Universidad del Rosario | Facultad de Economía

SERIE
DOCUMENTOS
DE TRABAJO

No. 314

Febrero de 2024

Nightlight, landcover and buildings: understanding intracity socioeconomic differences

Andres García-Suaza

Daniela Varela

Universidad del Rosario, Facultad de Economía
Cl. 12c #4-59. Bogotá, D. C., Colombia
PBX: (031) 2970200 ext. 4163

facultadeconomia@urosario.edu.co
https://www.urosario.edu.co/Facultad-de-Economia/Inicio/
https://repository.urosario.edu.co/handle/10336/20291

Nightlight, landcover and buildings: understanding intracity socioeconomic differences

Andres García-Suaza* Daniela Varela†

February 12, 2024

Abstract

Monitoring patterns of segregation and inequality at small-area geographic levels is extremely costly. However, the increased availability of data through non-traditional sources such as satellite imagery facilitates this task. This paper assess the relevance of data from nightlight and day-time satellite imagery as well as building footprints and localization of points of interest for mapping variability in socio-economic outcomes, i.e., household income, labor formality, life quality perception and household informality. The outcomes are computed at a granular level by combining census data, survey data, and small area estimation. The results reveal that non-traditional sources are important to predict spatial differences socio-economic outcomes. Furthermore, the combination of all sources creates complementarities that enable a more accurate spatial distribution of the studied variables.

JEL Classification: R12, E26, C21.

Keywords: Remote sensing, Satellite imagery, nightlights, points of interest, spatial segregation, urban footprints, informal housing.

*School of Economics, Universidad del Rosario. Email: andres.garcia@urosario.edu.co.

†Researcher at Universidad del Rosario. Email: danielavarelatabares@gmail.com.

This working paper is funded by the Colombia Científica-Alianza EFI Research Program, with code 60185 and contract number FP44842-220-2018, funded by the World Bank through the call Scientific Ecosystems, managed by the Colombian Ministry of Science, Technology, and Innovation. We also thank Juan Carlos Duque for providing data for our statistical analysis. Errors, opinions, and omissions are ours and do not compromise the institutions.

1 Introduction

Understanding the patterns and trends in the spatial distribution of socioeconomic conditions is essential for policymakers to allocate resources efficiently and develop sustainable cities. The availability of accurate and more disaggregated information on Household (HH) socioeconomic conditions is essential for this matter. Typically, official data sources provide information relevant to geographic units of an aggregated nature, such as municipalities, but data regarding intra-city spatial variations is still limited. Household surveys, for instance, generally offer a representative picture on an aggregate city level, whereas more granular analyses would require censuses that are far less frequent and more expensive to carry out (Duque et al., 2013; Kilic et al., 2017; Fujii and van der Weide, 2020). This is especially true for cities in low and medium-income countries (LMICs) where detailed information is lacking or is collected less frequently, several countries only conduct a few nationwide surveys to estimate poverty every ten years, leading them to depend on outdated figures (Chen and Nordhaus, 2011; Addison and Stewart, 2015; Abascal et al., 2022; Jean et al., 2016; Marty and Duhaut, 2024).

This issue has been addressed through to the increased availability of remote sensing and geospatial data (hereafter named as Remote Sensing Data, RSD), e.g. satellite imagery, by delivering proxy measures of multiple economic variables (see Kuffer et al., 2017). RDS present advantages for economic analysis since it facilitates access to scarce information, with greater spatial disaggregation and greater geographic coverage (Donaldson and Storeygard, 2016). Additionally, it is less costly to obtain information, with less lag time and facilitates comparisons between units and time periods (see Gibson and Boe-Gibson, 2021). As RSD have improved their ability to capture richer information, multiple uses of these data have emerged. For example, luminosity data from nighttime satellite images has been widely used to estimate economic activity level, poverty rate, population density, urban sprawl, among others (see Henderson et al., 2012; Keola et al., 2015; Kuffer et al., 2017; Gibson et al., 2021; Baragwanath et al., 2021; Pérez-Sindín et al., 2021).

Other sources such as daytime satellite imagery, urban footprints, points of interest, and many more have been used to compute geospatial indicators that capture relevant and somehow complimentary information (see Keola et al., 2015; Baragwanath et al., 2021; Kohli et al., 2016; Puttanapong et al., 2020). For instance, an extensive literature has shown that RSD is able to generate similar patterns than census for poverty maps (see Jean et al., 2016; Watmough et al., 2016; Lee and Braithwaite, 2022; van der Weide et al., 2023). The combination of these sources of information has the potential to study the spatial distribution of economic and social issues (see Saiz and Salazar Miranda, 2023a, for a comprenhensive overview of spatial data sources used in urban science and urban economics). Therefore, the association between variables from RSD and variables with economic relevance is a powerful tool for cost-effective monitoring of complex phenomena.

Numerous works have explored the relationship between remote sensing variables and official information sources at the regional or municipal level, yet relatively less attention has been paid to intra-city differences. This paper assesses the relationship between relevant economic and social outcomes in urban studies measured with official census and survey data, alongside variables constructed with RSD for the case of Medellin, Colombia.

In particular, the predictive ability of data from Nighttime Lights data (NLT), Daytime Satellite Imagery (DSI), Global Urban Foorprints (GUF), Points of Interest (POI), and Public Transport (PT) data is studied for understanding variables such as income level, job formality, informal housing and subjective wellbeing index. Our approach is close to Engstrom et al. (2022) and van der Weide et al. (2023) who have explored the effectiveness of RSD in reproducing official statistics.

While our study focuses on data at the census block level, certain outcomes cannot be directly observed at this level of detail. Therefore, we employed a two-stage methodology to address this limitation. Initially, Small Areas Estimation (SAE) techniques were used to link data from a representative household survey (the Quality of Life Survey, QLS) with census data. Then, regression models with variable selection, particularly LASSO, are used with the aim of predicting socioeconomic outcomes. The results show that the different sources of RSD are complementary to understand intra-city spatial differences. Our analysis applies to data for Medellin (Colombia), a city characterized by high levels of segregation and socioeconomic contrasts throughout the city. This is an interesting case to illustrate the ability of RDS to map spatial differences in socioeconomic outcomes and the implementation of SAE to increase survey resolution. Although it is complex to ensure that the estimated models can be extrapolated to other contexts, as discussed by Engstrom et al. (2022), it allows for providing evidence on the factors that determine crucial variables for the design of urban policies such as housing, transportation, employment, among others, in developing countries.

This study is related to recent research such as Kuffer et al. (2017) and Che and Gamba (2019) that analyze the ability of NLT data and land cover data to characterize intra-urban areas. Kuffer et al. (2017) evaluate and verify the capability of International Space Station (ISS) nightlight images compared to other datasets with lower resolution, such as DMSP-OLS, in analyzing intra-urban disparities and determining to what extent these are correlated with socio-economic characteristics (e.g., poverty) in the context of deprived areas (slums), that correspond to urban spaces little studied due to the scarcity of information. The findings emphasize the significance of contextual knowledge specific to each city for accurate interpretation, as the correlation coefficients between built-up areas, population density, and nightlight values vary across different cities. Che and Gamba (2019) propose a procedure to detect intra-urban changes in five different urban areas by jointly exploiting Sentinel-1 (S-1) SAR data and nighttime light data, and the experimental results, concluding that the detected patterns actually correspond to peculiar changes in the structure of urban blocks.

A vast part of studies using RSD have focused on the construction of alternative measures of GDP, particularly through the utilization of nightlight data. Although the association between NLT data and GDP has been extensively explored, it has revealed certain limitations that have been partially addressed through improved data availability and advancements in machine learning (see e.g., Gibson et al., 2021; Baragwanath et al., 2021). However, further studies have revealed that the relationship between NLT and GDP is context-dependent (Kuffer et al., 2017), e.g., in rural areas or in the presence of informality this relationship may be noisier (Rangel-Gonzalez and Llamosas-Rosas, 2019; Gibson et al., 2021; Pérez-Sindín et al., 2021). Similarly, Doll et al. (2006) reports that DSI correlates with regional GDP at different scales; while, Sutton and Costanza (2002) combines NSI and land cover to estimate economic activity at high spatial resolution (see

also Mellander et al., 2015).

Beyond economic activity, RSD is relevant to study other economic variables. For instance, Elvidge et al. (1997) and Lin and Shi (2020) analyze the relationship of NLT with energy consumption, while Li et al. (2020) analyze housing prices. Addison and Stewart (2015) include manufacturing value added, total population, urban population, among others; and Charris et al. (2019); Sherman et al. (2023) consider more complex target variables such as the human development index. Burchfield et al. (2006); Ch et al. (2021); Li et al. (2021) study urban sprawl and city size.

In addition to NLT, information from GUF has also been exploited to determine the presence of informal settlements and the growth of deprived urban areas (Durst et al., 2021; Abascal et al., 2022). These studies examine building morphology metrics to characterize size and shape, aiming to identify patterns related to different settlement types. (see also Jochem and Tatem, 2021). On the other hand, POI data have emerged as valuable assets for studying urban and regional economies. In this line, Jiang et al. (2015); Niu and Silva (2021) investigate urban land use, Ganter et al. (2022) demonstrate that POI data can effectively model urban inequality with a high level of granularity and accuracy. Finally, transportation data has been widely used to explain employment determinants and the spatial distribution of economic activities (see also Redding and Turner, 2015; Baum-Snow and Turner, 2017; Posada and García-Suaza, 2022; Akbar et al., 2023a).

In summary, existing literature shows that various sources of RSD are informative for predicting intra-city spatial patterns. However, it seems relevant to determine in which context it is most useful to use such sources. This has practical effects for the study of specific outcomes and, from a research perspective, offers measurement alternatives through proxy variables when the outcome cannot be observed accurately. By utilizing a diverse range of data sources for a common objective, it becomes possible to overcome the limitations inherent in each individual source. This approach leads to a more precise and comprehensive understanding of the socioeconomic conditions within a specific area. For instance, in contrast to nightlight data, which typically has lower resolution and is easily affected by noise and saturation, daytime images incorporate more features and can unveil intricate topographic details with greater precision (see Liu et al., 2021; Goldblatt et al., 2007). Alternatively, footprint data provides valuable insights into urban density measures that are crucial for analyzing urban patterns and dynamics. And, POI and PT complements this information by offering additional insights into the availability and allocation of amenities.

Our results suggest that the interaction between novel sources of information, such as RSD, and statistical models is effective for analyzing outcomes at a high level of spatial resolution. First, the use of SAE allows for obtaining a highly granular map of the variables of interest. Second, the use of RSD provides informative insights into important outcomes for the design of urban policies, which is especially advantageous in situations where the availability of census or survey data is limited. Specifically, it is found that variables from various sources have predictive ability for outcomes related to income, labor market, housing, and subjective well-being. This indicates that it is possible to monitor different dimensions of the socioeconomic status of a city by combining different sources of RSD.

In fact, when analyzing the sensitivity in the performance of predictive models, it is inferred that the data sources have a certain level of complementarity, which encourages the use of all sources. Interestingly, the feature selection algorithms result in a combination that includes variables from different sources. Although the information on transportation seems to have a lower contribution, in aggregate, it helps to understand the different outcomes. Our findings also support recent literature showing that, e.g., nighttime lights data requires additional information to enhance proxies for poverty or economic activity. This source of information is relevant but not the most relatively important. On the other hand, urban footprints show high relevance in predicting the outcomes of interest. This is expected in the case of housing informality, which is related to the city's shape and buildings, but in our case, it is shown as an important source for understanding spatial differences in other socioeconomic variables. These results contribute to a better understanding of urban growth, informality and quality of life at the intra-urban level.

The rest of the paper is organized as follows. Section 2 describes the sources of information and methods used both for measuring outcomes and and for variable selection. Section 3 presents the results of estimating variables of interest using SAE. Section 4 discusses the results of the variable selection process, and Section 5 presents concluding remarks.

2 Data and Methods

Medellin City is located in the northwest region of Colombia. It is also considered the second largest city in the country. The city has been characterized by high urban growth with a population of 2.4 million inhabitants[1]. Despite the notorious progress in recent decades, the poverty rate remains high. Regarding official statistics from DANE,[2] 27.6% of households were under poverty conditions in 2021. This is accompanied by high levels of quality housing deficits (or informal housing, 17.9%) and unemployment rates above the national average (Posada et al., 2022).

In such a context, it is crucial to constantly monitor the living conditions of households, as well as to understand whether the prevalence of these problems is homogeneous across the city. For this purpose, traditional sources of information such as censuses and surveys are combined with other non-traditional sources coming from RSD with two purposes: to obtain a granular snapshot and to identify easily captured variables that allow for tracking intra-city differences over time. That is, censuses and surveys allow for a comprehensive measurement of outcomes of interest, while RSD contain signals that relate to these outcomes and allow for the prediction of spatial differences.

Since there are missing relevant economic outcomes in the census that can be found in the survey, SAE was implemented to estimate them at a more granular level. Regarding RSD, we use NLT, DSI, POI, GUF, and PT data. The outcomes of interest are income

[1]Data according to results of the national Census 2018 at https://www.dane.gov.co/files/censo2018/informacion-tecnica/presentaciones-territorio/190709-CNPV-presentacion-medellin.pdf

[2]Data taken from https://www.dane.gov.co/files/investigaciones/condiciones_vida/pobreza/2021/Presentacion-pobreza-monetaria_2021.pdf

level, formal employment, perception of quality of life and housing quality as a proxy of informal housing. While the latter is computed using census data at the census block level, the other three are not directly observable in the Census and so are estimated using SAE. These variables together measure different dimensions of quality of life and allow inferences about poverty conditions. These outcomes are estimated using information for 2018.

2.1 Census and surveys

Our first data source is the Colombian National Census of 2018,[3] which was run by the Colombian National Administrative Department of Statistics (DANE). In the case of Medellin, data of individuals and households is georeferenced at census block level for a total of 14,327 units. The census contains information on housing characteristics (e.g., access to utilities and housing materials), sociodemographic variables such as age, sex, educational level, and general information on the employment status at the individual level. This source is exploited in three ways. First, it is used to compute a housing quality index, a proxy of informal housing, following the definition of informal housing discussed in UN-Habitat (2003). In particular, a housing unit is classified as informal if it exhibits at least one of the following characteristics: overcrowding, lack of adequate water and sewage connections, or the use of precarious building materials. Second, it is used to determine the unit of analysis which consists of census blocks. Finally, since the census does not report important variables such as income, type of work (formal or informal) or subjective welfare variables, they are used as input for the estimation of these outcomes by SAE.

On the other hand, outcomes are observed in the Quality of Life Survey[4] (QLS) of Medellin. This survey enables the tracking and measurement of the socio-economic conditions of residents in the 16 communes and 5 districts that make up the municipality of Medellín. As for 2018, the survey sample includes 9,228 households corresponding to 30,941 individuals and 295 neighborhoods, providing indicators on demographic and housing characteristics, education, health and social security, labor market status, citizen perception, among others. Three outcomes are observed in the QLS, namely: household income, formal job status and quality life perception. Household income is the total income of the household estimated in per-capita terms. In turn, labor formality is determined by workers reporting contributions to pension systems. Lastly, the quality life perception is constructed based on responses where household heads are asked to rate various aspects aspects including noise levels, waste collection system, visual contamination, and neighborhood tree planting. If any of those factors were perceived as low for the household head, the positive perception is deemed as zero.

[3]For details, see https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/censo-nacional-de-poblacion-y-vivenda-2018

[4]Data is available at https://www.medellin.gov.co/irj/portal/medellin?NavigationTarget=contenido/6916-Encuesta-Calidad-de-Vida-2018

2.2 Non-traditional data sources

Non-traditional data sources refer to databases that are usually unstructured and are not collected for the purpose of generating official statistics. But they can be transformed into structured data to measure physical characteristics or to obtain signals that relate to social and economic variables. Among these type of data, RSD is of particular interest because it facilitates obtaining information of a geographic space with high resolution. For our purposes, we explore NLT, DSI, GUF, POI, and PT data. These sources of information are a powerful source to respond to the increased demand for spatially disaggregated statistics, reducing reliance on surveys (see also Struijs et al., 2014).

Our first piece of information corresponds to NLT data that was obtained from the annual composite found in Earth Observation Group (EOG) website.[5] The image was collected by the Visible and Infrared Imaging Suite (VIIRS) sensor and subsequently processed from monthly cloud-free average radiance grids spanning 2012 to 2020. These composites cover virtually the entire globe at a resolution of 15 arc seconds ($\approx 500m^2$ at the Equator). We used the so called *average-masked* global image for 2018, from the NPP satellite (VIIRS Cloud Mask - Stray Light Removed) configuration. To obtain the MEAN-VIIRS variable, we extracted values from the raster file for all pixels within each block and calculated the average of these values.

In turn, DSI was extracted from Sentinel-2, a wide-swath, high-resolution, multi-spectral imaging mission.[6] Sentinel-2 carries the Multispectral Imager (MSI) that delivers 13 spectral bands ranging from 10 to 60-meter pixel size. We calculated the Normalized Difference Vegetation Index (NDVI) for quantifying green vegetation, which is defined as NDVI = (NIR - Red) / (NIR + Red), where Red and NIR stand for the spectral reflectance measurements acquired in the red (visible) and near-infrared regions, respectively (Observatory, 2000). This variable ranges from -1 to +1. Negative values correspond to water bodies; values close to zero (0.1) to areas of rock, sand or snow; low positive values (0.2 to 0.5) represent sparse vegetation; and high values (0.6 to 0.9) indicate dense vegetation. Similar to the nightlights data, we extracted the variable MEAN-NDVI for each block. We also included the Normalized Difference Built-up Index (NDBI) which uses the NIR and SWIR bands. This index is calculated by subtracting the near-infrared (NIR) band from the shortwave infrared (SWIR) band and dividing it by their sum. The values range from -1 to 1; negative values represent water bodies, low values reflect the presence of vegetation, and higher values indicate a higher probability of built-up areas or urbanization (Zha et al., 2003; Xu, 2007). These indices have proven to be useful in different fields in economics (see e.g. Galdo et al., 2021; Wang and Peng, 2021; Pan et al., 2022; Tang et al., 2022).

For POI, we consider establishments level location data collected from Google Place using Google Maps queries. This data correspond to the points of interest for Medellin in Akbar et al. (2023b). It was captured by searching keywords according to the following categories: work, school, exercise, transport, childcare, adultcare, goods, services, meals,

[5]Data is available at https://eogdata.mines.edu/products/vnl/

[6]It comprises a constellation of two polar-orbiting satellites placed in the same sun-synchronous orbit, phased at 180° to each other, that allow repeated surveys every 5 days. See for details https://sentinel.esa.int/documents/247904/1848117/Sentinel-2_Data_Products_and_Access

errands, recreation, healthcare, religion and other. Further details of the methodology can be found in Akbar et al. (2023a). In this study, we transformed the establishment data using an overlying operation of the establishment point's layer with the Medellin's census shapefile. And, by calculating the number of sites per block, we obtained the following variables: total number of sites, and sites related to work, transport and services, social, and care. Furthermore, we included three indicators proposed by Niu and Silva (2021) to examine zoning effects when aggregating POIs: POI density (number of POIs per square meter), POI class richness (number of POI classes present in a geographical unit), and POI class diversity (Shannon's Diversity Index).

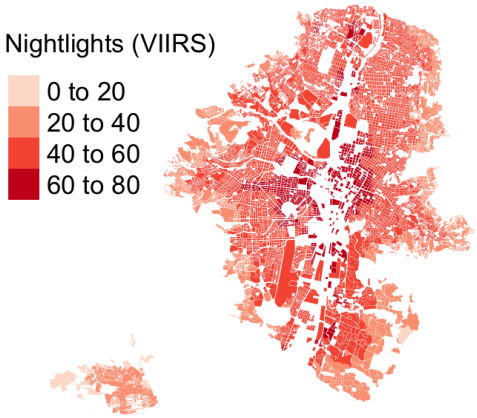
GUF is obtained from Bing Maps.[7] It was generated from imagery collected between 2020 and 2021, with subsequent Semantic Segmentation and Polygonization. We conducted block-level aggregation by computing the mean and the standard deviation for the following measures: area, perimeter, roundness, compactness, angle of rotation, and nearest neighbor distance (see Jochem and Tatem, 2021). Lastly, the PT data refers to the stations or stops of public transport, constructed based on records from the Secretaria de Movilidad de Medellin.[8] The features included from this source involved the minimum distance as a meaningful metric. Additionally, the number of transport points within a 500-meter radius from the centroid of each block.

Figure 1 presents some variables from the non-traditional sources. The results reveal two important facts. First, this variables show the presence remarkable spatial differences. For example, it is observed that nighttime luminosity is concentrated around the center of the city, while POIs are more dispersed with some concentration in the southern part of the city, which is characterized by higher income. Secondly, although there is some similarity, the variables exhibit dissimilar behavior that represents a certain level of complementarity between the sources.

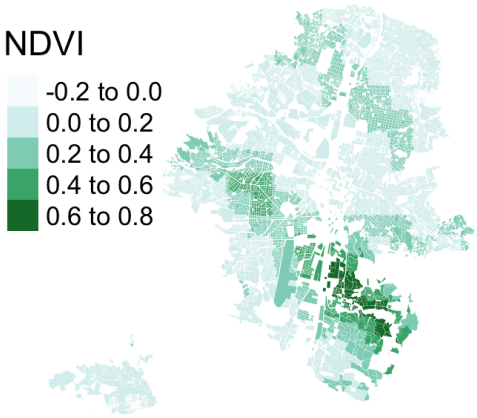
[7]Data is part of Microsoft Open source projects. Data is available at https://github.com/microsoft/SouthAmericaBuildingFootprints

[8]Data is available at https://www.medellin.gov.co/geomedellin/datosAbiertos/382

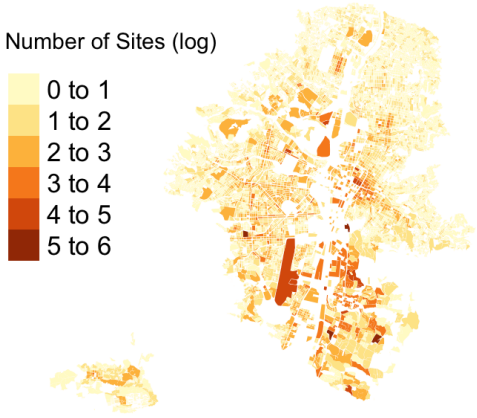
Figure 1: Spatial differences in characteristics from non traditional data sources



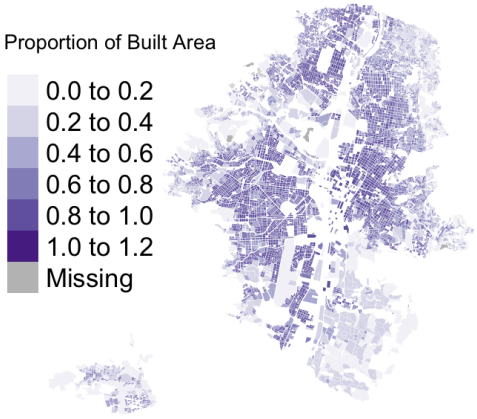
(a) VIIRS mean value by block.



(b) NDVI mean value by block.



(c) Logarithm of the total number of sites within each block.



(d) Footprint density: proportion of the block area that is built.

Source: Authors' calculations.

2.3 Small areas estimation

Although the census does not present problems of representativeness, the outcomes that can be studied are limited, in particular to study aspects related to poverty, since it lacks information on household income. In contrast, the QLS survey contains a larger number of variables, but its representativeness is limited to the comuna level, a larger administrative spatial unit than the neighborhoods and the census blocks. Consequently, these units are too aggregated to enable the effective examination of spatial patterns in intra-urban dynamics (Duque et al., 2013). Therefore, to leverage the richness of the survey data but in a more granular setting, we employed a SAE method by integrating both datasets. These methods are statistical techniques for improving the level of granularity of a data source without necessarily collecting additional data in the field (Rao and Molina, 2015; Asian Development Bank, 2020).

We used a unit-level model-based approach to link survey data, available for only

a sample of the target population, with census data that are available for the entire population (Kreutzmann et al., 2019). Specifically, we used the Elbers, Lanjouw and Lanjouw, that uses a nested-error model (Elbers et al., 2003). This method has been widely used to obtain small area estimates of poverty in many geographic contexts (see Molina and Rao, 2010; Elbers and van der Weide, 2014; van der Weide et al., 2023, for more details). The methodology starts by gathering the two datasets: the census dataset containing a set of covariates for all units in the population and the survey dataset containing both the variable of interest and the same set of covariates.

The basic unit-level model has a nested structure (nested error linear regression model) based on a Linear Mixed Model specification (Battese et al., 1988) as follows:

$$y_{ij} = x_{ij}^T \beta + \mu_i + \epsilon_{ij} \tag{1}$$

Where y_{ij} is the variable of interest for the j^{th} unit (individual/household HH) in the i^{th} geographical domain/area with $j = 1, ..., N$ and $i = 1, ..., D$. x_{ij} is a set of observable characteristics at the household level. And, μ_i is a area-specific parameter, while ϵ_{ij} to a individual level random errors. In practice, the model is fitted with the survey data, which results into estimates of the model parameters $\hat{\beta}$, $\hat{\sigma}_\mu^2$, $\hat{\sigma}_e^2$. The latter terms correspond to the estimated variances for μ_i and ϵ_{ij}, respectively. The underlying distributions are generated from the empirical distributions of the residuals, and are added to each estimated fitted value in the population (census) microdata in order to simulate the target variable. For this simulation, bootstrapping is used to generate B unit-level predictions each size N. Simulated outcome is obtained using the following expression:

$$y_{ij}^* = x_{ij}^T \hat{\beta} + \mu_i^* + \epsilon_{ij}^* \tag{2}$$

Finally, y_{ij}^* can be used to compute any area-specific target parameter. In our case, during each iteration of the bootstrapping, y_{ij}^* is aggregated as the mean by blocks and the average from the $B = 500$ repetitions is reported for each block. Taking into account the difference in the geographic unit between census and QLS, an assumption of how census blocks and neighborhoods relate to each other is required. In particular, it is assumed that the term μ_i is common to all census blocks located in the same neighborhood.

Our specifications consider a set of covariates typically used in the literature to explain household income, formal employment, and positive life quality perception. In particular, we considered socio-demographic variables including gender, age, marital status, and disability condition, along with educational and occupational aspects such as higher education level, employment status, education enrollment, and literacy status (see Furnham and Cheng, 2018; Terol-Cantero et al., 2023). Responses provided by the head of the household (HoH) were taken into account. Additionally, we also include potential predictors that were not directly associated with individuals but rather with households and physical infrastructure. These factors encompassed the dependency ratio, crowding index, housing informality index, as well as the availability of internet and gas services. The dependency ratio is a demographic measure that compares the size of the dependent population (usually children and the elderly) to the working-age population, and the crowding index is a continuous measure used to assess the level of overcrowding within a dwelling

by quantifying the number of people residing in relation to the number of bedrooms. The housing informality index was previously described in Section 2.

2.4 Feature selection in the outcomes prediction

To evaluate the importance of alternative data sources as significant inputs in mapping socioeconomic outcomes, regression models were conducted and the predictive capacity of the variables originated from each source was reported. We used Least Absolute Shrinkage and Selection Operator (LASSO) regression (Tibshirani, 1996), in the context of variable selection. This regularisation technique helps reduce the complexity of the model by shrinking the coefficients of less important variables to zero. It is particularly useful when dealing with a large number of variables, as it can help to identify the most important variables for the model. The restriction on the number of included covariates is carried out by means of adding a penalty factor to the classical ordinary least squares (OLS) problem (see Tibshirani, 1996; Freijeiro-González et al., 2022).

The selection of the penalty parameter λ, crucial for this process, is tuned through cross-validation. The conventional LASSO minimizes the following equation:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j| \tag{3}$$

where RSS refers to the Residual Sum of Squares. Similarly, the Adaptive LASSO represents an extension that addresses overfitting by penalizing large coefficients. In this case, the corresponding equation is expressed as follows:

$$RSS + \lambda\sum_{j=1}^{p}\hat{\omega}_j|\beta_j| \tag{4}$$

where $\hat{\omega}_j$ is called adaptive weights vector which is specific for each parameter and is defined as $1/(|\hat{\beta}_j^{ini}|)^{\gamma}$, with $\gamma = 1$ and $\hat{\beta}_j^{ini}$ the initial estimates for β_j. It is worth mentioning that lower levels of λ imply a lower penalty, and therefore a less restrictive selection. When λ is equal to zero, then the model becomes the Ordinary Least Squares regression. Besides, some of the coefficient estimates are to be exactly equal to zero when the tuning parameter λ is sufficiently large.

The primary objective behind LASSO is to identify the most robustly significant predictors in determining the outcomes of interest. Consequently, our analysis aims to compare whether predictors are useful across all outcomes. This enables us to gain valuable insights into the underlying patterns and drivers of the observed spatial variations. Finally, in order to consider alternative variable selection algorithms we also implemented Adaptive LASSO (Zou, 2006). This selection algorithm extends the LASSO regularization term by incorporating the magnitude of the regression coefficients contained using OLS. This modification allows to address the selection problem when predictors are highly

correlated. We implemented these algorithms using a total of 40 predictors which are described in Table 6.

3 Estimating the main outcomes

Our first set of results corresponds to the implementation of SAE to compute the outcomes of interest. Specifically, linear mixed models are estimated for household income, labor formality and quality of life perception using survey data and then each variable is simulated at the census block level. To provide some insights on the expected results of the relationship between control variables and outcomes, Table 1 presents the average of each variable within samples divided by the median income and groups defined under low and high perception of quality of life as well as formal and informal employment. Regarding income, higher income levels are associated with a lower proportion of youth, women, and higher education levels. From a household perspective, it is observed that access to services such as gas and internet, higher housing quality associated with the informality index, and overcrowding are related to higher income levels. In general, these findings are intuitive.

These same associations are observed in the case of the quality of life perception, with less pronounced variations in the proportion of head of household with higher education or households with access to gas and internet. In turn, in the case of type of job (formal and informal) it is observed a strong relation with educational level and marital status.

Tabla 1: Descriptive statistics of covariates.

	Income		Labor		Quality of life perception	
	Low	High	Informal	Formal	No positive	Positive
Age (Young, <29)^p	0.08	0.06	0.44	0.25	0.07	0.06
Gender (Female)^p	0.51	0.46	0.59	0.42	0.49	0.46
Marital status (Uncommitted)^p	0.49	0.48	0.70	0.56	0.49	0.47
Education enrollment (No)^p			0.71	0.90		
Higher education^p	0.30	0.54	0.35	0.77	0.47	0.50
Literacy status (Illiterate)^p	0.05	0.02	0.11	0.00	0.03	0.02
Disability condition (Disable)$^{p+}$			0.10	0.05	0.23	0.21
Work status (Working)^p	0.48	0.49			0.49	0.48
Natural gas (No)^h	0.33	0.17	0.22	0.17	0.23	0.20
Internet (No)^h	0.57	0.29	0.36	0.21	0.38	0.33
Informality index^h	0.10	0.03	0.06	0.04	0.05	0.05
Crowding index^h	1.41	0.87	1.23	1.08	1.04	0.96
Dependency ratio^h	0.27	0.28	0.27	0.18	0.27	0.29
Socioeconomic stratum (Low)^h	0.73	0.39	0.53	0.43	0.54	0.40

p: taken at individual level in labor model, and from the HoH in the other two models.
h: taken at household level for the three models.
p+: considered positive if at least one household member presents the condition.

When estimating regression models for the three outcomes of interest, obtained parameters are, in general, intuitive and consistent with previous findings in the literature. Both individual and household characteristics have proven to be relevant in determining

these outcomes (see Table 2). Specifically, household income shows a positive relation with higher levels of education, access to gas and internet, and better housing conditions. Similarly, formal employment tends to be more prevalent among adults, men, individuals with higher education, and those living in better housing conditions. In contrast, the quality of life perception presents a lower number of significant factors and with a difference in the sign of the education variable indicating that the better perception is associated with households with a less educated head of household.

Tabla 2: Estimated coefficients of the linear mixed models

	Dependent variable:		
	Household income	Labor formality	Positive perception
Age (Young, <29)^p	−0.05 (0.04)	−0.47*** (0.04)	−0.01 (0.09)
Gender (Female)^p	−0.07** (0.03)	−0.89*** (0.03)	−0.11* (0.06)
Marital status (Uncommitted)^p	0.04* (0.03)	−0.01 (0.03)	−0.05 (0.05)
Education enrollment (No)^p		0.98*** (0.05)	
Higher education^p	0.18*** (0.02)	1.51*** (0.03)	−0.14*** (0.05)
Literacy status (Illiterate)^p	−0.14** (0.07)	−1.96*** (0.18)	0.02 (0.14)
Disability condition (Disable)$^{p+}$		−0.72*** (0.06)	−0.14** (0.05)
Work status (Working)^p	0.06** (0.02)		−0.05 (0.05)
Natural gas (No)^h	−0.05* (0.03)	−0.08* (0.04)	−0.08 (0.06)
Internet (No)^h	−0.28*** (0.02)	−0.47*** (0.04)	0.01 (0.05)
Informality index^h	0.16*** (0.05)	−0.07 (0.08)	0.36*** (0.11)
Crowding index^h	−0.45*** (0.02)	−0.09*** (0.03)	−0.11*** (0.04)
Dependency ratio^h	0.11*** (0.04)	−1.47*** (0.07)	0.03 (0.08)
Socioeconomic stratum (Low)^h	−0.32*** (0.03)	0.14*** (0.04)	−0.45*** (0.07)
Constant	13.24*** (0.04)	−1.60*** (0.07)	0.46*** (0.09)
Observations	9228	30941	9228
Log Likelihood	-13246.45	-13246.45	-13246.45
Akaike Information Criterion	26522.9	26984.22	12439.21
Bayesian Information Criterion	26629.85	27109.32	12546.16
$\hat{\sigma}^2_\mu$	0.24	0.19	0.47

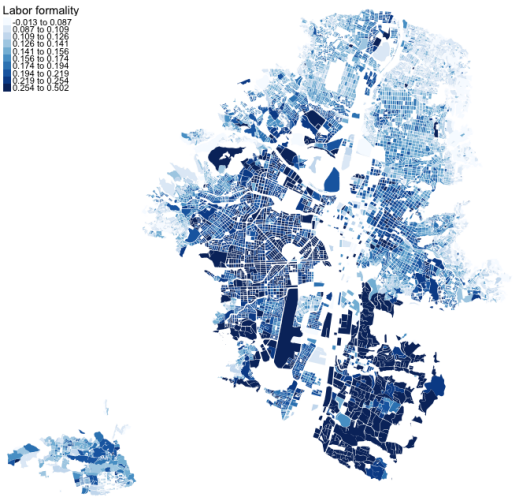
Significance codes: '***' 0.001 '**' 0.05 '*' 0.1
p: taken at individual level in labor model, and from the HoH in the other two models.
h: taken at household level for the three models.
p+: considered positive if at least one household member presents a disability.

Using the results of regression models, the simulation is implemented to estimate the outcomes at the block level. The results are presented in the Figure 2 where spatial differences are highlighted. All outcomes are presented in positive terms, meaning higher levels represent better socioeconomic conditions. Therefore, housing informality is presented as the probability of not observing deprivation in housing quality. Specifically, it is shown, as expected, that the south of the city concentrates census blocks with high levels in each of the outcomes, while the eastern and western slopes concentrate the lowest levels. It is also observed that the correlation between the outcomes is not perfect. For example, the northwest zone of the city, in relative terms, shows high levels of the housing formality

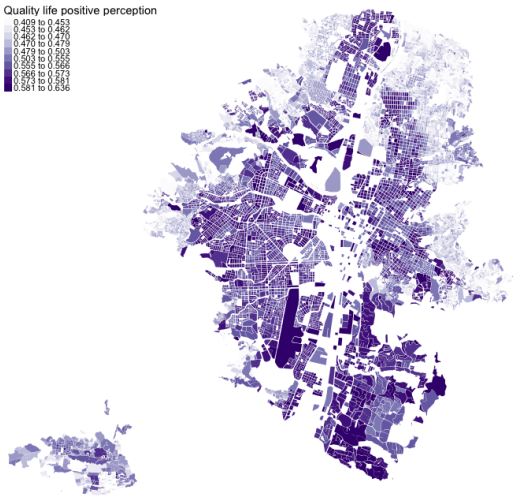
index but less intensity in labor formality.

Figure 2: Estimated outcomes using SAE

(a) Household income



(b) Formal labor



(c) Quality life perception

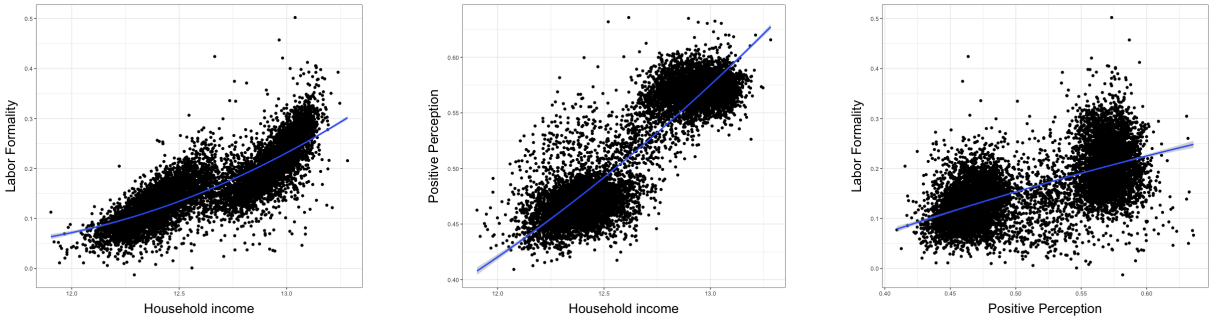
(d) Housing Formality Index (No deprived)

Source: Own calculations using regression models and Census.

A potential concern regarding the estimated outcomes is that the use of common predictors may generate a high level of dependence, and therefore the predicted values could be showing similar information. However, the results reveal that spatial patterns are diverse, and although there is a positive relationship, it is not perfect. In order to verify whether each outcome represents different spatial patterns, Figure 3 shows the pair-wise relationship between the outcomes. An expected result is that these variables are positively related, i.e., census block with higher average incomes tend to exhibit higher levels of formality and positive perceptions of quality of life. In addition, there seems to be a less strong relationship for the formality and perception pair. An essential element is that although the predicting equations include a similar set of information,

the relationship between predictors and outcomes varies in each case, which incorporates degrees of freedom to capture different spatial patterns. For example, marital status or work status is relevant for exploiting household income, but is not statistically significant for perceived quality of life. Similarly, the coefficients associated with the Formality index and Crowding index differ significantly in magnitude.

Figure 3: Correlations between estimated outcomes



(a) Household income vs labor formality

(b) Household income vs life quality perception

(c) Life quality perception vs labor formality

Source: Own calculations.

4 Selecting predictors of socioeconomic outcomes

From the estimated outcomes, the LASSO and Adaptive LASSO predictor selection algorithms are used. This requires two steps: firstly to estimate the model including the penalty term that allows to identify the relevant predictors, and secondly to implement OLS using the subset of selected predictors. In the initial stage, both LASSO methods employ regularization to shrink the model parameters, effectively reducing some coefficients to zero. Table 3 shows the results for the minimum value of λ found for both selection algorithms, indicating that in general, Adaptive LASSO has a more restricted selection of predictors, choosing between 26 and 29, while the other selects more than 35 in all cases. In particular, Adaptive LASSO significantly reduces the number of predictors, eliminating up to 14 (35%) predictors. Furthermore, The selected predictors do not perfectly match among the outcomes. In fact, while all variables are relevant for predicting income in the Lasso algorithm, a few variables are excluded for the other outcomes. Interestingly, the excluded variables for these outcomes are not the same.

Tabla 3: Summary of selected variables

Variable	Minimum λ		Number of variables	
	Regular Lasso	Adaptive Lasso	Regular Lasso	Adaptive Lasso
HH Income	0.0002	0.0104	40	26
Labor Formality	0.0016	0.0036	36	29
Positive Perception	0.0011	0.0016	35	27
No Deprivation	0.0017	0.0055	35	27

Once the predictors are selected, OLS regression models are estimated to compare the magnitudes of the estimated parameters. To do so, all the predictors are centered (mean is substracted) and scaled (divided by the standard deviation). Figures 4 and 5 display the results for Lasso and Adaptive Lasso, respectively, ordering the variables in descending order according to their relationship with household income. In general, there is a match across outcomes, but with some nuances in terms of magnitude. In addition, the predictors from the different sources show relevance, with public transport variables being of lesser relative importance. For instance, focusing the analysis on the results for adaptive Lasso, which results in a more parsimonious model, variables associated with the size and shape of the buildings, the incidence of POI, and both the intensity of nightlights has a positive relevance for predicting income. A negative relationship is found with the density of points of interest and population density, which can be explained by the fact that Medellin is a highly segregated city where neighborhoods with lower incomes are characterized by high agglomerations.

Comparing between outcomes, the greatest difference is identified for the no deprivation index where changes are registered in the signs in variables. Particularly, it excludes variables such as distance to transportation and population density per built area, and includes additional characteristics associated with the shape of buildings and features of the POI. Interestingly, it changes the sign of the diversity of points of interest to negative. This may be related to the fact that areas with higher deprivation are generally slums with a lack of planning with atypical and less uniform shapes that are captured by GUF.

Another interesting finding is that nighttime lights are not the primary determinant of the studied outcomes, which is consistent with recent studies discussing the relevance of using this variable as a proxy for economic variable and that remains as an inconclusive debate. Concerns have been raised about its inaccuracy in low-income settings and rural areas, as well as the inevitable limitations arising from varying lighting technologies and levels of light pollution, potentially resulting in inconsistencies (see Jean et al., 2016; Saiz and Salazar Miranda, 2023b; Pérez-Sindín et al., 2021; Puttanapong et al., 2020, for further discussions). Moreover, our results imply that proxies should include other sources that provide additional information to complement what nighttime lights are capable of capturing. Finally, the biggest difference in variable selection between Adaptative LASSO and LASSO correspond to variables from GUF and POI.

The main message is that a combination of variables is crucial for establishing intra-city differences, and therefore all sources are relevant as they increase the diversity of information that allows capturing spatial differences. For example, variables from GUF provide predictors with both positive and negative relationships that enhance the models' ability to generate diverse predictions. One way to show the relevance of information sources is to measure changes in the model fit level in two scenarios: first, by removing one of the sources, and second, by comparing fit metrics using only one of the sources. The results of these sensitivity exercises using AIC and adjusted R^2 as criteria are presented in Tables 5 and 4. In the columns identified as "in," it indicates the level of fit if only a specific information source is included, while "out" refers to the case where that source is excluded. The results corresponding to all sources serve as a benchmark for comparison.

For all outcomes, the individual source with the best fit is GUF, implying that it could serve as a good proxy for these socioeconomic variables. However, the actual values of AIC

and adjusted R^2 are far from those obtained using all information sources. In order, the following sources in relevance are variables associated with density and information from satellite images. In contrast, variables associated with public transport seem to contribute less to prediction. These findings coincide with what is observed when performing a leave-one-out exercise. That is, the worst relative performance occurs when variables associated with GUF are excluded, while the change is minimal in the case of public transport. Overall, it is evident that from non-traditional sources, it is possible to capture intra-city differences, and although the relevance of each source varies, the best result is obtained when all are combined.

5 Concluding remarks

Valuable information on segregation patterns and spatial discrepancies are crucial for local government decision making. Unfortunately, this information is not easy to build, and depends on the availability of highly regular information. For local governments it is expensive to recurrently update surveys that make it possible to measure the living conditions of the inhabitants. Therefore, it is imperative to identify differences at the spatial level to make more effective decisions. In this context, alternative sources, such as remote sensing data, have the ability to offer a complete estimated map of inequalities within a region or city.

By combining small area estimation techniques and variable selection algorithms, evidence is provided that alternative sources to census and surveys make it possible capture spatial differences in socioeconomic. Our estimates allow us to argue that building attributes, variables associated with points of interest, and satellite images allows to have a clear insight into the living conditions of the inhabitants of a city.

Interestingly, there is no prevalence of one type of information, on the contrary the combination of nightlight intensity, points of interest, building shape and land use indices are relevant. The use of alternative sources presents different advantages for policy makers. For example, it provides a higher resolution of patterns within a municipality or city, and allows a more constant monitoring of socioeconomic variables. This does not imply that this type of sources replaces traditional measurement instruments. On the contrary, they complement them, and in fact, their effectiveness depends on relating and integrating these two types of information.

References

Abascal, A., Rodríguez-Carreño, I., Vanhuysse, S., Georganos, S., Sliuzas, R., Wolff, E., and Kuffer, M. (2022). Identifying degrees of deprivation from space using deep learning and morphological spatial analysis of deprived urban areas. *Computers, environment and urban systems*, 95:101820.

Addison, D. M. and Stewart, B. (2015). Nighttime lights revisited: the use of nighttime lights data as a proxy for economic variables. *World Bank Policy Research Working Paper*, (7496).

Akbar, P., Couture, V., Duranton, G., and Storeygard, A. (2023a). Mobility and Congestion in Urban India. *American Economic Review*, 113(4):1083–1111.

Akbar, P. A., Couture, V., Duranton, G., and Storeygard, A. (2023b). The fast, the slow, and the congested: Urban transportation in rich and poor countries. Technical report, National Bureau of Economic Research.

Asian Development Bank (2020). Introduction to Small Area Estimation Techniques:: A Practical Guide for National Statistics Offices. Technical report, Asian Development Bank, Manila, Philippines. Edition: 0 ISBN: 9789292622237 9789292622220.

Baragwanath, K., Goldblatt, R., Hanson, G., and Khandelwal, A. K. (2021). Detecting urban markets with satellite imagery: An application to india. *Journal of Urban Economics*, 125:103173.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association*, 83(401):28–36. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Baum-Snow, N. and Turner, M. A. (2017). Transport infrastructure and the decentralization of cities in the people's republic of china. *Asian development review*, 34(2):25–50.

Burchfield, M., Overman, H. G., Puga, D., and Turner, M. A. (2006). Causes of sprawl: A portrait from space. *The Quarterly Journal of Economics*, 121(2):587–633.

Ch, R., Martin, D. A., and Vargas, J. F. (2021). Measuring the size and growth of cities using nighttime light. *Journal of Urban Economics*, 125:103254.

Charris, C., Velilla, R., and Chaves, L. (2019). Mapping the human development index using nighttime lights inside brazil. *XVII ENABER—Encontro Nacional da Associação Brasileira de Estudos Regionais e Urbanos. https://brsa. org. br/wp-content/uploads/wpcf7-submissions/990/manuscript_Iden. pdf.*

Che, M. and Gamba, P. (2019). Intra-urban change analysis using sentinel-1 and nighttime light data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(4):1134–1142.

Chen, X. and Nordhaus, W. D. (2011). Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences*, 108(21):8589–8594.

Doll, C. N., Muller, J.-P., and Morley, J. G. (2006). Mapping regional economic activity from night-time light satellite imagery. *Ecological Economics*, 57(1):75–92.

Donaldson, D. and Storeygard, A. (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, 30(4):171–198.

Duque, J. C., Patino, J., Ruiz, L., and Pardo, J. (2013). Quantifying Slumness with Remote Sensing Data.

Durst, N. J., Sullivan, E., Huang, H., and Park, H. (2021). Building footprint-derived landscape metrics for the identification of informal subdivisions and manufactured home communities: A pilot application in hidalgo county, texas. *Land Use Policy*, 101:105158.

Elbers, C., Lanjouw, J. O., and Lanjouw, P. (2003). Micro-Level Estimation of Poverty and Inequality. *Econometrica*, 71(1):355–364.

Elbers, C. and van der Weide, R. (2014). Estimation of normal mixtures in a nested error model with an application to small area estimation of poverty and inequality. *World Bank Policy Research Working Paper*, (6962).

Elvidge, C. D., Baugh, K. E., Kihn, E. A., Kroehl, H. W., Davis, E. R., and Davis, C. W. (1997). Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption. *International Journal of Remote Sensing*, 18(6):1373–1379.

Engstrom, R., Hersh, J., and Newhouse, D. (2022). Poverty from space: Using high resolution satellite imagery for estimating economic well-being. *The World Bank Economic Review*, 36(2):382–412.

Freijeiro-González, L., Febrero-Bande, M., and González-Manteiga, W. (2022). A Critical Review of LASSO and Its Derivatives for Variable Selection Under Dependence Among Covariates. *International Statistical Review*, 90(1):118–145. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/insr.12469.

Fujii, T. and van der Weide, R. (2020). Is predicted data a viable alternative to real data? *The World Bank Economic Review*, 34(2):485–508.

Furnham, A. and Cheng, H. (2018). Social-demographic indicators, cognitive ability, personality traits, and region as independent predictors of income: findings from the uk household longitudinal study (ukhls). *Journal of Intelligence*, 6(2):19.

Galdo, V., Li, Y., and Rama, M. (2021). Identifying urban areas by combining human judgment and machine learning: An application to india. *Journal of Urban Economics*, 125:103229.

Ganter, M., Toetzke, M., and Feuerriegel, S. (2022). Mining Points-of-Interest Data to Predict Urban Inequality: Evidence from Germany and France. *Proceedings of the International AAAI Conference on Web and Social Media*, 16:216–227.

Gibson, J. and Boe-Gibson, G. (2021). Nighttime lights and county-level economic activity in the united states: 2001 to 2019. *Remote Sensing*, 13(14):2741.

Gibson, J., Olivia, S., Boe-Gibson, G., and Li, C. (2021). Which night lights data should we use in economics, and where? *Journal of Development Economics*, 149:102602.

Goldblatt, R., Heilmann, K., and Vaizman, Y. (2007). Can Medium-Resolution Satellite Imagery Measure Economic Activity at Small Geographies? Evidence from Landsat in Vietnam.

Henderson, J. V., Storeygard, A., and Weil, D. N. (2012). Measuring economic growth from outer space. *American economic review*, 102(2):994–1028.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794. Publisher: American Association for the Advancement of Science.

Jiang, S., Alves, A., Rodrigues, F., Ferreira, J., and Pereira, F. C. (2015). Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems*, 53:36–46.

Jochem, W. C. and Tatem, A. J. (2021). Tools for mapping multi-scale settlement patterns of building footprints: An introduction to the r package foot. *PLoS One*, 16(2):e0247535.

Keola, S., Andersson, M., and Hall, O. (2015). Monitoring economic development from space: using nighttime light and land cover data to measure economic growth. *World Development*, 66:322–334.

Kilic, T., Serajuddin, U., Uematsu, H., and Yoshida, N. (2017). Costing household surveys for monitoring progress toward ending extreme poverty and boosting shared prosperity. *World Bank Policy Research Working Paper*, (7951).

Kohli, D., Sliuzas, R., and Stein, A. (2016). Urban slum detection using texture and spatial metrics derived from satellite imagery. *Journal of Spatial Science*, 61(2):405–426. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14498596.2016.1138247.

Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., and Tzavidis, N. (2019). The *R* Package **emdi** for Estimating and Mapping Regionally Disaggregated Indicators. *Journal of Statistical Software*, 91(7).

Kuffer, M., Sliuzas, R., van Maarseveen, M., Pfeffer, K., and Baud, I. (2017). City nighttime light variations using iss images. In *2017 Joint Urban Remote Sensing Event (JURSE)*, pages 1–4. IEEE.

Lee, K. and Braithwaite, J. (2022). High-resolution poverty maps in sub-saharan africa. *World Development*, 159:106028.

Li, C., Zhu, H., Ye, X., Jiang, C., Dong, J., Wang, D., and Wu, Y. (2020). Study on average housing prices in the inland capital cities of china by night-time light remote sensing and official statistics data. *Scientific reports*, 10(1):1–20.

Li, F., Liu, X., Liao, S., and Jia, P. (2021). The modified normalized urban area composite index: A satelliate-derived high-resolution index for extracting urban areas. *Remote Sensing*, 13(12):2350.

Lin, J. and Shi, W. (2020). Statistical correlation between monthly electric power consumption and viirs nighttime light. *ISPRS International Journal of Geo-Information*, 9(1):32.

Liu, H., He, X., Bai, Y., Liu, X., Wu, Y., Zhao, Y., and Yang, H. (2021). Nightlight as a Proxy of Economic Indicators: Fine-Grained GDP Inference around Chinese Mainland via Attention-Augmented CNN from Daytime Satellite Imagery. *Remote Sensing*, 13(11):2067. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

Marty, R. and Duhaut, A. (2024). Global poverty estimation using private and public sector big data sources. *Scientific Reports*, 14(1):3160. Number: 1 Publisher: Nature Publishing Group.

Mellander, C., Lobo, J., Stolarick, K., and Matheson, Z. (2015). Night-time light data: A good proxy measure for economic activity? *PloS one*, 10(10):e0139779.

Molina, I. and Rao, J. (2010). Small area estimation of poverty indicators. *Canadian Journal of statistics*, 38(3):369–385.

Niu, H. and Silva, E. A. (2021). Delineating urban functional use from points of interest data with neural network embedding: A case study in Greater London. *Computers, Environment and Urban Systems*, 88:101651.

Observatory, T. E. (2000). Measuring Vegetation (NDVI & EVI). Publisher: NASA Earth Observatory.

Pan, Y., Chen, J., Yan, X., Lin, J., Ye, S., Xu, Y., and Qi, X. (2022). Identifying the spatial–temporal patterns of vulnerability to re-poverty and its determinants in rural china. *Applied Spatial Analysis and Policy*, 15(2):483–505.

Pérez-Sindín, X. S., Chen, T.-H. K., and Prishchepov, A. V. (2021). Are night-time lights a good proxy of economic activity in rural areas in middle and low-income countries? examining the empirical evidence from colombia. *Remote Sensing Applications: Society and Environment*, 24:100647.

Posada, H. M., García, A., Londoño, D., et al. (2022). The external effects of public housing developments on informal housing: The case of medellín, colombia. Technical report.

Posada, H. M. and García-Suaza, A. (2022). Transit infrastructure and informal housing: Assessing an expansion of Medellín's Metrocable system. *Transport Policy*, 128:209–228.

Puttanapong, N., Martinez, A. M., Addawe, M., Bulan, J., Durante, R. L., and Martillan, M. (2020). Predicting poverty using geospatial data in Thailand. Working Paper 630, ADB Economics Working Paper Series.

Rangel-Gonzalez, E. and Llamosas-Rosas, I. (2019). An alternative method to measure non-registered economic activity in mexico using satellite nightlights. In *Presentation given at the 7th International Monetary Fund Statistical Forum, November*, volume 14.

Rao, J. N. and Molina, I. (2015). *Small area estimation*. John Wiley & Sons.

Redding, S. J. and Turner, M. A. (2015). Transportation costs and the spatial organization of economic activity. *Handbook of regional and urban economics*, 5:1339–1398.

Saiz, A. and Salazar Miranda, A. (2023a). Understanding urban economies, land use, and social dynamics in the city: Big data and measurement. *MIT Center for Real Estate Research Paper*, (23/19).

Saiz, A. and Salazar Miranda, A. (2023b). Understanding Urban Economies, Land Use, and Social Dynamics in the City: Big Data and Measurement.

Sherman, L., Proctor, J., Druckenmiller, H., Tapia, H., and Hsiang, S. M. (2023). Global high-resolution estimates of the united nations human development index using satellite imagery and machine-learning. Technical report, National Bureau of Economic Research.

Struijs, P., Braaksma, B., and Daas, P. J. (2014). Official statistics and big data. *Big Data & Society*, 1(1):2053951714538417.

Sutton, P. C. and Costanza, R. (2002). Global estimates of market and non-market values derived from nighttime satellite imagery, land cover, and ecosystem service valuation. *Ecological economics*, 41(3):509–527.

Tang, B., Liu, Y., and Matteson, D. S. (2022). Predicting poverty with vegetation index. *Applied Economic Perspectives and Policy*, 44(2):930–945.

Terol-Cantero, M. C., Martín-Aragón Gelabert, M., Costa-López, B., Manchón López, J., and Vázquez-Rodríguez, C. (2023). Causal attribution for poverty in young people: Sociodemographic characteristics, religious and political beliefs. *Social Sciences*, 12(5):308.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288. Publisher: [Royal Statistical Society, Wiley].

UN-Habitat (2003). The challenge of slums: Global report on human settlements.

van der Weide, R., Blankespoor, B., Elbers, C., and Lanjouw, P. (2023). How Accurate is a Poverty Map Based on Remote Sensing Data? An Application to Malawi.

Wang, G. and Peng, W. (2021). Detecting influences of factors on gdp density differentiation of rural poverty changes. *Structural Change and Economic Dynamics*, 56:141–151.

Watmough, G. R., Atkinson, P. M., Saikia, A., and Hutton, C. W. (2016). Understanding the evidence base for poverty–environment relationships using remotely sensed satellite data: an example from assam, india. *World Development*, 78:188–203.

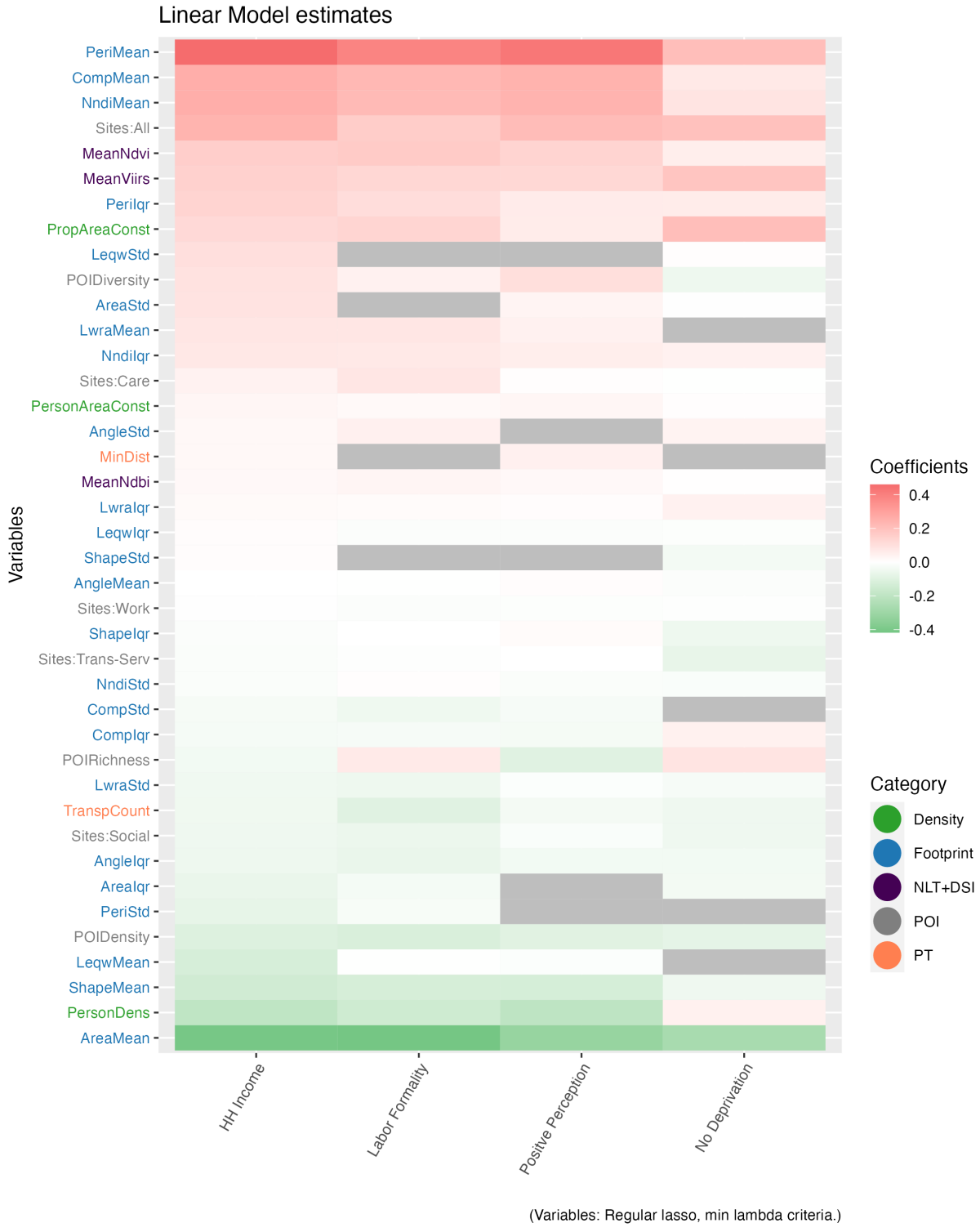
Xu, H. (2007). Extraction of Urban Built-up Land Features from Landsat Imagery Using a Thematicoriented Index Combination Technique. *Photogrammetric Engineering & Remote Sensing*, 73(12):1381–1391.

Zha, Y., Gao, J., and Ni, S. (2003). Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *International Journal of Remote Sensing*, 24(3):583–594. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01431160304987.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

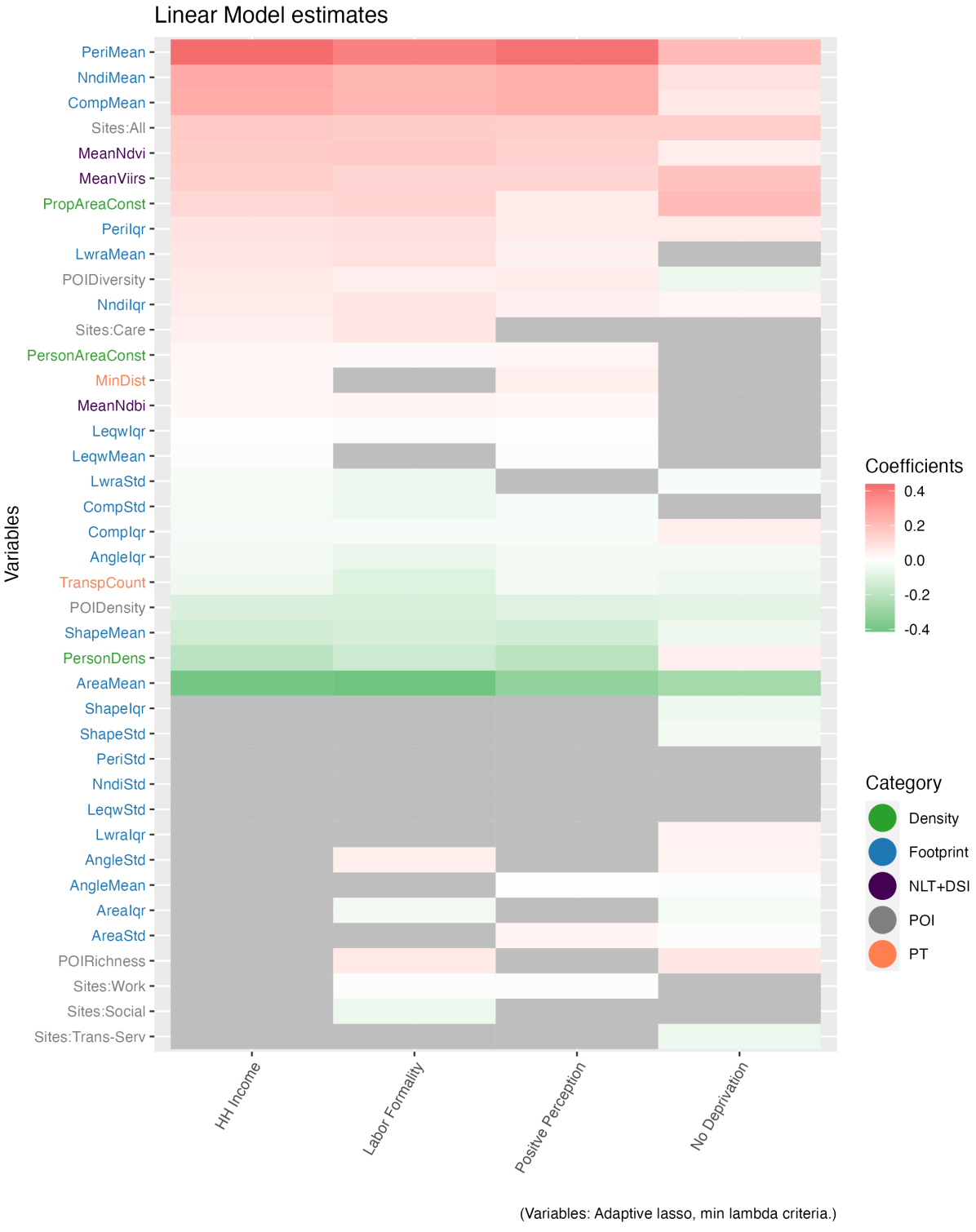
Appendix

Figure 4: Coefficients of selected variables using LASSO



Source: Own calculations.

Figure 5: Coefficients of selected variables using Adaptative LASSO



Source: Own calculations.

Tabla 4: Sensitivity of include information using AIC criterion

Source	HH Income		Labor Formality		Positive Perception		No Deprivation	
	AIC.in	AIC.out	AIC.in	AIC.out	AIC.in	AIC.out	AIC.in	AIC.out
All	-4074	3851	-31856	-26680	-36404	-30246	-15474	-13490
NLT + DSI	661	-3319	-28884	-31315	-32663	-35928	-14654	-15182
POI	850	-3640	-28790	-31458	-32619	-36203	-13894	-15411
PT	3601	-4007	-26882	-31740	-30403	-36327	-13554	-15454
Density	150	-3477	-28978	-31559	-33329	-35895	-14668	-15166
GUF	-1851	-2562	-30251	-30924	-34908	-35209	-14705	-15311

Source: Own calculations. AIC.in refers to the case where only the variables from the corresponding group are estimated, while AIC.out considers the case where that group is excluded.

Tabla 5: Sensitivity of include information using R^2 adjusted criterion

Source	HH Income		Labor Formality		Positive Perception		No Deprivation	
	R2.in	R2.out	R2.in	R2.out	R2.in	R2.out	R2.in	R2.out
All	0.55	0.00	0.41	0.00	0.46	0.00	0.18	0.00
NLT + DSI	0.27	0.52	0.20	0.37	0.22	0.44	0.11	0.16
POI	0.26	0.53	0.19	0.38	0.21	0.45	0.04	0.18
PT	0.02	0.55	0.02	0.40	0.02	0.46	0.01	0.18
Density	0.31	0.52	0.21	0.39	0.27	0.44	0.11	0.16
GUF	0.44	0.48	0.30	0.35	0.38	0.39	0.12	0.17

Source: Own calculations. R2.in refers to the case where only the variables from the corresponding group are estimated, while R2.out considers the case where that group is excluded.

Tabla 6: Variable definitions

Variable	Description	Source
Sites:All	Count of POI in the block regardless on the category.	POI
Sites:Trans-Serv	Count of POI in Transport and Services categories.	
Sites:Care	Count of POI in Care category which includes school, healthcare, childcare and adultcare.	
Sites:Social	Count of POI in Social category which includes recreation, other, meals, exercise, goods and errands.	
Sites:Work	Count of POI in Work category.	
POIRichness	POI Richness	
POIDiversity	POI Diversity	
POIDensity	POI Density	
MeanNdvi	Mean NDVI (Daytime Satellite data)	NLT + DSI
MeanNdbi	Mean NDBI (Daytime Satellite data)	
MeanViirs	Mean Viirs (Nightlights data)	
AreaMean	Area Mean	GUF
AreaStd	Area Standard deviation	
Arealqr	Area Interquartile Range	
AngleMean	Angle Mean	
AngleStd	Angle Standard deviation	
Anglelqr	Angle Interquartile Range	
CompMean	Compactness Mean	
CompStd	Compactness Standard deviation	
Complqr	Compactness Interquartile Range	
LwraMean	Length-width Mean	
LwraStd	Length-width Standard deviation	
LwraIqr	Length-width Interquartile Range	
LeqwMean	Length-equivalent-width Mean	
LeqwStd	Length-equivalent-width Standard deviation	
LeqwIqr	Length-equivalent-width Interquartile Range	
NndiMean	Nearest Neighbour Distance Mean	
NndiStd	Nearest Neighbour Distance Standard deviation	
Nndilqr	Nearest Neighbour Distance Interquartile Range	
PeriMean	Perimeter Mean	
PeriStd	Perimeter Standard deviation	
PeriIqr	Perimeter Interquartile Range	
ShapeMean	Shape Mean	
ShapeStd	Shape Standard deviation	
Shapelqr	Shape Interquartile Range	
PersonAreaConst	Person built area: Number of Persons per built area (footprint total area)	Density
PropAreaConst	Built area proportion: Total footpint area divided by total block area	
PersonDens	Persons dentsity: Number of persons divided by total block area	
MinDist	Minimum distance to a public transport station (bus or metro)	PT
TranspCount	Count of public transit stations (bus or metro)	