A methodological approach to estimate population density in a land-use cover change model: a case study in Bogota - Colombia

ABSTRACT

In rapidly growing cities on developing countries, urban sprawl has been a trend. The city of Bogota and its adjacent municipalities have undergone densification processes during the last decades, especially on peripheral locations where low-income settlements have established. This study uses ordinary least squares (OLS) and geographically weighted regression (GWR) models to include density analysis in a Land-use cover change (LUCC) model based on cellular automata. Variables related to residential land-use and transport and urban oriented variables on housing built area were used to predict future population densities. The global regression model (OLS), is used to estimate statistically significant explanatory variables and overall model significance, while the local regression model (GWR), estimates parameters differently depending on spatial weighting among neighboring geographical zones. An indirect methodological approach was established to calculate population density on a low-resolution scale, using a well-specified GWR model that determine housing built area. The methodology proposed was tested in a case study in the city of Bogota. Results for Bogota suggest the methodology proposed improved the capability of LUCC models to derive densities. Further research is needed to better understand densification processes in the context of sustainable development for the city of Bogota and its region.

Keywords: Housing built area, Population density, Residential land-use, Geographically weighted regression

1. Introduction

Globally, urbanized land area is expanding faster than population. However, urban expansion is often poorly planned and not well integrated with land use and transportation networks. Rapid urban population growth could lead to sprawl that is poorly connected to urban activity centers, exacerbating congestion forces (Kaw & Roberts, 2016). To enhance livability in cities and improve opportunities for prosperity; planners, government policy makers, and stakeholders need to better manage the spatial city structure and give access to land in a sustainable way thinking in smart cities concepts.

LUCC models have been used widely to foster urban planning. Most models developed up-to-date have strong methodologies to predict changes in land uses based on regulation, accessibility and real market conditions. However, limited literature exists today in ways to include, in addition to land-use changes, predicting population density changes over-time. This paper presents a methodology to incorporate population density changes in a LUCC model.
Considering density in future land demands is a vital step for urban planning and land development. Increasing density in urban areas and promoting compact urban development can facilitate agglomeration economies, improve access to services, and generate property tax revenues. If implemented well, density and spatial structure-shaping actions can also lead to more efficient transportation networks. On the other hand, unbridled increase in density could lead to congestion of roads and other basic services, unaffordable commercial and residential land, and polluted air (Kaw & Roberts, 2016). Urban growth is a management problem all over the world. In most big cities planning or policies are unable to control growth (Lahti, 2008). This leads to uncontrolled urban sprawl, characterized by a low-density expansion. The causes of sprawl are typically connected to growth in population and economy, but also preferences in living, changes in demography, price of land and cultural traditions (Batty, et al., 1999).

Overly dispersed land-use patterns with homogeneous, low densities and urban sprawl usually require more land for development, and infrastructure costs per person are higher than where densities are higher. The notion of the compact city, which is often referred as the city of short distance, contrasts the car-oriented urban sprawl of many modern cities and can be characterized by combining efficient, multifunctional and multi-modal transport systems while fostering relatively high population densities (De Vries & et al., 2016). This city morphology is desirable, due to urban sprawl force a high demand for a limited supply of land which at the same time compels to informal land development (Masum & et al, 2016). Besides, inadequate housing in informal settlements (or slums) is known to be an obstacle to economic growth in cities (Weldesillassie & et al., 2016).

Density shortens distances between people and the places they need access to, therefore the question of density is one of how much activity, population and built area can be concentrated into a given urban area. Population densities focus on measures of people, dwellings and jobs per hectare. What concentrations of built area make sense in terms of shortening the distances between attractions while maintaining amenity? How close can we get to where we want or need to be? (Australian Research Council, 2014).

For planners and researchers to support the decision-making process of spatial and land-use planning, it is necessary to understand the causes and consequences of land-use change process.

Therefore, the effort of this research consists to develop and test processes to predict land-use cover change (LUCC). The testing environment is an existing model called Bogota Land Development (BoLD) that uses cellular automata (CA) software Metronamica (Paez & Escobar, 2016).

2. Research background
   2.1. Literature review

   A few number of studies have examined the changes in population density and built area in Bogota, for instance Bocarejo et al. (2013) evaluate the impact of TransMilenio BRT system on densification and land use in the city. According to this study, urban sprawl has been the general growth pattern in most developing cities worldwide, but the
city of Bogotá has experienced a densification process in specific areas during the last decade. Overall density in Bogotá has increased by 8% between 2001 and 2008, which means an increase of around 1400 inhabitants/km². On other study, Paez (2012) argues that apparently, there is not a clear understanding of what conditions (both social and environmental) Bogota offers as a very high-density city at a micro-level. It is important to understand if increased densification improves the population’s wellbeing and what is the environmental effect (Paez, 2012). Additionally, a detailed examination of density facilitates the identification of urban behaviors such as commission housing, where the provision of housing has particular characteristics depending on national and local policies (Paez, 2012).

Batty et al. (1999) paper present urban models whose dynamics are based on theories of development associated with cellular automata (CA) where temporal processes of change are presented through entirely local actions which take place in the immediate proximity of the various objects which influence in the spatial environment. In spatial systems, the objects are defined as cells which can take on various states and which are influenced by other cells in their immediate neighborhood. Dynamic change is thus conceived as change in a cell being a function of the states of those cells which involve its first nearest neighbors (Batty et al., 1999).

The analysis carried out for urban areas in the Netherlands by Broitman and Koomen (2015), explored residential development and density changes in relation to prevailing spatially explicit policies; using detailed geographical data about land-use and residential densities from 2000 onwards. First, they define densification share as the net increase in housing units in period \( T_0 - T_1 \) within the urban area of \( T_0 \) expressed as percentage of the total increase in housing units. Density increase as the increase in residential density in the urban area of \( T_0 \) expressed as percentage of the initial residential density in the area. And expansion density as the average density of housing units built in new residential areas in period \( T_0 - T_1 \), expressed as increased number of housing units per hectare (Broitman & Koomen, 2015). Based on this schematic representation, a quantitative understanding of the process of residential expansion is obtained through a GIS-based analysis where transitions to residential land-use in the observation period are performed by making a pixel-by-pixel comparison of two subsequent land-use data sets.

The notion of compact city as a strategy to reduce urban sprawl is the focal point in Chhetri et al. (2013) investigation, where they consider urban residential density as a surrogate measure for urban compactness, and empirically examine the cadastre database that contains details of every property in order to capture changes in urban residential density patterns in Melbourne, Australia using geospatial techniques. Findings indicate that urban densities across the buffer zones around Melbourne CBD are significantly different, and rapid densification of the inner outer zone is surprising (Chhetri, Han, Chandra, & Corcoran, 2013).

A case study in Wuhan city, PR China, presents a spatial data analysis method to seek and model major determinants of urban growth between the period 1993–2000. According to Cheng and Masser (2003) the method comprises exploratory data analysis and spatial logistic regression technique. The former can visually explore the spatial impacts of each explanatory variable, while the latter can provide a systematic confirmatory approach to comparing the
variables. The study concludes that the major determinants of urban growth are urban road infrastructure and developed area (Cheng & Masser, 2003).

A variety of approaches have been used to model the spatial process of urban growth, including potential models, Markov chains, and spatial logistic regression (He, Okada, Zhang, Shi, & Zhang, 2006). Recently, there is growing literature on applications of cellular automata (CA) models in urban growth and land-use cover change. CA models have the strong ability to represent non-linear, spatial and stochastic processes. Many works have already demonstrated the CA model’s capability to simulate spatial pattern and process of urban expansion in a very realistic way. He et al. (2006) developed a new urban expansion scenario model by coupling one “bottom-up” cellular automata based model and one “top-down” system dynamics based model. Results of this investigation suggest that urban expansion is restricted and limited by water resource availability and environment deterioration (He et al., 2006).

Density analysis lies at the core of studies on urban expansion; however, many methods in urban land density analysis are arbitrary and suffer from the lack of an established foundation. Jiao (2015) took as samples 28 major cities at three time points in China, to illustrate the general trend of urban land density variation from the urban center to the outside periphery, and propose a mathematical formulation of the rule using a nonlinear least squares fitting method (Jiao, 2015). Subsequently, the paper indicates the properties of the defined function and discuss how to use it to characterize urban form and the dynamics of urban sprawl.

According to Ku (2016) there are many CA models that apply ordinary least squares (OLS) regression model to estimate the weights of land-use driving factors. In these models, however, the dependent variable, which is normally the amount of change in land-use, may be correlated with each other spatially. This would result in spatial dependency in residuals causing higher standard deviations of estimations that might influence overall performance of the model (Ku, 2016). Thus, to solve the problem, this paper applies a spatial regression model to explain the amount of land-use change in a CA model.

In other study, a GIS-based cellular automata model was developed for exploring the vertical complexities of urban growth. Considering a series of variables, such as accessibility, population density and building density and height (Lin, Huang, Chen, & Huang, 2014). As a result, building height patterns were successfully simulated, and the model demonstrates that urban vertical growth patterns across space from the city center, through the fringe, and to the periphery and hinterland zones can be effectively modeled using the technique developed in this study.

In the context of promoting new urbanization, Zeng et al. (2015) integrate remote sensing, a geographical information system, and spatial analysis techniques to monitor and model urban expansion with a spatially explicit and multi-scale perspective in Wuhan, China. For the exploration of the driving mechanisms underlying urban expansion, 20 explanatory variables are categorized into three groups: characteristics, density, and proximity. Moran’s I index is later used to test the spatial autocorrelation in the percentage of urban built-up land area and the residuals. Finally, a spatial lag model and a spatial error model are applied to explore the causal factors of urban built-up land area (Zeng, Zhang, Cui, & He, 2015).
3. Methodology

The main idea of this research is to develop a regression model that allows to explain housing built area, based on the 100x100 meters residential cells of BoLD model. The first step to achieve this, was to identify and extract the residential cells form the 2014 land-use map. Secondly, assign them the sum of all the constructed area of residential properties that fall inside of each cell, based on the cadastral dataset. And thirdly, consolidate a shapefile with the residential polygons of the same size of the raster land-use map, so it is ready to run regression models. The study area for the regression analysis excludes the 6 municipalities considered in BoLD and concentrates only on Bogota’s urban perimeter delimited by the UPZs.

The integrated model framework developed in this paper is shown in Fig. 1. The methodology can be divided in three major parts including the exploratory regression, the specification of an Ordinary Least Squares (OLS) model, and the application of the Geographically Weighted Regression (GWR). The detail of each part will be further discussed.

![Integrated model framework](image)

Based on a regression analysis, which is the most commonly used statistic method in social sciences, a methodological approach was developed to examine the built area for residential land use in 2014. This method is used to evaluate relationships between two or more variables. Identifying and quantifying the effect of one or more explanatory variables over a dependent variable of interest allow to understand, predict, or simply examine the causes of certain phenomenon.

The exploratory regression is a data mining tool that will try all possible combinations of explanatory variables to see which models pass all the necessary OLS diagnostics. By evaluating all possible combinations of the candidate explanatory variables, the probability of finding the best model to explain the dependent variable increases. Exploratory regression is similar to stepwise regression, rather than only looking for models with high Adjusted R² values, the tool seeks for models that meet all of the requirements and assumptions of the OLS method (Anderson et al., 2014). Once all the plausible explanatory variables are identified through this process, it is possible to continue with the determination of a well-specified global regression model using OLS.

An OLS model is a linear regression model that is used in this research for the purpose of estimating the linear relationship between the dependent and independent variables. The coefficients in this regression model are estimated using least squares method. The form of the OLS regression model used in this paper can be described as Equation (1):

\[ y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon \]  

(1)

where \( y \) is the dependent variable, that in this case is the residential built area in 2014; \( x_1 \) and \( x_n \) represent the independent or explanatory variables; \( \beta_1 \) and \( \beta_n \) are the coefficients of the regression model; \( \beta_0 \) is the intercept of the model; and \( \varepsilon \) represent the residuals.

Ordinary least squares is the proper starting point for all spatial regression analyses, because it provides a
global model of the variable or process of interest; the result of OLS is a single regression equation to represent that variable (Ku, 2016). OLS regression is a straightforward method that has well-developed theory behind it, and several effective diagnostics to verified the validity and significance of the model. Nevertheless, OLS is only effective and reliable when the data and regression model comply with the assumptions inherently required by this method theory. Spatial data often violates OLS assumptions and requirements, so it is important to use the diagnostics regression tools in conjunction with appropriate spatial tools (Spatial Autocorrelation Moran’s I index) that can assess whether the data is being well implemented and the regression model is well-specified. In order to achieve this, ArcGIS OLS tool was used to determine the global regression model and the diagnostics tests to evaluate the validity and effectiveness of the model.

The diagnostics tests are: 1) Model performance: assess the model performance through the multiple $R^2$ and adjusted $R^2$. 2) Explanatory Variables significance and multicollinearity: assess by the $t$-statistic, p-value, and the Variance Inflation Factor (VIF) value. 3) Model significance: assess by the Joint F-Statistic and Joint Wald Statistic. 4) Model stationarity: assess by the Koenker (BP) Statistic which determine whether the explanatory variables in the model have a consistent relationship to the dependent variable both in geographic space and in data space. 5) Model bias: assess by the Jarque-Bera statistic which indicates whether the residuals are normally distributed. 6) Model spatial autocorrelation: assess by the Spatial Autocorrelation Moran’s I index, which ensure that the regression residuals are spatially random.

Finally, the spatial regression used in this paper is a Geographically Weighted Regression (GWR). OLS regression models with statistically significant non-stationarity are often good candidates GWR analysis (Chiou, Jou, & Yang, 2015). GWR is one of several spatial regression techniques increasingly used in geography and other disciplines due to it provides a local model by fitting a regression equation to every feature in the dataset. GWR builds these separate equations by incorporating the dependent and explanatory variables of features falling within the bandwidth of each target feature. The advantage of specifying a local model for each geographic analysis zone, is that allows to compare and identify spatial patterns and changes among the explanatory variables as well as for the dependent variable, the way GWR do this is by estimating parameters differently depending on spatial correlations among neighboring regions.

The GWR estimates the coefficients for each define geographic location using a weighted least squares method, as shown in the Equation (2) (Su, Lei, Li, Pi, & Cai, 2017). GWR extends the OLS Equation (1) by incorporating spatial dependence, which is determined by the kernel bandwidth. The relative weight of the samples within the spatial scope is determined by a Gaussian distance decay function Equation (3) (Su et al., 2017), where nearer samples have greater influences.

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{n} \beta_k(u_i, v_i)x_{ik} + \epsilon_i$$

Where $y_i$ is the dependent variable at location $i$; $x_{ik}$ is the value of the $k^{th}$ explanatory variable at location $i$, $(u_i, v_i)$ represent the coordinates of position $i$; $k$ is the number of independent variables; $\beta_0(u_i, v_i)$ and $\beta_k(u_i, v_i)$ are the intercept parameter and local regression coefficient for the $k^{th}$ explanatory variable,
respectively at location \( i \); and \( \varepsilon_i \) represents the residuals or random error at location \( i \).

\[
w_{ij} = \exp\left(-\frac{d_{ij}^2}{h^2}\right) \tag{3}
\]
where \( w_{ij} \) is the weight for sample \( j \) surrounding sample \( i \); \( d_{ij} \) represents the distance between the sample \( i \) and \( j \), and \( h \) is the kernel bandwidth.

4. Case study: Bogota

4.1. Study area

Bogota is one of the most important metropolitan areas in South America, and is the most important economic and industrial center of Colombia and the largest and most populous city in the country. It is a dense and compact city with a rapidly growing population (Bocarejo et al., 2013). Urban and transport planning must play a crucial role in developing a sustainable city that seeks to supply the needs for housing, land, and transport infrastructure. At this moment, the city is suffering land scarcity, particularly on the urban core, which might induce urban sprawl. According to Buxton & Tieman (2005), the debate over the costs and benefits of urban consolidation has perhaps never been more important given the rise in the importance of sustainability principles as a basis of planning, the growth of mega cities and development pressure on urban fringes.

In 2015, Bogota had 7.9 million inhabitants spread on an urban area of 365 km\(^2\), resulting in an average population density close to 21,600 inhabitants per square kilometers. During the last two decades, population migration has encouraged a fast and uncontrolled growth of the city’s boundaries, and illegal settlements on the urban fringes became the most common way for developing low-income housing. Regarding the urban built environment Bogota is characterized as a low-rise, with reduced green areas and public spaces. In average, the number of floors is around 2.05 and for predominantly residential constructions it goes up to 2.3 (Bocarejo et al., 2014). Despite this low-rise structure, the average population density remains high due to the poor average indoor space per capita of around 25 m\(^2\)/inhab (Bocarejo et al., 2014). As outcome the population density in low-income zones is around 23,100 inhab/km\(^2\), 12,700 inhab/km\(^2\) in medium-income zones and 7,500 inhab/km\(^2\) in the wealthiest zones (Bocarejo et al., 2014).

Bogota’s population growth has been significant for the past 20 years. It has moved from around 5 million in 1990 to almost 7.2 million by 2010. Population projections by the National Administrative Department of Statistics (DANE) estimates that Bogota will grow at an average rate of 1.1% per year in the next 25 years. The city is divided into 19 urban communities which are in turn divided into 112 planning zones. These zones are named Unidades de Planeacion Zonal (UPZ) and determine by the planning department of Bogota. The purpose of the planning zones is to define and regulate urban land-use and management at a detailed scale (Bocarejo et al., 2013).

Nowadays the city urban growth trends present a concentration of low-income households in the south and west, meanwhile the high-income households are located in the north and CBD of the city (Bocarejo et al. 2013). Land scarcity to expand and high land prices outbound low-income population to peripheral zones of the city (Oviedo Hernandez & Dávila, 2016). Releasing land for development within the city jurisdiction is vital to avoid urban growth occurring on the surrounding municipalities of Bogota, especially
on Soacha. The Soacha municipality in the south of Bogota is notorious for rapid growth of informal settlements on lower-priced land resulting from the expansion of Bogota (Oviedo Hernandez & Dávila, 2016).

BoLD is a LUCC model conceived to address the need to understand global impacts on the urban development of transport infrastructure projects. The modelling was done using the Cellular Automata (CA) software Metronamica, developed by the Research Institute for Knowledge Systems (RIKS). As Fig. 2 shows, the model incorporates 6 municipalities in addition to Bogota, these are: Funza, Mosquera, Madrid, Facatativa, Cota and Soacha. The latest and Facatativa are zones where mainly low-income residential land-use is consolidated. Besides, the other municipalities have predominantly medium-income residential land-use. The model was calibrated using two complete datasets of land-use cover; for the baseline year (2005) and the calibration year (2014). Although the time lapse is just below what is recommended in the literature (10 years), the rapid growth experienced in Bogota during those years has resulted in enough amount of LUCC as to properly calibrate the model (Paez & Escobar, 2016). The spatial resolution used for the CA modelling was cells of 100x100 meters, which means that each cell covers an area of 1 hectare, and according on previous experiences of LUCC models this resolution is appropriate for cities with similar characteristics to Bogota (Escobedo, Clerici, Staudhammer, & Corzo, 2015).

The model was then used for predicting the urban patterns between 2014 and 2040, during four scenarios of transportation infrastructure: highway and train-based development with restricted and unrestricted natural reserve. It is important to acknowledge some limitations of the selected modelling; first, not all potential growth municipalities outside Bogota were included, especially those in the north such as Chia and Cajica, which are particularly attractive for high income residential areas. And second, the interaction among the municipalities on the west side of Bogota with others outside the study area are not included.

In general terms, a LUCC model based on CA requires inputs in the following areas: (1) Future land demands. (2) Current and future land zoning changes and suitability conditions. (3) Neighboring relationships between land-uses, and (4) Accessibility analysis based on transport infrastructure (Paez & Escobar, 2016). Regarding the estimation of future land demands for residential land-use, three major assumptions were made: i) High-income residential land-use will always correspond to 5% of the total cell of residential land-use. ii) Poverty will decrease 5% in 9 years and the medium income will increase at the same rate during that period. iii) Population densities
remain constant over time (Paez & Escobar, 2016). The latter assumption represents a modeling limitation and an opportunity to enhance the BoLD model, and is the main motivation for this research.

4.2. Spatial data exploration

Before initiating the model framework aforementioned, it is important to explain the spatial processes and tools used to build-up the explanatory variables employed in the regression modelling. All the spatial analysis was made based on BoLD land-use map of the year 2014, with special remark on the residential cells classified by the three income groups; described in greater detail in Paez and Escobar (2016) study. Also, the public transport network, main roads, parks and public space shape files for that year were considered in the analysis, in order to examine the influence of those infrastructures on housing built area.

As a first and exploratory analysis, difference in population density was depicted between 2005 and 2014 (See Fig. 3), in order to evaluate the zones were densification has taken place. Population density was measured using the number of inhabitants divided by the urban area for each UPZ. This concept of population density differs from gross density, that takes population divided by the total area, which is the sum of the private buildable areas plus the public land area (Paez, 2012). The data needed to determine the urban population density per UPZ was provided by Bogota’s Urban Planning Office (SDP). A hot spot analysis was performed to revealed the clustering pattern of high and low changes in population density.

**Fig. 3. Differences in population density: spatial distribution and hot spot analysis**
values. As Fig. 3 shows, high values (densification) are concentrated in the western zone of the study area, with the highest values located in the municipalities Funza and Mosquera which are adjacent to the western city border. On the contrary, the lowest changes are clustered along central-eastern zones of the city, where density mainly keeps steady over time.

The primary explanatory variables used for running regression models were measured or quantify from each 100x100 meters residential cell, these variables are; the public transport network (TransMilenio) coverage, measure as the quantity of km of trunk and feeder line per square kilometer for each residential cell; the main roads coverage, measure as the quantity of km of road per square kilometer for each residential cell; and the distance to parks and public space, measure as the average km from each residential cell to a polygon classified as park or public space. This last type of variable was determined using Euclidean distance tool in ArcGIS, which calculates the distance from a defined grid of cells to the closest source of parks and public space polygons. Then, each distance is aggregated with zonal statistics tool to each residential cell. The first and second type of variables were calculated using the ArcGIS Line Density tool that estimates the density of linear features in the neighborhood of each output raster cell, this density is calculated in units of length per unit of area. As Fig. 4 shows, public transport network and main roads coverage was determined for 2014, in order to calculate the coverage of these infrastructures on residential cells and then used them as variables for the regression model that helps to explain housing built area.

Fig. 4. Public transport and main roads network line density 2014
4.3. Exploratory regression analysis results

The explanatory variables obtained from the exploratory analysis were, public transport network line density; main roads line density; average distance from residential land-use to public transport stations; average distance from residential land-use to parks and public space and lastly average distance from residential land-use to first level commercial and industrial centralities. These variables were found by trying every combination between the potential explanatory variables, in order to achieve well-specified candidates OLS models. Three possible OLS models were found meeting the following threshold criteria: (1) Minimum acceptable Adjusted \( R^2 = 0.3 \); (2) Maximum Coefficient p-value Cutoff = 0.05; (3) Maximum VIF Value Cutoff = 7.5, this value reflects how much multicollinearity (redundancy) among the explanatory variables will be tolerated, values higher than 7.5 indicate multicollinearity can make the model unstable; and (4) Minimum Acceptable Jarque-Bera p-value = 0.1, smaller p-values might indicate the model residuals are not normally distributed and the model could be biased.

4.4. Ordinary Least Squares (OLS) results

The best specified model found according to the criteria aforementioned is the following:

\[
\log BA_{(i,j)} = \beta_0 + \beta_1 \times \log RP_{(i,j)} - \beta_2 \times \log RPT_{(i,j)} + \beta_3 \times \log RMR_{(i,j)}
\]

where \( BA \) is the residential built area; \( RP \) represents the average distance to parks and public space from residential cells; \( RPT \) is the public transport network line density coverage measure from the residential land-use; and \( RMR \) represents the main roads line density coverage measure as well from the residential land-use.

Table 1 and Table 2 show the diagnostics tests for the OLS model. The three explanatory variables are statistically significant with a p-value (p<0.01) and do not present multicollinearity (redundancy) among them based on the fact that VIF value for each value is less than 7.5. The Weld statistics and Joint F-Statistics, which represents the Fisher test (F-test) on regular statistical and econometric models are all significant, suggesting the existence of overall model significance accompanied on the fact that the Koenker (BP) Statistic is statistically significant, or the model is non-stationary, which means that the spatial process and relationships represented by the explanatory variables change spatially in the study area. Regarding the Jarque-Bera statistic, as it is statistically significant it indicates the model is biased due to the residuals are not normally distributed, which can be caused by the residuals spatial autocorrelation, and result in a misspecification in the OLS model. To verified residuals spatial autocorrelation, the Spatial Autocorrelation (Moran’s I) tool was run on the regression residuals (see Fig. 5), to find that they are spatially random, thus there is no problem with the model specification.

![](image.png)

**Fig. 5. Spatial Autocorrelation (Moran’s I) index on the regression residuals**
### Table 1
**Estimated OLS model parameters**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>t-Statistic</th>
<th>p-Value</th>
<th>VIF</th>
<th>Standardized Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.255***</td>
<td>-8.07</td>
<td>0.00071</td>
<td>----</td>
<td>0</td>
</tr>
<tr>
<td>Parks and public space distance</td>
<td>0.110***</td>
<td>8.20</td>
<td>0.00005</td>
<td>1.007</td>
<td>0.066</td>
</tr>
<tr>
<td>Public transport network line density coverage</td>
<td>-0.102***</td>
<td>-10.42</td>
<td>0.00081</td>
<td>1.024</td>
<td>-0.085</td>
</tr>
<tr>
<td>Main roads line density coverage</td>
<td>0.235***</td>
<td>20.38</td>
<td>0.00043</td>
<td>1.019</td>
<td>0.165</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses

*** p<0.01

### Table 2
**Descriptive statistics for the OLS model**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint F-Statistic</td>
<td>187.76</td>
<td>Prob(&gt;F), (3, 112) degrees of freedom: 0.0000*</td>
</tr>
<tr>
<td>Joint Wald Statistic</td>
<td>588.08</td>
<td>Prob(&gt;chi-squared), (3) degrees of freedom: 0.0000*</td>
</tr>
<tr>
<td>Koenker (BP) Statistic</td>
<td>187.64</td>
<td>Prob(&gt;chi-squared), (3) degrees of freedom: 0.0000*</td>
</tr>
<tr>
<td>Jarque-Bera Statistic</td>
<td>1659.94</td>
<td>Prob(&gt;chi-squared), (2) degrees of freedom: 0.0000*</td>
</tr>
</tbody>
</table>

* Statistically significant p-value (p<0.01)

### 4.5. Geographically Weighted Regression (GWR) results

In contrast to estimations using global models, local models represent each local area. Particularly, this GWR model explains the impact of characteristics from each 100x100 meters residential cell on the housing built area for each define polygon on the study area. Furthermore, the GWR model can clearly discern the level of impact of the explanatory variables in each one of the geographic zones. To compare the GWR results with those by global OLS model, the same three significant independent variables from the global linear regression were adopted as explanatory variables. Thus, Fig. 6 and Fig. 7 show the model performance and predictive spatial pattern, indicating that GWR model explains a 33% more the housing built area for each cell than simply assuming the average constructed area.

### Table 3
**GWR Statistics**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neighbors</td>
<td>422</td>
</tr>
<tr>
<td>Residual Squares</td>
<td>7325.59</td>
</tr>
<tr>
<td>Effective Number</td>
<td>410.77</td>
</tr>
<tr>
<td>Sigma</td>
<td>0.71</td>
</tr>
<tr>
<td>AICc</td>
<td>32475.54</td>
</tr>
<tr>
<td>R²</td>
<td>0.489</td>
</tr>
<tr>
<td>R²-Adjusted</td>
<td>0.475</td>
</tr>
</tbody>
</table>
Model statistics and estimations (local R² and coefficients) are shown in Table 3 and Fig. 8 - Fig. 11 to visualize the spatial non-stationary relationships identified by GWR. It can be seen that the local R² present great spatial variations, implying that the exploratory ability of GWR model varies within each residential cell, and the predictive power is concentrated along TransMilenio BRT system trunk lines and on the northeast region where the greatest housing built area is located. The standardized residuals spatial distribution shows that over and under predictions are randomly distributed and there is no statistically significant clustering of high and/or low residuals that might indicate that the GWR model is not well-specified (Fig. 8). The estimated parameters and identified relationships vary across space, indicating that GWR might be a proper tool to analyze the key factors influencing spatial changes on residential built area in the future.

![GWR model observations vs. predictions](image)

**Fig. 6. GWR model observations vs. predictions**

![GWR Spatial distribution for residential built area: observed and predicted 2014](image)

**Fig. 7. GWR Spatial distribution for residential built area: observed and predicted 2014**
As Fig. 9 shows, raster coefficient surfaces among the study area depict significant high and positive correlations for the public transport network coverage over the residential land-use, mainly among the trunk line corridors and the zones covered by the feeder lines. This behavior is desirable and expected due to TransMilenio BRT System since its implementation over 15 years ago has increased the housing built area on areas near its trunk lines. The only two TransMilenio trunk lines that do not show the same behavior are “Calle 80” and “Suba”, located in the northwestern region of the city. The overall effect of this variable in the GWR model goes according with urban and mobility dynamics in the city and also is able to characterized the zones where there is a negative correlation with housing built area illustrated by the blue areas in the raster map. Regarding the influence of the main roads line density coverage, the raster coefficient surface shown in Fig. 10 portrayed the positive association of the main road corridors over the central and northwestern zones of the city where there is a high line density of roads. Besides, a negative contributor is illustrated over the western peripheral zones and some specific spots in central regions where mostly low-income residential land-use is settle down. Lastly, the raster coefficient surface shown in Fig. 11 depicts the positive correlation that exists over the central region between the residential built area and the distance to parks and public space. Likewise, the strength of this variable is highly notorious on two spots; one located in the west, where a major park is on the boundary of the urban perimeter, and the other one positioned on the south region where there is lack of green and public space. All these findings demonstrate the spatial non-stationary relationships between the housing built area and the local characteristics of each residential cell.

Fig. 8. GWR Spatial distribution for standardized residuals and local $R^2$
Fig. 9. GWR Spatial distribution for public transport network coverage and raster coefficient surface.

Fig. 10. GWR Spatial distribution for main roads network coverage and raster coefficient surface.
The main objective of the GWR model was to develop a local linear equation that could be able to estimate built area for future residential cells modeled in defined scenarios with BoLD model. Once a well-specified local regression model is achieved as shown previously, it is time to use its predictive power. Taking the raster coefficient surface for each explanatory variable and estimating the independent variables with the line density coverage provided by the forecasted infrastructure transport and road projects, it is possible to assign the corresponding values for each future residential cell and calculate the foreseeing built area in 2040 (See Fig. 12). The expected transport infrastructures considered for the predictive analysis were the metro first line, three new TransMilenio trunk lines, and the west Regiotram system. Regarding the road infrastructure, all the expected and stipulated highways and roads on the land-use regulation or POT (Plan de Ordenamiento Territorial) were considered.
Table 4
Housing built area cells statistics: 2014 and 2040

<table>
<thead>
<tr>
<th></th>
<th>#Cells 2014</th>
<th>%Cells 2014</th>
<th>#Cells 2040</th>
<th>%Cells 2040</th>
<th>Built Area Density</th>
<th>#Cells 2014</th>
<th>%Cells 2014</th>
<th>#Cells 2040</th>
<th>%Cells 2040</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>High Built Area</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Density</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Medium Built Area</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low Built Area</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-Income</td>
<td>1359</td>
<td>9.05%</td>
<td>1926</td>
<td>7.89%</td>
<td>1007</td>
<td>74.10%</td>
<td>1266</td>
<td>65.73%</td>
<td></td>
</tr>
<tr>
<td>Residential</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>346</td>
<td>25.46%</td>
<td>545</td>
<td>28.30%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>0.44%</td>
<td>115</td>
<td>5.97%</td>
<td></td>
</tr>
<tr>
<td>Medium-Income</td>
<td>9007</td>
<td>59.97%</td>
<td>18507</td>
<td>75.81%</td>
<td>1412</td>
<td>15.68%</td>
<td>1897</td>
<td>10.25%</td>
<td></td>
</tr>
<tr>
<td>Residential</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5058</td>
<td>56.16%</td>
<td>7298</td>
<td>39.43%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2537</td>
<td>28.17%</td>
<td>9312</td>
<td>50.32%</td>
<td></td>
</tr>
<tr>
<td>Low-Income</td>
<td>4652</td>
<td>30.98%</td>
<td>3978</td>
<td>16.30%</td>
<td>275</td>
<td>5.91%</td>
<td>130</td>
<td>3.27%</td>
<td></td>
</tr>
<tr>
<td>Residential</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1052</td>
<td>22.61%</td>
<td>1942</td>
<td>48.82%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3325</td>
<td>71.47%</td>
<td>1906</td>
<td>47.91%</td>
<td></td>
</tr>
</tbody>
</table>

Based on Fig. 7 and Fig. 12 housing built area statistics for modeled 2014 and predicted 2040 residential cells were elaborated with the scope to display the percentage changes and differences among the three groups of residential cells which at the same time are classified by the resulting built area density considering each cell area of 1ha. High built area density cells were defined as those with a constructed area greater than 10,000m² per hectare; medium built area density cells were defined as those with a residential constructed area greater than 4,500m² and less than 10,000m² per hectare; and low built area density cells were defined as those with a constructed area less than 4,500m² per hectare. These intervals were determined according to the geometrical deviation of the housing built area distribution.

As shown in Table 4, the medium-income residential land-use consolidates on the study area as it cells increased over 15% from 2014 to 2040, while low-income and high-income residential cells diminished around 14% and 2% respectively. Therefore, it can be inferred that most of low-income residential cells are expelled from the urban city perimeter and located on the surrounding municipalities. Regarding the built area density, it can be figure out that high density is basically exclusive for high-income residential cells (65%), due to the percentage of this segment for medium and low-income residential cells is less than 10% and 3% respectively, in 2040. Lastly, medium and low built area density tend to increase in all residential cells from 2014 to 2040, with the exception of the 16% and 24% decrease of medium-income cells and low-income residential cells with a constructed area between 4,500m² - 10,000m² and less than 4,500m² per hectare respectively.

5. Discussion

For regions where there are micro-level demographic data limitations, such as Bogota, it is very important that model structures effectively incorporate all available data and information. Producing reliable estimates of population density for areas that lack direct samples is a problem of interest not only for urban planners but as well for topics such
as resource allocation, risk management, and smart cities analysis. Modeling estimations are vital for decision makers operating in regions limited by incomplete or unreliable census data, regardless of the lack of information from previous time periods that is needed for calibrating and validating models.

In the OLS log-log model presented in the results section, the coefficients can be used to determine the impact of the explanatory X variables on the housing built area which in this case is the dependent Y variable. The coefficients in this log-log model represent the elasticity of the Y variable with respect to each X variable. In other words, the coefficient is the estimated percent change in the dependent variable for a percent change in your independent variable, this means that if we change $X_i$ by 1%, $Y$ is expected to change by $\beta_i$ percent. Further, as depicted in Table 1, OLS model coefficients were standardized by converting them into standard deviations, so it is possible to compare the effect each explanatory variable on the residential built area. Interpretations of coefficients, however, can only be made in light of the standard error, due to they indicate how likely is to get the same coefficients when a resample of data and recalibration of the model is made for several number of times. In this case OLS model exhibited relatively small standard errors shown in brackets, indicating that coefficients would be fairly consistent.

Geographically weighted regression (GWR) model overcomes the limitations of OLS, by quantifying local-specific relationships. In this case, GWR produce a series of spatially varying coefficients for transport and road infrastructure line density and distance to parks and public space explanatory variables that help understanding the spatial non-stationary changes for housing built area in Bogota’s urban perimeter. The raster coefficients surfaces shown previously, are a very useful tool to visualize the effect and spatial contribution of each explanatory variable on the dependent variable. It is important to mention that zones that appeared to have a negative correlation or contribution produced by a certain independent variable on the residential built area, suggest that on these locations a specific explanatory variable might decrease residential built area or simply that particular variable has a positive effect on other land-use, such as commercial or industrial, rather than in residential land-use.

Although the main ambition of this research was to quantify changes in population density, it could not be achieved directly due to lack of micro-level population data. The last census was made in 2005 and nowadays demographic projections for Bogota have a UPZs geographical scale, so it was not possible to quantify population density for each 100x100 meters residential cell. Despite of this limitation, a global and local regression model was developed for examining housing built area at a 100x100 meters resolution, so then it could be used to estimate population density based on the average area per dwelling for each income combined with the average number of households per dwelling and the average number of persons per household. According to the Territorial Dynamics Observatory (SDP, 2012) the average area for low-income (Strata 1 and 2) residential dwellings is 49,53m$^2$, the average area for medium-income (Strata 3 and 4) residential dwellings is 72,37m$^2$, and the average area for high-income (Strata 5 and 6) residential dwellings is 124,03m$^2$. Combining these average dwelling areas with the information presented in Table 5 taken form the Multipurpose Survey on Households (SDP, 2014) developed by
Bogota’s Urban Planning Office in cooperation with National Statistics Department (DANE), it is possible to calculate the average population density for each 100x100 meters residential polygon or cell employed in the GWR model. Therefore, to determine the population density for each cell, the residential built area should be divided by the corresponding average dwelling area, depending of the income residential classification. Then, the result should be multiplied by the average number of households per dwelling respectively, and finally multiplied by the average number of persons per household. In this way, population density for each 100x100 meters cell could be spatially mapped throughout the study area.

<table>
<thead>
<tr>
<th>Socioeconomic Strata</th>
<th>Average # of households per dwelling</th>
<th>Average # of persons per household</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>CV</td>
</tr>
<tr>
<td>Strata 1</td>
<td>1.02</td>
<td>0.3</td>
</tr>
<tr>
<td>Strata 2</td>
<td>1.03</td>
<td>0.4</td>
</tr>
<tr>
<td>Strata 3</td>
<td>1.02</td>
<td>0.3</td>
</tr>
<tr>
<td>Strata 4</td>
<td>1.01</td>
<td>0.3</td>
</tr>
<tr>
<td>Strata 5</td>
<td>1.00</td>
<td>0.1</td>
</tr>
<tr>
<td>Strata 6</td>
<td>1.00</td>
<td>0.1</td>
</tr>
</tbody>
</table>

This study has some constraints. Firstly, both OLS and GWR coefficient results are mainly applicable to determine residential built-up areas. Secondly, the municipalities surrounding the city, were not considered in the study area due to lack of trustworthy built area and cadastral information from those municipalities. This might have an impact on densification processes inside the urban perimeter, because of the great interaction among the city and those municipalities. And thirdly, even though the regression models do not determine changes on population density, those changes can be estimated indirectly using the GWR built area model.

6. Conclusions

To better understand and quantify changes in population density, this paper respectively applies global (OLS) and local (GWR) regression models to identify the significant determinants between housing built intensity and transport and urban oriented variables for residential land-use cells considered in BoLD model. Even though population density was not directly determined by global or local regression models, a methodological approach was established taking into account a well-specified GWR model that allows to calculate indirectly; through housing built area, the spatial patterns of population density over the study area. In terms of explanatory power, this model performs well along TransMilenio trunk lines and on the northeast region where the greatest built area is located, suggesting that local coefficient estimations have a high predictive power on those regions of the study area. In this research, variables of the global linear model were maintained in GWR for comparison purposes. Future studies can alter the variables or increase the variety of variables, considering floor area ratios, jobs density and land value.

This research confirms the notion that housing built area can be generally associated with local transport and road infrastructure characteristics, as well as with distance to parks and public space. This let us conclude that these variables might be major contributors on processes of densification on the inside urban perimeter of the city of Bogota.

Further research is needed to better understand densification processes in the context of sustainable development for the city of Bogota and its region.
References


