and 0.023 higher in average SBDS compared to toilets with flush to piped sewer system, and 4.4% lower median building area and 14.85% lower average area for pit latrines without slab or open pit.

For roofing materials compared to no roofing, using roofing shingles is correlated to 19% higher median and average building area. In addition to this, having electricity is correlated to 24.6% higher median building area, 25.1% higher mean building area, and 0.0065 lower average SBDS.

At cluster level we use the proportions of specific major types of toilets, roofs, etc. as independent variables for the cluster, and therefore we do not need to convert the categorical variables into dummy variables as we did in the individual level model and compare the results to the baseline. The coefficients in cluster level model are meaningful in their own. The cluster level model did not show significant differences in performance from the individual level.

Table 5.6: Estimation of Building Indicators by Living Conditions at Cluster Level

	Building Area Model 13			SBDS Model 17		
	log_median	average	log_SD	log_median	log_average	log_SD
$water\_source\_piped\_yard$	-0.1018^{*}	0.9692	0.2202^{***}	1.6004^{***}	0.4285^{***}	0.2102^{***}
	(0.0414)	(3.1059)	(0.0565)	(0.1886)	(0.0600)	(0.0309)
$water\_source\_public\_tap$	-0.2039^{***}	-8.2075^{*}	0.3212^{***}	2.2029^{***}	0.4839^{***}	0.3074^{***}
	(0.0451)	(3.3850)	(0.0616)	(0.2056)	(0.0654)	(0.0336)
$water\_source\_bottled\_water$	0.0681	13.5757^{***}	0.3991^{***}	1.8617^{***}	0.3563^{***}	0.1346^{***}
	(0.0507)	(3.8016)	(0.0692)	(0.2309)	(0.0735)	(0.0378)
$water\_source\_sachet\_water$	-0.2859^{***}	-2.1207	0.4495^{***}	2.6411^{***}	0.7178^{***}	0.4058^{***}
	(0.0433)	(3.2484)	(0.0591)	(0.1973)	(0.0628)	(0.0323)
$toilet\_flush\_pit\_latrine$	-0.0987	-22.3575^{***}	-0.5143^{***}	-0.4318	0.2476^{***}	0.2097^{***}
	(0.0512)	(3.8446)	(0.0700)	(0.2335)	(0.0743)	(0.0382)
$toilet\_open\_pit$	0.0748	-17.1121^{**}	-0.2649^{*}	-1.1734^{**}	-0.0157	0.0506
	(0.0837)	(6.2818)	(0.1143)	(0.3815)	(0.1214)	(0.0624)
$roof\_metal$	-0.0841	11.0881^{*}	0.3629^{***}	0.4177	0.3491^{***}	0.2352^{***}
	(0.0629)	(4.7197)	(0.0859)	(0.2866)	(0.0912)	(0.0469)
$roof\_wood$	-0.2097^{*}	3.0159	0.5353^{***}	0.8728^{*}	0.3342^{*}	0.2178^{**}
	(0.0931)	(6.9830)	(0.1271)	(0.4241)	(0.1350)	(0.0694)
$roof\_roofing\_shingles$	0.5427^{***}	58.5384^{***}	0.5860^{**}	0.6448	0.0400	-0.0259
	(0.1327)	(9.9571)	(0.1812)	(0.6047)	(0.1925)	(0.0989)
$constant$	4.0080^{***}	86.3073^{***}	4.5983^{***}	-9.2111^{***}	-1.9170^{***}	-0.5645^{***}
	(0.0734)	(5.5088)	(0.1003)	(0.3345)	(0.1065)	(0.0547)
RSE	0.3516	26.3744	0.4801	1.6017	0.5099	0.2620
R^2	0.1554	0.1518	0.1759	0.1868	0.1391	0.2466
$AdjustedR^2$	0.1485	0.1449	0.1692	0.1802	0.1322	0.2405
$Observations$	2615	2615	2615	2615	2615	2615

$^{*}\ p < 0.05$; $^{**}\ p < 0.01$; $^{***}\ p < 0.001$

For drinking water sources, each 25% increase in the proportion of cluster residents using public tap or standpipe is correlated with a 2.1 square metre decrease in the average building area of the cluster and a 0.13 increase in the average SBDS. Also a 25% increase in the proportion of

cluster residents using bottled water as their primary source of drinking water is correlated to 3.4 square metres increase in average building area and 0.09 increase in average SBDS. For the type of latrine, if the proportion of latrines flush to pit latrine decreases by 25% is correlated with an increase of 5.6 square metres in average building area and a decrease of 0.062 in average SBDS within cluster. A 25% increase in the proportion of roofing shingles within cluster is associated with a 14.6 square metre increase in mean building area.

Subsequently, Table 5.7 to 5.10 shows how we fit each drinking water source, toilet type, roofing material, and the proportion of appliances and transportation owned in the cluster using building metrics with remotely sensed data at the cluster level, and we find that some of the models have high R-squared, which means that we can capture fluctuations in the questionnaire or interview survey data through processed remotely sensed data alone, with the models can be used for estimation.

The performance of the predictive models for 6 most common drinking water sources is shown in Table 5.7, we find that the R-squared of piped into dwelling model and sachet water model are higher than 50%, which means that the fluctuations of our remotely sensed data can already capture more than half of the fluctuations of the related survey data. Aachet water model using the building area metric's has lower AIC and BIC compared to the SBDS metrics, which means that the model with the building metrics has better performance when considering the number of parameters, avoiding overfitting and underfitting simultaneously. The RMSE of the model is only 0.13, which can be considered as a relatively accurate predictive model.

Table 5.7: Estimation of Water Sources by Building Indicators

	piped_dwelling		piped_yard		public_tap		bottled_water		sachet_water	
	BA	SBDS	BA	SBDS	BA	SBDS	BA	SBDS	BA	SBDS
AIC	2857	2846	1448	1448	1018	991	1121	1122	619	626
BIC	2936	2920	1528	1522	1098	1065	1201	1195	698	700
RMSE	0.2793	0.2788	0.1833	0.1835	0.1585	0.1550	0.1687	0.1695	0.1352	0.1375
R^2	0.5970	0.5984	0.2128	0.2115	0.2234	0.2572	0.3359	0.3297	0.5298	0.514
Obs	3299	3299	3299	3299	3299	3299	3299	3299	3299	3299

In Table 5.8, the R-square of the model for the proportion of toilet flush to piped sewer system and pit latrine toilets is 58% and 48%, and the RMSE of the model for the pit latrine toilets with the SBDS indicator is only 0.15, which can be estimated with greater accuracy.

Table 5.8: Estimation of Toilet Facility by Building Indicators

	flush_piped_sewer_system		flush_pit_latrine		pit_latrine	
	BA	SBDS	BA	SBDS	BA	SBDS
AIC	2726	2708	459	459	1130	1097
BIC	2806	2782	539	532	1210	1170
RMSE	0.3049	0.304	0.0962	0.0965	0.1519	0.1501
R^2	0.5809	0.5833	0.2248	0.2202	0.4725	0.4852
$Observations$	3299	3299	3299	3299	3299	3299

The models have the better performance for rooftop materials, and the relevant results are shown in Table 5.9, where we find that the R-squared of the models for metal and wood is more than 50%, and the cement and rustic mat models have more than 70% R-squared. The RMSE of the wood and rustic mat models are only 0.05 and 0.07, which means that we can predict the proportion of the above two types of roofs in the cluster fairly accurately.

Table 5.9: Estimation of Roof Material by Building Indicators

	metal		cement		ceramic_tiles		wood		rustic_mat	
	BA	SBDS	BA	SBDS	BA	SBDS	BA	SBDS	BA	SBDS
AIC	2037	1929	1400	1446	528	527	246	245	108	107
BIC	2111	1998	1475	1515	602	595	321	313	183	176
RMSE	0.3169	0.3034	0.2150	0.2215	0.1464	0.1465	0.0785	0.0796	0.051	0.0527
R^2	0.5299	0.5689	0.7405	0.7247	0.2986	0.2980	0.5702	0.5577	0.706	0.6867
Obs	2228	2228	2228	2228	2228	2228	2228	2228	2228	2228

For the other models presented in Table 5.10, we find that the models with electricity access percentage also have good performance, with better AIC and BIC for the SBDS metrics, R-squared higher than 50% while the RMSE is only 0.08.

Table 5.10: Estimation of Electricity and Transportation by Building Indicators

	electricity		radio		television		fridge		motor	
	BA	SBDS	BA	SBDS	BA	SBDS	BA	SBDS	BA	SBDS
AIC	304	282	2657	2639	809	798	2614	2566	2629	2629
BIC	384	355	2737	2712	888	871	2693	2640	2709	2702
RMSE	0.0817	0.0816	0.2313	0.2309	0.1173	0.1161	0.2144	0.2109	0.2249	0.2247
R^2	0.5137	0.5153	0.6557	0.6569	0.3279	0.342	0.3989	0.4183	0.5607	0.5612
Obs	3299	3299	3299	3299	3299	3299	3299	3299	3299	3299

Observing the performance of the model 12 and model 16 models, we find that the SBDS metrics and the building area metrics do not differ significantly in their fit to the different areas, and thus can be used as a reference at the same time.

6. DISCUSSION

This study acknowledges several inherent limitations, particularly concerning our data selection, where we encountered significant constraints. The most recent Open Buildings version 3 dataset has yet to provide comprehensive coverage of all developing nations, notably omitting building footprints from pivotal regions such as the Middle East, South Asia, and Eastern Europe. Additionally, the Demographic and Health Surveys, funded by the United States Agency for International Development, carries an official affiliation with the U.S. government. Due to international relations and geopolitical considerations, the DHS Survey has not been implemented in certain countries, leading to data gaps spanning recent decades. Consequently, our research was limited to countries covered by both datasets. Of the over 100 developing economies worldwide, our study encompasses only 27. Despite covering 467 million people, the omission implies that a substantial segment of the urban population in the developing world remains outside the purview of our analysis.

As the Open Buildings project continues to evolve, it is anticipated that future research will incorporate a broader spectrum of cities. Our confidence in the expansive coverage of Google's Open Buildings is underscored by its remarkable growth, the version 1 data covered a mere 19.4 million square meters 2 years ago, which has since surged to 58 million square meters in the latest release. It is a reasonable projection that comprehensive coverage of developing economies will be achieved in the near future. On the socio-economic data front, future studies should consider diversifying data sources beyond the DHS. Instruments such as the UNICEF Multiple Indicator Cluster Surveys could be integrated to ensure a more extensive data coverage. In this study, we employed population as the primary criterion for city selection, adopting a people-centric approach. However, this might have inadvertently overlooked other pivotal factors. For instance, from an urban spatial perspective, the built-up area of a city, and from an economic standpoint, the city's total output, can also significantly influence its importance. Future research may consider contemplate integrating these elements when determining city selection.

Furthermore, Open Buildings does not explicitly disclose the specific dates of the high-resolution satellite images utilized, and images might originate from different years. While all images are relatively recent, there may be discrepancies in the timing of satellite imagery across various cities. Concurrently, our study predominantly employs DHS data post-2010, with two-thirds stemming from 2015 onwards. However, certain developing countries, characterized by a low economic baseline and rapid economic fluctuations, might witness significant growth or decline in socio-economic indicators within a few years. This study, regrettably, does not fully account for these nuances, often treating the data as if it represents a singular point in time. Additionally, there might be temporal mismatches between the two datasets, some DHS data might exhibit a temporal lag compared to Open Buildings data. In future research, the DHS dataset offered a substantial amount of historical data. If Open Buildings provide building footprints extracted from high-definition satellite imagery at various historical time points, it would be feasible to establish panel regression at the city level. This approach would allow for the incorporation and consideration of temporal effects.

Multiple linear regression was used extensively in this study. A significant limitation of this

approach is the extent to which the assumptions of multiple linear regression are met. There exists potential contention regarding whether certain models satisfy one or several of the five core assumptions of multiple linear regression. For instance, Model 15, focuses on the cluster-level average years of education using the SBDS metric. To ensure the model adhered to the no multicollinearity assumption, a Variance Inflation Factor (VIF) check was conducted. After testing, we excluded variables with excessively high variance inflation factors, such as the total population count. Consequently, we believe that there is no multicollinearity among the model's parameters, although this perspective might be subject to debate. To verify the independence assumption, we employed the Durbin-Watson test to check for autocorrelation. The model's D-W Statistic was found to be 1.9, which being close to 2, indicates that the model does not exhibit autocorrelation and thus aligns with the assumption. Figure 6.1 displays the Residuals vs. Fitted, Normal Q-Q, Scale-Location, and Residuals vs. Leverage plots.

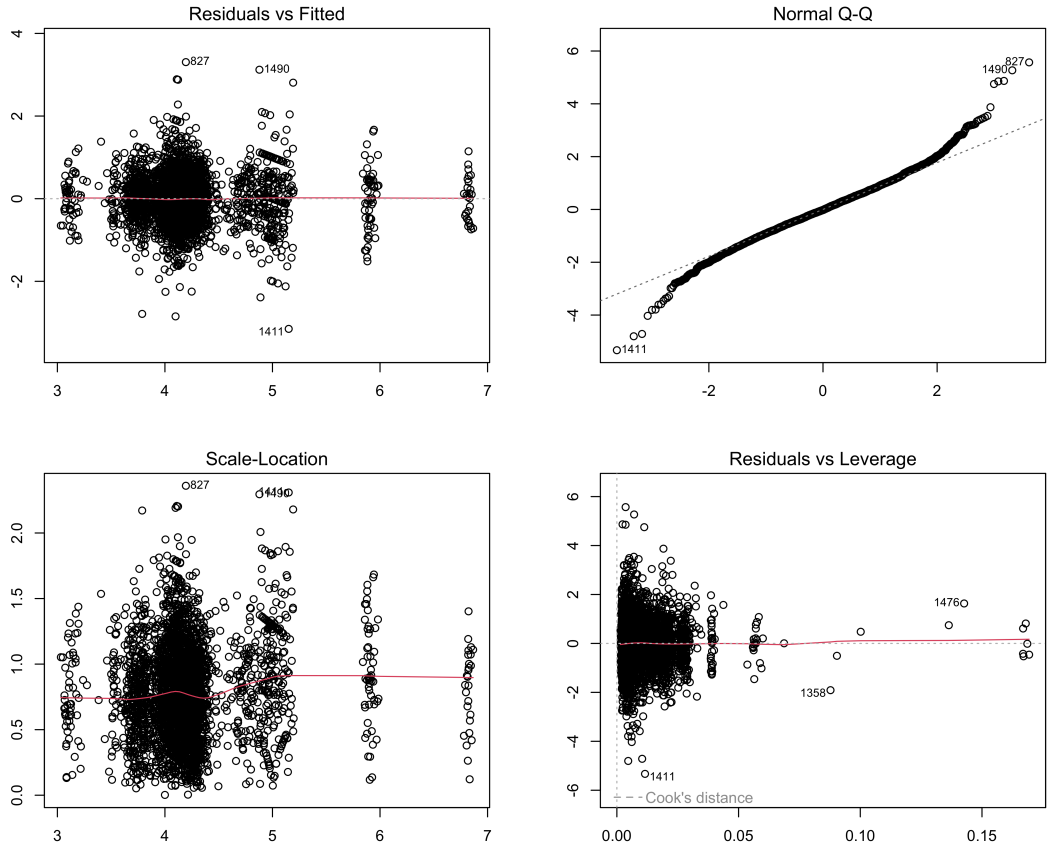

Figure 6.1: *R*egression Plots for Model 15

In the Residuals vs. Fitted plot, we examine the distribution of residuals across different fitted values to ascertain the linear relationship between predictor variables and the outcome variable. We observe that the residuals spread on either side of an almost horizontal red line, suggesting the absence of non-linear relationships, which tests the linearity assumption. Within the Normal Q-Q plot, we test the normality assumption, ensuring that the residuals predominantly lie along the diagonal line to confirm their normal distribution. The plot reveals that, except for a few extreme values at the tails, the majority conform to a normal distribution, leading us to conclude that the assumption is met. The Scale-Location plot is employed to test the homoscedasticity assumption, ensuring that residuals exhibit constant variance across different fitted values. In this plot, we note distinct clusters around an average education level of 4, and slightly below 6 and 7. This clustering might be attributed to varying educational policies across countries. For

instance, primary education stages in countries like India and Bangladesh last for 5 years, while in the DR Congo and the Philippines, they span 6 years. Consequently, we observe similar clustering in the fitted values. However, the residuals around 6 and 7 years are lower than those for 5 years and earlier. Given that the majority of the data is distributed around 5 years and earlier, we believe that homoscedasticity is achieved, though this perspective might be controversial. In the Residuals vs. Leverage plot, we aim to identify potential influential outliers that should be filtered to prevent undue influence on the results. The plot does not reveal any potential outliers beyond Cook's distance.

Recognizing the paramount importance of the foundational assumptions in multiple linear regression models, we meticulously employed rigorous statistical techniques to identify and address any latent discrepancies or anomalies. However, akin to the scenario presented in Model 15, other models utilized in this study also exist potential contention regarding the fulfillment of these assumptions, despite our unwavering commitment to ensuring model robustness. Nevertheless, the decision to employ multiple linear regression was not made lightly. It was primarily influenced by the model's intrinsic capability to yield interpretable results, which facilitates a comprehensive and meaningful quantitative analysis, enabling us to extract valuable insights and draw substantive conclusions from the data.

In future research endeavors, if the primary objective is to predict socio-economic data using remote sensing data, without delving into the relationship between architectural metrics and corresponding socio-economic indicators, one might consider the deployment of machine learning models known for superior predictive performance. Potential choices could encompass advanced decision tree models or neural network models. While the interpretation of these sophisticated models may lack tangible significance, they can effectively enhance predictive accuracy and adaptability, thereby facilitating the precise acquisition of pertinent socio-economic data through the approach proposed in this study.

7. CONCLUSION

To identify informal residential areas in major cities of developing countries, we processed building footprint data recognized through deep learning models based on high-resolution satellite imagery, as provided by Open Buildings. In our multifaceted approach, we employed a 500-meter grid overlay across 63 cities, calculating the average building area within each grid. This data was subsequently utilized to develop an interactive visualization platform, enabling users to explore potential informal settlements within these urban landscapes. Notably, the Small Building Density Scores (SBDS) layer, post-convolution with a Gaussian kernel, exhibited superior performance in delineating slums. By considering both the weighted area of individual structures and their proximate buildings, this method accentuates zones characterized by high-density, small-footprint constructions, effectively highlighting even minuscule aggregations. This technique demonstrates exceptional precision at the boundaries of informal constructions, ceasing abruptly near expansive roads or green spaces, thereby minimizing misclassifications. By calibrating various thresholds, we can generate accurate polygons of informal constructions based on the SBDS layer, furnishing invaluable data on informal urban residential zones for subsequent research endeavours.

Our analysis unveiled a robust correlation between architectural indicators and certain socio-economic metrics. Economically speaking, the Estimated Individual Product exhibited a positive correlation with the median building area surrounding respondents, while it inversely correlated with the standard deviation of the SBDS. Although we cannot definitively ascertain the direction of the effect or establish causality, the mere existence of this correlation bridges architectural and economic indicators, offering the potential to estimate regional economic conditions through architectural metrics. Intriguingly, within individual countries, the same indicators manifested varying degrees of correlation. This perhaps suggests that architectural characteristics, especially in major cities, are influenced by country-specific policies. Consequently, for a more precise prediction of local economic conditions, country-specific models might outperform a singular, globalized composite model in terms of accuracy. However, the explanatory power of such tailored models might be inferior to a holistic, global model.

At the educational echelon, the explanatory power provided by architectural indicators is markedly inferior to that of economic metrics. Decision tree models further elucidated that educational attainment is predominantly orchestrated by pertinent policies, exhibiting pronounced disparities across different nations. Notably, these national variations in education overshadow differences attributable to the architectural characteristics of respondents' residences. To circumvent the pitfalls of omitted variable bias, we employed a fixed-effect panel regression, leveraging country-specific variables to account for the heterogeneity in educational attainment induced by disparate policies. This adjustment facilitated a more accurate estimation of the effect based on architectural indicators. While individual variances in years of education are substantial, making it challenging to predict educational tenure at the individual level based solely on architectural metrics, these indicators exhibit discernible explanatory power at the cluster level. We discerned a positive correlation between an increase in average building area and years of education. Conversely, an augmentation in the standard deviation of building area and the SBDS correlates negatively with educational tenure, suggesting that heightened

architectural inequality within a region is concomitant with a reduced average educational span for its populace.

Regarding living conditions, our analysis discerned that architectural indicators offer substantive explanatory power for variables such as sources of drinking water, types of sanitation facilities, roofing materials, and the prevalence of various household appliances and transportation means. By deploying logistic regression models, we were able to investigate the proportion of specific types of living conditions within defined clusters. Our findings indicate that certain living conditions can be predicted with a commendable degree of accuracy. For instance, within the domain of drinking water sources, categories like bottled water and sachet water emerged as predictable. Similarly, sanitation facilities such as pit latrines and flush pit latrines, roofing materials like wood or rustic mats, and the prevalence of electricity and television ownership in households were also reliably forecasted. The robust correlation between these living conditions and architectural indicators ensures that the models exhibit superior fit and performance.

In summation, our research demonstrates the feasibility of exclusively employing remote sensing data, in conjunction with machine learning methodologies, to predict socio-economic metrics, as delineated in Figure 7.1. Initially, we harnessed the building footprints dataset, based on high-resolution satellite imagery, provided by Open Buildings, which encompasses the vast majority of developing countries. To extract richer architectural metrics, we applied a convolution using the Gaussian kernel to this dataset. Concurrently, we sourced remote sensing data from various repositories, capturing variables such as temperature, the global human footprint, and vegetation indices. Leveraging survey data from platforms like the DHS Survey and UNICEF MICS, we constructed a suite of models in tandem with the aforementioned remote sensing data. Post-validation, these models stand poised to predict socio-economic parameters. This approach significantly amplifies data accessibility, offering a cost-effective alternative to traditional surveys. While there might be a slight trade-off in terms of accuracy, this method might effectively bridge the data chasm prevalent in many developing nations.

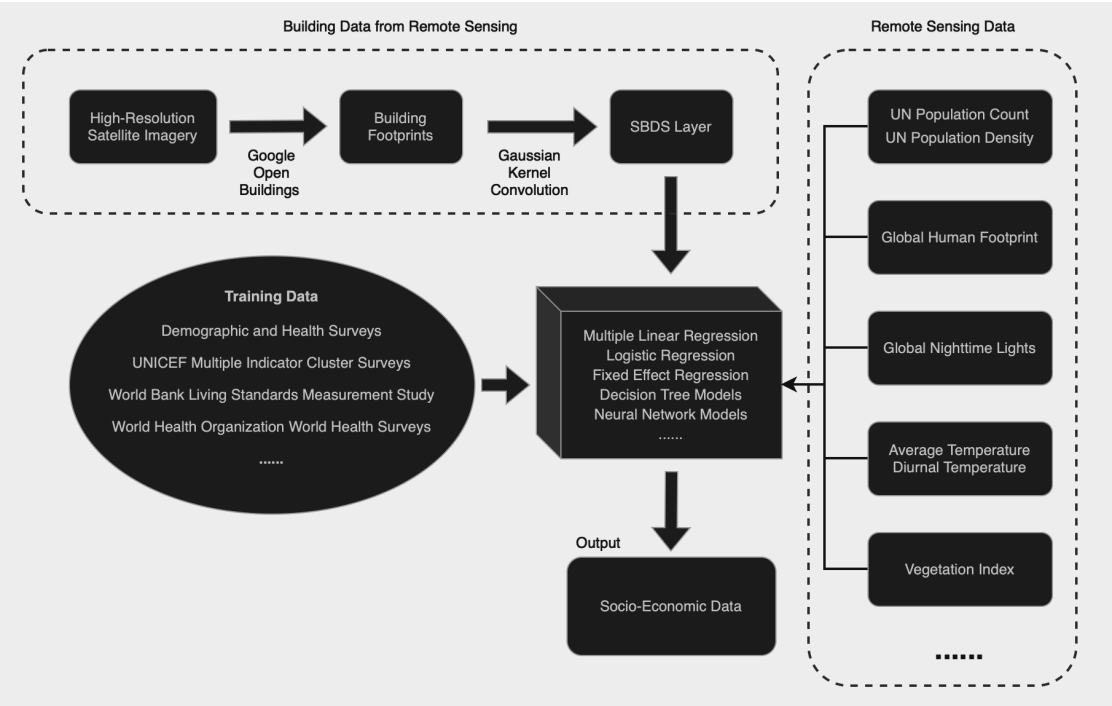

Figure 7.1: *F*low Chart for Acquiring Socio-Economic Data Using Remote Sensing Data

BIBLIOGRAPHY

Bakibinga, P., Kabaria, C., Kyobutungi, C., Manyara, A., Mbaya, N., Mohammed, S., Njeri, A., Azam, I., Iqbal, R., Mazaffar, S., Rizvi, N., Rizvi, T., Ur Rehman, H., Shifat Ahmed, S., Alam, O., Khan, A., Rahman, O., Yusuf, R., Odubanjo, D., Ayobola, M., Fayehun, F., Omigbodun, A., Owoaje, E., Taiwo, O., Diggle, P., Aujla, N., Chen, Y., Gill, P., Griffiths, F., Harris, B., Madan, J., Lilford, R., Oyobode, O., Pitidis, V., De Albequerque, J., Sartori, J., Taylor, C., Ulbrich, P., Uthman, O., Watson, S. and Yeboah, G. (2019), 'A protocol for a multi-site, spatially-referenced household survey in slum settings: Methods for access, sampling frame construction, sampling, and field data collection', *BMC Medical Research Methodology* **19**(1).

Ballinger, O. (n.d.).
URL: *https://ollielballinger.users.earthengine.app/view/ism*

Bird, J., Montebruno, P. and Regan, T. (n.d.), 'Life in a slum: understanding living conditions in nairobi's slums across time and space', **33**(3), 496–520. _eprint: https://academic.oup.com/oxrep/article-pdf/33/3/496/18169819/grx036.pdf.
URL: *https://doi.org/10.1093/oxrep/grx036*

Center for International Earth Science Information Network (2018), 'Gridded population of the world, version 4 (GPWv4): Population count adjusted to match 2015 revision of UN WPP country totals, revision 11'.
URL: *https://doi.org/10.7927/H4PN93PB*

DHS (2023), 'The dhs program'.
URL: *https://dhsprogram.com/Data/*

Diou, C., Lelekas, P. and Delopoulos, A. (2018), 'Image-based surrogates of socio-economic status in urban neighborhoods using deep multiple instance learning', *Journal of Imaging* **4**(11).
URL: *https://www.mdpi.com/2313-433X/4/11/125*

Duque, J. C., Patino, J. E. and Betancourt, A. (2017), 'Exploring the potential of machine learning for automatic slum identification from vhr imagery', *Remote Sensing* **9**(9).
URL: *https://www.mdpi.com/2072-4292/9/9/895*

Faisal, K. and Shaker, A. (2014), 'The use of remote sensing technique to predict gross domestic product (gdp): An analysis of built-up index and gdp in nine major cities in canada', *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **XL-7**, 85–92.
URL: *https://isprs-archives.copernicus.org/articles/XL-7/85/2014/*

Hecht, R., Meinel, G. and Buchroithner, M. (2015), 'Automatic identification of building types based on topographic databases – a comparison of different data sources', *International Journal of Cartography* **1**(1), 18–31.
URL: *https://doi.org/10.1080/23729333.2015.1055644*

Jia, C., Liu, Y., Du, Y., Huang, J. and Fei, T. (2021), 'Evaluation of urban vibrancy and its relationship with the economic landscape: A case study of beijing', *ISPRS International Journal of Geo-Information* **10**(2).
URL: *https://www.mdpi.com/2220-9964/10/2/72*

Mayala, B., Fish, T. D., Eitelberg, D. and Dontamsetti, T. (2018), 'The geospatial covariate datasets manual demographic and health surveys'.

NCEI (2019), 'Global nighttime lights annual and monthly composites using viirs day night band'.
URL: *https://ngdc.noaa.gov/eog/viirs*

Quinn, J. (2021), 'Mapping africa buildings with satellite imagery'.
URL: *https://blog.research.google/2021/07/mapping-africas-buildings-with.html*

Salvati, L. and Carlucci, M. (n.d.), 'Shaping dimensions of urban complexity: The role of economic structure and socio-demographic local contexts', **147**(1), 263–285.
URL: *https://doi.org/10.1007/s11205-019-02156-2*

Sirko, W., Kashubin, S., Ritter, M., Annkah, A., Bouchareb, Y. S. E., Dauphin, Y., Keysers, D., Neumann, M., Cisse, M. and Quinn, J. (2021), 'Continental scale building detection from high resolution satellite imagery'.

UN-Habitat (2003), 'Slums of the world: The face of urban poverty in the new millennium'.
URL: *https://unhabitat.org/slums-of-the-world-the-face-of-urban-poverty-in-the-new-millennium*

UN-Habitat (2022), 'Envisaging the future of cities world cities report 2022'.
URL: *https://unhabitat.org/wcr/*

United Nations (2023), *World Economic Situation and Prospects 2023*.
URL: *https://desapublications.un.org/publications/world-economic-situation-and-prospects-2023*

Warth, G., Braun, A., Assmann, O., Fleckenstein, K. and Hochschild, V. (2020), 'Prediction of socio-economic indicators for urban planning using vhr satellite imagery and spatial analysis', *Remote Sensing* **12**(11).
URL: *https://www.mdpi.com/2072-4292/12/11/1730*

Widayani, P. (n.d.), 'Aplikasi object-based image analysis untuk identifikasi awal permukiman kumuh menggunakan citra satelit worldview-2', **32**(2), 162.
URL: *https://jurnal.ugm.ac.id/mgi/article/view/32306*

Wildlife Conservation Society and Center for International Earth Science Information Network (2005), 'Last of the wild project, version 2 global human footprint dataset'.
URL: *https://doi.org/10.7927/H4M61H5F*

World Bank (2022), 'Gini index'.
URL: *https://data.worldbank.org/indicator/SI.POV.GINI*

World Bank (2023), 'Gdp per capita in current us dollar'.
URL: *https://data.worldbank.org/indicator/NY.GDP.PCAP.CD*

World Population Review (2023), 'World city populations 2023'.
URL: *https://worldpopulationreview.com/world-cities*

WorldPop (2023), 'Open spatial demographic data and research'.
URL: *https://www.worldpop.org/*

Wu, B., Yu, B., Yao, S., Wu, Q., Chen, Z. and Wu, J. (2019), 'A surface network based method for studying urban hierarchies by night time light remote sensing data', *International Journal of Geographical Information Science* **33**(7), 1377–1398.
URL: *https://doi.org/10.1080/13658816.2019.1585540*

Wurm, M., Weigand, M., Schmitt, A., Geiß, C. and Taubenböck, H. (2017), Exploitation of textural and morphological image features in sentinel-2a data for slum mapping, *in* '2017 Joint Urban Remote Sensing Event (JURSE)', pp. 1–4.

APPENDIX

Appendix 1: Lists of Meetings with Supervisor

Meeting 1: 2023 April 12, Online. Discussed building quantitative model to analysis the satellite image recognition model output, the supervisor suggested Google Open Building and recommended the Cloud-Based Google Earth Engine tools.

Meeting 2: 2023 May 24, CASA Office. Discussed the use of Open Buildings data, the data processing methods using Google Earth Engine, and the choice of city. Set the timeline afterward. The supervisor recommends the DHS data as a dependent variable.

Meeting 3 & 4: 2023 July 17, CASA Office. Demonstrated visualization of gridded data, and discussed next steps forward, supervisor suggested using a 2km buffer. Discussed merging DHS survey data with spatial data, considering effects at different levels, and model selection.

Meeting 5: 2023 August 7, CASA Office. Discussed the design of the variables, the final model selection and the preliminary results of the model, and the supervisor suggested the additional use of Gaussian Kernel outputs as metrics and comparing metrics differences.

Meeting 6: 2023 August 22, Online. The supervisor gave some feedback based on dissertation drafts and discussed some potential revisions.

Appendix 2 Reproducible Codes

Reproducible SAS, R, Google Earth Engine codes for this study can be found in the GitHub Repository. Part of the data can be found in the data folder, but DHS data is not allowed to be shared publicly, but can be requested, details are in the README file.

Appendix 3 List of Cities in the Study

Region	Country	City
Central & Southern Africa	Angola	Luanda
	Cameroon	Douala
		Yaounde
	DR Congo	Kinshasa
		Lubumbashi
		Mbuji-Mayi
	South Africa	Cape Town
		Durban
		Ekurhuleni
		Johannesburg

Appendix 3 – Continued

Region	Country	City
		Pretoria
Latin America	Colombia	Barranquilla
		Bogota
		Cali
		Medellin
	Dominican Republic	Santo Domingo
	Guatemala	Guatemala City
	Haiti	Port-au-Prince
	Peru	Lima
Nothern & Eastern Africa	Egypt	Alexandria
		Cairo
	Ethiopia	Addis Ababa
	Kenya	Nairobi
	Madagascar	Antananarivo
	Tanzania	Dar es Salaam
	Uganda	Kampala
	Zambia	Lusaka
South Asia	Bangladesh	Chittagong
		Dhaka
	India	Ahmedabad
		Bangalore
		Chennai
		Coimbatore
		Delhi
		Hyderabad
		Indore
		Jaipur
		Kanpur
		Kochi
		Kolkata
		Kozhikode
		Lucknow
		Malappuram
		Mumbai
		Nagpur
		Pune
		Surat
		Thrissur
	Pakistan	Faisalabad
		Gujranwala
		Karachi
		Lahore
		Multan
		Peshawar
		Rawalpindi
Southeast Asia	Cambodia	Phnom Penh
	Philippines	Manila
Western Africa	Burkina Faso	Ouagadougou

Appendix 3 – Continued

Region	Country	City
	Ghana	Accra
		Kumasi
	Guinea	Conakry
	Ivory Coast	Abidjan
	Nigeria	Abuja
		Ibadan
		Kano
		Lagos
		Port Harcourt
	Senegal	Dakar

Appendix 4 List of DHS data

Region	Country	City
Central & Southern Africa	Angola	Angola: Standard DHS, 2015-16
	Cameroon	Cameroon: Standard DHS, 2018
	DR Congo	DR Congo: Standard DHS, 2013-14
	South Africa	South Africa: Standard DHS, 2016
Latin America	Colombia	Colombia: Standard DHS, 2010
	Dominican Republic	Dominican Republic: Standard DHS, 2013
	Guatemala	Guatemala: Standard DHS, 2014-15
	Haiti	Haiti: Standard DHS, 2016-17
	Peru	Peru: Continuous DHS, 2009
Nothern & Eastern Africa	Egypt	Egypt: Standard DHS, 2014
	Ethiopia	Ethiopia: Standard DHS, 2016
	Kenya	Kenya: Standard DHS, 2022
	Madagascar	Madagascar: Standard DHS, 2021
	Tanzania	Tanzania: Standard DHS, 2015-16
	Uganda	Uganda: Standard DHS, 2016
	Zambia	Zambia: Standard DHS, 2018
South Asia	Bangladesh	Bangladesh: Standard DHS, 2017-18
	India	India: Standard DHS, 2019-21
	Pakistan	Pakistan: Standard DHS, 2017-18
Southeast Asia	Cambodia	Cambodia: Standard DHS, 2021-22
	Philippines	Philippines: Standard DHS, 2022
Western Africa	Burkina Faso	Burkina Faso: Standard DHS, 2010
	Ghana	Ghana: Standard DHS, 2014
	Guinea	Guinea: Standard DHS, 2018
	Ivory Coast	Cote d’Ivoire: Standard DHS, 2011-12
	Nigeria	Nigeria: Standard DHS, 2018
	Senegal	Senegal: Continuous DHS, 2019