ORIGINAL PAPER

Modeling population density guided by land use-cover change model: a case study of Bogotá

Luis A. Guzman[1] · **Ricardo Camacho[2]** · **Arturo Rodriguez Herrera[3]** ·
Carlos Beltrán[1]

Abstract

Population densities provide valuable spatial information to identify populations at risk, quantify mobility, and improve our understanding of future urban settlements. Advancements in machine learning algorithms open up new horizons to face these challenges. This research proposes a supervised machine learning approach, Random Forest, for population density appraisal in a large and dense developing city. We studied Bogotá, where functional integration with neighboring municipalities exists, although they have different governments and uncoordinated urban development plans. As a starting point, we use simulated residential land-use patterns, classified according to socioeconomic levels, from a cellular automata-based model. We estimate population density with reliable land-use change models and nine simple representations of the urban structure, such as land values and the distance to urban amenities. Therefore, combining a cellular automata model with a classification model, considering both continuous and categorical variables, demonstrates this methodology's potential and promises a reliable assessment of population density. Finally, we present a trip generation model integrated with densities and spatial location. A comprehensive results discussion suggests this study's importance in urban planning and the accuracy of the proposed methodology to support decision-making processes and policy evaluation.

✉ Luis A. Guzman
la.guzman@uniandes.edu.co

Extended author information available on the last page of the article

Introduction

Population estimations have a long history in the social sciences in supporting and informing decision-making (Archila Bustos et al., 2020) and planning for the future. Existing methods typically produce population estimations in a large-scale way (Gerland et al., 2014) using three major approaches: the cohort component method, trend extrapolation methods, and structural models (Smith et al., 2013). However, the growing interest in small areas population analysis highlights the importance of disaggregated projections (Chi, 2009; Raymer et al., 2012), particularly as it relates to transport and urban planning. In addition, several factors influence urban population growth and density, generally in unpredictable ways, especially in developing cities. The rapid urbanization process makes it imperative to understand the main urban characteristics, such as land development and population growth, that play a crucial role in urban sustainability (Bassolas et al., 2019).

Despite a large body of evidence showing that a compact population causes a decrease in the costs of urban public services (Carlino et al., 2007; Fernández-Aracil & Ortuño-Padilla, 2016), reduces car use, and encourages the use of active transport modes (Boulange et al., 2017; Lewis, 2018), the benefits of urbanization and density are far from a settled issue. The growth of high-population urban areas in Latin America and Southeast Asia with an increasing population density has increased congestion and air pollution (Güneralp et al., 2020). The increase in urban population presents both challenges and opportunities: although density improves the efficiency of cities in many ways, it also can encourage crime, congestion and pollution, and more resources end up being used (Bettencourt et al., 2007). Understanding population growth and changing population densities is crucial to supporting territorial decision-making.

Urban population estimation is gaining importance due to its implications for transport planning, travel demand forecasts, urban facilities, and public space measures, which have become important development and well-being indicators of cities. Despite the growing demand for local-scale population projections, few disaggregate population projections exist in developing cities. Zonal-level population projections are typically only available through the (scarce) official data or private companies and frequently only are estimated under particular goals and scenarios. The lack of rigorous small-area population projections according to urban growth trends and land-use changes has hampered our understanding of urban and mobility patterns (Gao & O'Neill, 2020). Therefore, emerging technologies offer tools to access and process good-quality data that might improve our understanding of this urban phenomenon. Joining information from emerging data sources with simulation models is crucial to improving the capacity to forecast future scenarios (Guzman et al., 2021).

Consequently, innovative approaches are required to adapt to situations where limited information is available. This paper aims to estimate population density and travel demand involving spatial and transport infrastructure factors at a residential block level in Bogotá, Colombia. We are doing so using a Random Forest

approach (Breiman, 2001) jointly with a land-use change simulator, a cellular automata-based (CA-based) land use-cover change model (Guzman et al., 2020a). We use different land-use pattern scenarios simulated from the CA-based model. Since this tool is limited to a 2D analysis, meaning that the density was missing from its results, as it is in most CA-based models, this constitutes a research gap that motivates our study. Therefore, we propose a classification model (Random Forest), to estimate population density, using supervised machine learning algorithms with 2D land-use model based on CA, which incorporates residential land occupation processes' spatial and temporal dimensions. This hybrid tool's development will allow the prediction of the influence that a planned urban development and new transport infrastructures will have on Bogotá's residential population density and how this redistribution will affect the travel demand.

State of the art

Population density and travel demand are essential issues in urban planning. The estimation changes in the concentration of people in cities are relevant to a broad range of issues related to the quality of urban life, including economics, infrastructure, and transport systems. Notably, these topics are crucial in designing public transport projects that depend on population growth, land values, and urban structure. Depending on the context, generally, density is estimated based on census or GIS data (Liu et al., 2008), other studies led to a more general expression based on gamma models that considered the relationship between population and land use (Batty & Longley, 1994), or use a fractal structure approach to study the complexity of population density in cities (Chen, 2010), which has worked well in cities with population clusters. Also, recent developments in urban planning require more accurate population forecasts at small scales to deal with population growth, land-use changes, and environmental and travel demand effects (Chi, 2009). Well-established methods for estimating and forecasting population and population density at those scales exist (Chi, 2009; Chi & Voss, 2011; Chi et al., 2011; Smith et al., 2013). Regarding travel demand, also well-established methods exist according to the context, scale, and data availability (Ortúzar & Willumsen, 2011).

Nevertheless, in several urban areas, the availability of census, GIS, historical data, or any population data is one of the biggest challenges when estimating the density (Onda et al., 2019). This limitation is particularly challenging when the study area's information is unavailable or even when there is uneven availability of information in different administrative jurisdictions (Guzman, 2019). To tackle these problems, different methods, such as geostatistical (Liu et al., 2008; Wu & Murray, 2005) or radar images (Kajimoto & Susaki, 2013), are often utilized.

Since these population density formulations are based on empirical analysis and statistical data for particular areas, they cannot be entirely replicated among different places. Only a handful of studies have tried to do so, where Qiang et al. (2020) stand out with their Metropolitan Statistical Areas using travel times to city centers. A major impediment to estimating population densities is that the methodologies, parameters, and available data used may not account for

specific characteristics of developing cities and their specific contexts. Considering the great challenges in planning faced by the cities in the Latinamerican region (Sarmiento et al., 2021), the potential of the supervised machine learning approach such as Random Forest models to improve planning is not to be understated.

Random Forest (RF) is a robust classification and regression algorithm (Breiman, 2001). The RF algorithm allows us to predict a Y vector over an X's matrix of previously trained features. The RF algorithm creates a series of estimators, also called decision trees. Each of these estimators represents a weak decision, and all the uncorrelated estimators work together, making the model powerful (H2O.ai, 2021). The resulting tree average makes the global prediction of individual decisions made by each estimator. The global prediction uses a series of "trees" as estimators, which is why the algorithm is called "forest." The features used to train the model can be both categorical and continuous. This characteristic positions the RF regression usefulness above Ordinary Least Squares (OLS) and Geographically Weighted Regression (GWR) when dealing with big data (Jiao et al., 2021). Other authors also reveal RF successfully handles high data multicollinearity, being insensitive to overfitting in remote sensing (Belgiu & Drăguţ, 2016).

RF usually are used to rank the importance of variables in a classification problem naturally. In the last few years, the ability and robustness of the RF models have been evaluated in different contexts and with different purposes. In this context, Rahmati et al. (2016) developed an RF model for mapping potential groundwater sources in the Mehran Region (Iran), or Khosravi et al. (2018), and their flash flood susceptibility mapping model. Recently, research has been published where RF was used in urban contexts, such as the classification of urban areas into residential regular vs. irregular settlement types (Jochem et al., 2018), the classification of urban green spaces (Puissant et al., 2014), or a creation an application of a high-resolution population grid from district-level data in India (Onda et al., 2019). Another example is the combination of RF and CA in a model to simulate urban growth in Zimbabwe highlights the potential of these tools for these tasks (Kamusoko & Gamba, 2015). This RF-CA combined model showed that RF improved the CA model's usefulness and potential to improve urban growth modeling.

The use of RF models to estimate gridded population models has grown in the last few years. Such is the case of Stevens et al. (2015), who developed an RF model for mapping the population in a country-wide pixel-level map, the development of a high-resolution population distribution map using ancillary data RF predictions (Gaughan et al., 2013), and the modeling and prediction of the population at the grid level (Sinha et al., 2019). Consequently, using a unique dataset collected in Bogotá, we attempt to investigate the effect and relative importance of the selected urban attributes at different scales concerning affecting population density estimation. Then, we also estimate the travel demand produced by this population density.

This estimation is crucial in developing cities, where city activities change faster than in developed cities (Cervero, 2013). In terms of the urban environment and living conditions, periphery and central areas experienced unequal living standards for a wide range of population densities. Such is the case in Bogotá, where high-density areas present more disadvantageous conditions than low-densities (Guzman

& Bocarejo, 2017; Guzman et al., 2017b). Additionally, the availability and accuracy of data are not uniform across urban areas and years.

Study area and data

As the most populous city in Colombia and one of the largest and densest on the continent, Bogotá has experienced rapid changes in land use and rapid urbanization beyond its original geographic jurisdiction boundaries. Over the past few decades, migration and population growth rates have encouraged people to occupy unplanned and informal urban settlements on the urban periphery characterized by high population densities (Guzman et al., 2017a). This growth process has segregated the population, with the city center and east edge occupied primarily by wealthy people (and economic opportunities) and the periphery by the poor. Besides, Bogotá has not been able to contain its growth within its existing boundaries, with mainly the low-income population spilling over into neighboring municipalities while restricting land development within its limits. This particular spatial pattern encourages uneven urban living conditions, negatively affecting the quality of urban life, and lengthens travel distances.

Besides economic land uses, residential areas are traditionally divided into six categories according to socioeconomic and urban characteristics. Bogotá is divided into homogeneous physical and socioeconomic residential areas, locally known as socioeconomic strata (SES). SES 1 corresponds to the neighborhoods of the lower-income population and poorest urban characteristics, and SES 6 corresponds to wealthy neighborhoods. This classification also represents the characteristics of the built environment and is considered an acceptable proxy for income (Cantillo-García
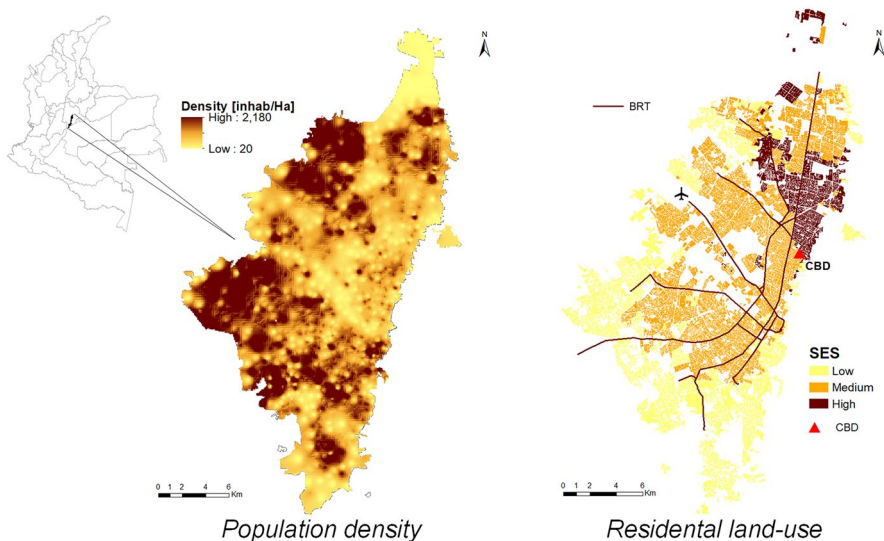

Fig. 1 Population density and residential land-use categories

et al., 2019). In this case, we used three categories based on SES: low-income residential areas (SES 1 and 2), medium-income areas (3 and 4), and high-income areas (5 and 6), as shown in Fig. 1.

As this research is a step further from the CA model developed by Guzman et al. (2020a), we considered the same data sources. This data includes urban attributes at different scales. Since population density is highly context-dependent and could be influenced by economic, sociodemographic, and regulatory factors, we have focused on economic (land price) and physical factors related to urban structure and divided them into four categories.

First, we collected data on the cadastral land value, population, and SES for Bogotá's 35,796 residential blocks in 2016 from the Bogotá Urban Planning Office (SDP). Hence, we have the residential cadastral land value (COP per square meter), the calculated population density for each block, and the SES category for each residential block (see Fig. 1). Second, there is data about the location of critical facilities such as educational services, health services, public parks, and the CBD from the city's Spatial Data Infrastructure (UAECD, 2019). Third, we have the locations of transport infrastructure, such as the main road network, Bus Rapid Transit (BRT) stations, and feeding routes, also from UAECD. Fourth, we gathered the resulting land-use cover simulations from a CA-based land-use change model developed in previous research by Guzman et al. (2020a).

To estimate travel demand, we use the 2019 Household Travel Survey (HTS) to gather Bogotá's travel patterns. The trip generation analysis considered the approximately 16,099,700 trips generated in Bogotá on a typical day before the pandemic. One of this study's objectives is to predict the home-based travel demand based on density estimations, so the model excluded 7,152,100 trips for "returning home." Thus, the trip generation model was calibrated with 8,947,600 trips. To avoid calibration errors, we did not include the trips outside the initial training dataset.

Methodology and procedures

The country-wide pixel approach of the RF models used in previous research aims to understand big-scale behavior. However, these approaches lack precision in small-scale predictions due to the ecological fallacy that pixel-level predictions introduce when a country's census data is not detailed enough. This ecological fallacy is based on the assumption that small-scale features behave like big-scale features (Sinha et al., 2019). The fallacy arises when the data inside a feature (e.g., a residential block) has a unique value within the block, which is often the case with census data. The solution to this fallacy is to use the predictions of the RF model as a weight layer to a dasymetric mapping scenario (Stevens et al., 2015). The approach of this study is to use geographically referenced data of the built environment for predicting the small-scale gridded population density. In this case, the data is accomplished by a combination of distances, census, and cadastral data at the residential block level, so a dasymetric mapping technique is not needed for the distance calculation.

Thus, we describe the data processing and empirical strategy used. The prediction model's validation should be appropriately performed to obtain sufficient land value

and population density prediction accuracy. There are three stages in this proposed methodology: the CA land-use scenarios, the RF model specifications (land value and population density), their model validation, and the travel demand model. The first component is a CA-based land-use model used to estimate different 2D land-use pattern scenarios. Then, land values and population densities will be estimated using RF models. Finally, the simulated scenarios serve to estimate travel demand.

Data processing

After gathering the data, we established ten urban variables, nine of which are explicative, one of which is a dependent variable. After a Variance Inflation Factor (VIF), a Moran Index, and a collinearity analysis, summed with the mix of continuous and categorical variables, it was determined that none of these variables have redundancy among them and that neither OLS nor GWR was an option for this study. The selected variables and data statistics values are summarized in Table 1, where Land Value (LV) is in thousand COP per square meter. The distances are all in meters, and the population density is in inhabitants per hectare.

Additional to this, we have another variable: Socioeconomic strata (SES), which is a categorical variable with three categories that vary according to household income (low, medium, and high). Every pixel in the study area has all the attributes mentioned earlier (Table 1).

We also developed an ArcGIS toolbox that calculates the mean nearest distance between a single objective layer and a series of grid vector layers. The complete process also has some valuable corrections to improve the performance of the toolbox, but they are not relevant to this study. The toolbox was designed for use in general situations, consisting of a table filled with distance fields. Figure 2 illustrates this procedure and summarizes the CA-based model integration that came before this research.

Table 1 Variable statistics

| Explicative variables		Mean	Standard deviation	Min	Max	Median
Land value	[LV×1000]	1,156	805.0	6.0	12,000.0	1,096.1
Distance to CBD	[CBD]	10,705.8	4903.9	0.0	22,721.4	10,833.9
Distance to BRT feeder lines	[FEED	852.1	1,377.5	3.9	8,315.5	237.4
Distance to zonal public parks	[PP]	1,321.5	866.6	1.9	4,564.6	1,150.5
Distance to BRT stations	[BRT]	1579.3	1143.5	31.8	6,888.8	1,331.9
Distance to the main road network	[MRN]	648.0	744.4	4.3	3,730.4	329.4
Distance to health services	[HS]	701.5	466.4	19.5	6,589.8	602.8
Distance to educational services	[ES]	188.8	133.5	7.7	1,778.6	157.4
Population density (dependent variable)	[PD]	620.9	427.4	0.02	5,701.6	550.8
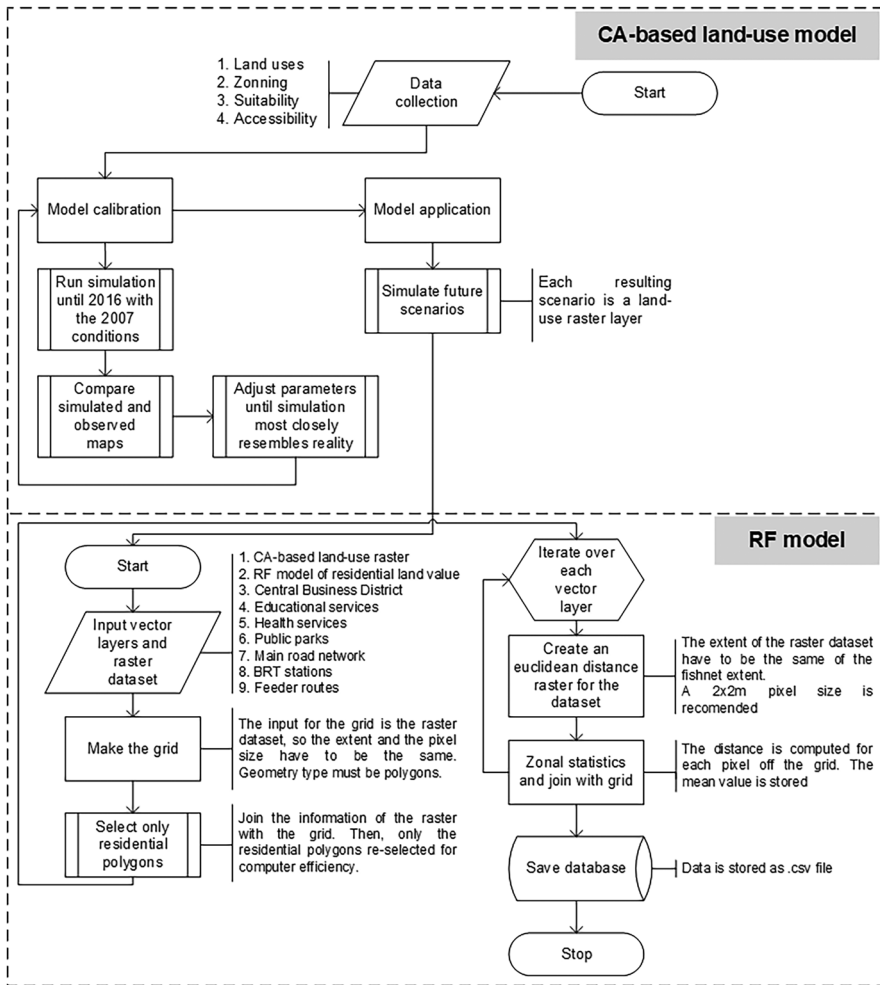
Fig. 2 Flowchart for the generation of the definite input table

Regarding the travel demand model, as trips reported by the HTS are at the household level, while the explanatory variables are at the block level, the analysis unit was homogenized. Thus, having each sampled household georeferenced, each trip was assigned to a block through a spatial joint using ArcGIS software's spatial analysis tool. This restructuring resulted in a database of 5,523 blocks with information on the number of trips generated, SES, population density, and the distance to the CBD. This aggregation assumes that the representativeness of each block coincides with that of the trips.

CA model description

The Bogotá CA-based land-use model was developed and calibrated using the CA-based model utilized in Metronamica® (van Delden & Vanhout, 2018). The results obtained from this model consist of a 60×60 m pixel raster, where for each pixel, there is an integer value with a land-use assigned to it. This pixel size was used due to the available satellite images' resolution and the cadastral information available in Bogotá to assign the land-use category to each pixel (Guzman et al., 2020a). Each pixel's state change potential is calculated in discrete steps while following a set of neighborhood rules that depend on spatial features such as neighborhood potential, proximity to transport infrastructure (accessibility), zoning, and suitability (Guzman et al., 2020b).

Residential land use has an assigned location based on the change potential derived from the transition rules mentioned before. At each moment, residential land use is located in places with the most significant potential for development. The model was calibrated using historical data, which establishes the model's suitability to reproduce current land-use dynamics (Zheng et al., 2015). Then, by adjusting the transition rules between accessibility and land-use types, the model attempts to reproduce the land-use dynamics of the studied area. Finally, the assessment of model accuracy using Kappa indices showed a substantial agreement between simulated and real land-uses, calibrating the model (Guzman et al., 2020a). This tool provides a dynamic modeling environment that can simulate 2D land-use changes over time (van Vliet et al., 2012). In this case, this tool was used to obtain two simulated land-use patterns, as explained before.

Since the available data was structured in residential blocks and the resulting land-use cover predictions are in a 60×60 m raster (pixel) layer, we recomputed the available data into squared polygons that are adjusted with the raster layer. This data restructuring meant that we had to recalculate every feature for each pixel. The raster has information on land use itself, so the SES is easily determined. The residential land value feature will change depending on the land-use distribution. We developed a second RF model based on the cadastral land value's importance in the population density model to predict the cadastral land value for each pixel. We needed a second RF model because the cadastral land value data for projected scenarios will never be available.

Random forest model structure

The output RF model resolution was selected based on preliminary developments using the CA model described earlier. The supervised classification uses an RF approach and relies on multiscale feature indices calculated from the observed population density distribution and nine explicative variables (Table 1). RF is a non-parametric ensemble-based prediction model, with a robust classification and regression algorithm. The RF algorithm creates a series of estimators. Each of these estimators represents a weak decision, and all of the uncorrelated estimators work

together, making the model powerful. The average represents the global prediction of individual decisions made by each estimator.

The features used to train the model can be both categorical and continuous. This flexibility allows the use of both types of variables (categorical, such as SES, and continuous such as land value). We chose Mean Absolute Error (MAE) and Root of Mean Squared Error (RMSE) as validation metrics because both are suited for the land value and population density RF models (Stevens et al., 2020). The advantage of using MAE and RMSE is that the metric units are the same as the predicted variable. We can compare the metric directly with the basic statistics of the predicted variable. We used these statistics to determine the best parameters that make the model error acceptable. These metrics are helpful in regression problems because the cost associated with an error in the model will be higher when the difference between the prediction and the real value is higher. We iterated different proportions of validation/training in our database until we found the best proportion for training the model. The data split was chosen due to the number of hyper-parameters that RF models require.

Another essential factor that needs to be modified in the RF model is the categorical variable encoding. The encoding is how the model converts string values to numeric values. The loss functions (metrics) will always work with numbers, so a string will result in an error that will stop the training process. There are several ways to transform from categorical strings to numbers. We applied the method most commonly used in machine learning: *One Hot Encoding*. It enumerates the different strings in a dataset and then converts each enumeration class into a new Boolean variable. The SES variable was converted into three different dummy variables in this specific case because SES can only take three different exclusive values (high, medium, or low).

Once the models are trained, they can perform predictions on any dataset using population density or land value. However, it is essential to validate the results using the validation dataset. The objective is to compare the observed dataset to the predicted one. The validation of the models is made by comparing descriptive statistics. As the number of trees increases, the error metrics will decrease for the training dataset. We can further create a model with the highest number of trees, resulting in a minimal error on the training dataset. This process could lead to overfitting in the model, and the algorithm will not recognize trends. To avoid overfitting in the model, we stopped the training processes when we found convergence in the validation dataset. When the validation dataset is used, the error will not decrease significantly. This procedure was used for both the land value and the population density RF models.

The cadastral land value is available in the training dataset, and we used it to train the population density RF model. However, in simulated scenarios "Proposed scenarios", the land value comes from a second RF model. Thus, this model's explicative variables were the same as described earlier. Having the georeferenced blocks, we calculated the distance to the closest explicative variable for each of them.

Travel demand model structure

The travel demand estimation associated with the densities and their locations is calculated based on an OLS model. We adjusted the OLS estimators calculated per pixel to be consistent with the units of analysis, to perform the trip estimation based on each 60×60 m pixel of the population density RF model raster. This adjustment is based on dividing each estimated parameter by the number of pixels within Bogotá corresponding to a residential block. This methodology for rescaling estimators is valid because the model standardizes the dimensions of the analysis units by including the density as an explanatory variable. The number of trips in each pixel is predicted according to its characteristics with the rescaled parameters.

We tested several model specifications to obtain the best fit for expected signs, magnitude, and statistical significance. The proposed model analyzes the relationship between the trip rate generation per pixel and its density by SES and the corresponding distance to the CDB. This specification is a quadratic model that considers the interaction between density by SES and how it is controlled by the distance of each pixel to the CBD. The following equation describes the model:

$$T_i = \beta_0 + \beta_1\left(hSES_i * PopD_i\right) + \beta_2\left(mSES_i * PopD_i\right) \\ + \beta_3 PopD_i + \beta_4 \mathrm{CBD} + \beta_5 CBD^2 \tag{1}$$

where T_i is the trip produced in each residential pixel i, $PopD_i$ is the population density per pixel i, CBD_i is the distance of each pixel i to the CBD (Fig. 1), $hSES_i$ and $mSES_i$ are dummy variables that correspond to each residential land-use category (high and low, respectively), with $lSES_i$ (low) as the base category. For instance, the density effect on high-SES trips is $\beta_3 + \beta_1$.

Proposed scenarios

Two simulated scenarios are studied and processed as input data for the RF models. These scenarios represent the projected land-use demands in the study area, and for this case, we consider only the residential land-use distribution results for the year 2050. The two scenarios considered in this research and their characteristics are summarized below:

- Scenario 1: Bogotá keeps its growth delimitation. Full restrictions to prevent residential use on agricultural land are applied. There are neither significant road infrastructure changes nor new developments in the public transport infrastructure under this scenario.
- Scenario 2: Bogotá (especially in the north) allows residential developments, enabling suburban land expansion. Restrictions on the occupation of agricultural capacity are eliminated. New BRT lines, the first metro line, and two regional train corridors are created, improving the regional road network.

Results and analysis

The results presented in this section summarise the main outputs obtained from the proposed methodology. The model parameters had to be adjusted to find the best combination to tackle the population density and land value problem. Based on the data, we evaluated several RF parameters to determine the best training/validation ratio. We found that a ratio of 80% training and 20% validation gives the best RMSE value. Then, the integration between the former and the CA-based models is presented, including the study of two scenario results. We used the same parameters for the land value RF model.

We determined a reasonable number of trees for which the computational cost would not be too high without directly affecting the outputs and avoiding overfitting. Once the best parameters were determined, we trained both RF models with their corresponding observed datasets. The parameters for the model were:

- Number of trees: 2000
- Depth (range): [1, 20]
- Random variables: Random for each tree, minimum 1, maximum 9
- Minimum rows: 1

Due to the randomness of the algorithm, the random value for the random seed was set to 503,080. This number was set to a fixed value to be able to compare different models. After determining the best parameters for both RF models, we executed the Land Value RF model so that its output could be an input for the Population Density RF model, as described before.

Population density estimation

The models were trained with a sample of our datasets. We found that the best proportion for training the model is 20/80. Using the Land Value RF model results and the explicative variables described earlier, we trained the Population Density model and then, proceeded with the validations. This process was made by comparing the observed population density of each pixel against the predicted density. We found that the population density model is especially sensitive to the number of trees, so we decided to iterate over the number of trees until we found convergence.

Figure 3a shows the accuracy of the Population Density model. As seen, the model can predict with high accuracy; however, the regression slope in the graph shows that in most cases, the model is sub-estimating the population density. There, some outer points do not follow the regression correctly. In cities like Bogotá, where a significant part of the peripheral urban development had an informal origin (Guzman et al., 2017a), implying a lack of planning and control, density estimation can be tricky. This unusual behavior can be explained as a limitation of the models' extrapolation capabilities (Chi et al., 2011), followed by the larger-than-Bogotá area created by considering the adjacent municipalities,
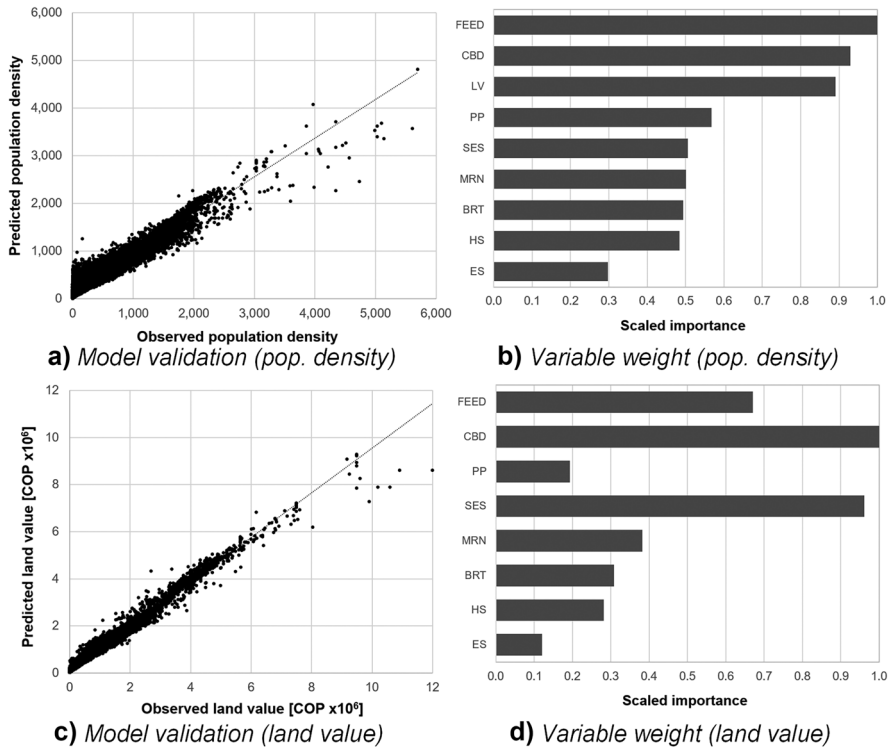
Fig. 3 Validation of the population density and land value RF models and variable importance. Variable acronyms as described in Table 1

which we are doing. Once the trained model was validated, we checked each of the explicative variables' weights, as shown in Fig. 3b. The variables' weight allowed us to better comprehend their impact on population density estimation. This variable importance led to a variable analysis, avoiding redundancy.

The validation for the Population Density RF model resulted in an MAE (155.70), RMSE (244.27), and R^2 (0.673), which shows the accuracy of the trained model (Fig. 3a). RMSE is only 0.91 standard deviations from the average value of all pixels in the resulting model. This acceptable performance is undermined by the extrapolation errors caused by the surrounding municipalities. When the Land Value RF model was validated, we obtained an MAE (119,713.55) and RMSE (232,947.32), giving us insight into the predictive model's accuracy. Other descriptive statistics, such as the R^2 value of 0.916, confirm this. The median of the land value in the validation dataset is only 1.05 standard deviations from the RMSE.

From the observed population density of each of Bogotá's 35,796 residential blocks from the Bogotá Urban Planning Office (SDP), we calculated a population of 5.8 million inhabitants, which was the starting data point. Since the projections were made for the year 2050, there are no official estimates at the aggregation

level of this study. At the aggregate level, official projections made in 2018 by the Bogotá's SDP estimate a population of approximately 11 million inhabitants by the year 2050 (SDP, 2018). The SDP projections are compatible with this study's estimates of 11.4 million inhabitants for Scenario 1 and 12.9 million inhabitants for Scenario 2.

When the model tries to extrapolate over the training dataset, the return value will be the value of the last node split. This value is determined in the algorithm implementation. The more variables the model has, the smaller the impact of extrapolation. For instance, in a model with ten variables where only one variable is extrapolating, the results will be influenced by the variability of the other nine variables.

Travel demand estimation

The final result was set from interactions of the categorical variables of SES with the population density, using the low-SES variable as a base, as shown in Eq. (1). We added each SES coefficient to the base coefficient to identify the total effect of density on medium-SES and high-SES trips. Table 2 presents the aggregate effect of density in each SES and the result of the joint significance tests, showing all the coefficients are statistically different from zero at the 1% level of statistical confidence.

This model indicates a positive relationship between population density and the number of trips generated in each pixel, i.e., the density and number of trips are more strongly correlated as SES increases. The effect of the distance to the CBD is positive but decreases as the distance increases because of the negative coefficient of the CBD variable in its quadratic form. The results show that the greater the distance to the CBD, the more trips are produced. However, after 13 km, the effect of

Table 2 Travel demand model results

Variables	Model
β_0 (constant)	350.0***
	(115,4)
β_1 (density in low-SES pixel)	0.484***
	(0.0764)
β_2 (density in medium-SES pixel)	0.712***
	(0.0966)
β_3 (density in high-SES pixel)	2.049***
	(0.407)
β_4 (distance to CBD)	0.176***
	(0.0244)
β_5 (distance to CBD-squared)	–6.77e-06***
	(1.17e-06)
Observations	5523
R-squared	0.032

Standard errors in parentheses

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$

Table 3 Total trips (model-observed comparison)	SES	Observed 2019	Model 2019	M/O ratio
	High	479,528	413,953	0.86
	Medium	4,062,324	4,307,825	1.06
	Low	4,405,762	4,225,835	0.96
	Total	8,947,614	8,947,614	1.00

distance on trip generation is zero. This implies that although a person who lives in a high-SES zone makes more trips on average, the lower-SES zones, located in the city periphery, have more significant trip generation due to the higher number of people living there.

Table 3 shows the observed trip generation values by SES (from HTS), the model results, and the ratio between them, complementing the observed trip production by SES with the fitted modeled values in the base year. As expected from the proposed model, the aggregate modeled values tend toward the observed values, given the normal distribution of the error term, with its mean equal to zero. The travel demand estimation process results were compared by SES level for the study area.

The results show that the travel demand model correlates well on an aggregate basis with observed data from the household survey and the data observed.
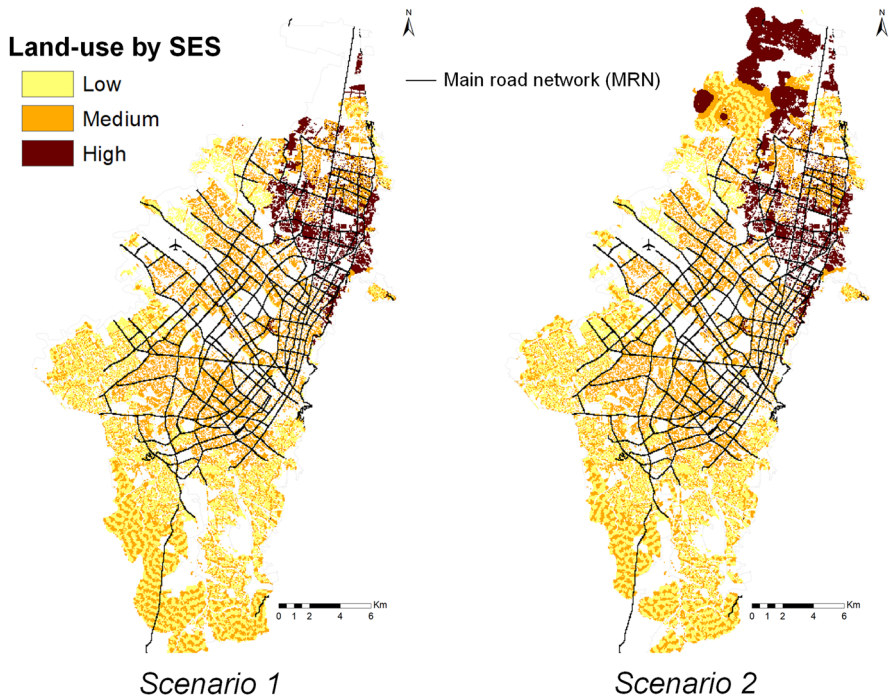

Fig. 4 Residential land-use test scenarios for the population density prediction model 2050

Scenario evaluation

The distribution of the projected residential land by SES level for both scenarios is displayed in Fig. 4. It is crucial to highlight a limitation of this model. Even though the CA model studies Bogotá and its surrounding municipalities, both the population density model and the trip generation model were trained and validated only for Bogotá due to data availability.

From these distributions of residential land use and transport infrastructure, we will apply the RF model to estimate land values and then, estimate population densities and travel demand. Figure 5 displays the population density assigned to each pixel for both of the proposed scenarios from Fig. 4. Both maps clearly show a redistribution of density in the city's periphery. In Scenario 2, the population density is distributed differently in the northern part of the city and discourages new urbanization in the extreme south.

We observe that the average population density in the high-SES areas increases for Scenario 1 and increases even more in Scenario 2. However, in Scenario 1, the population density is concentrated in a much smaller area, caused by land-use restrictions. In Scenario 2, we can see a steady incremental increase in the high-SES in both areas and average population density, which is accurate considering the new residential developments planned in the northern part of Bogotá under this scenario (see Fig. 6).
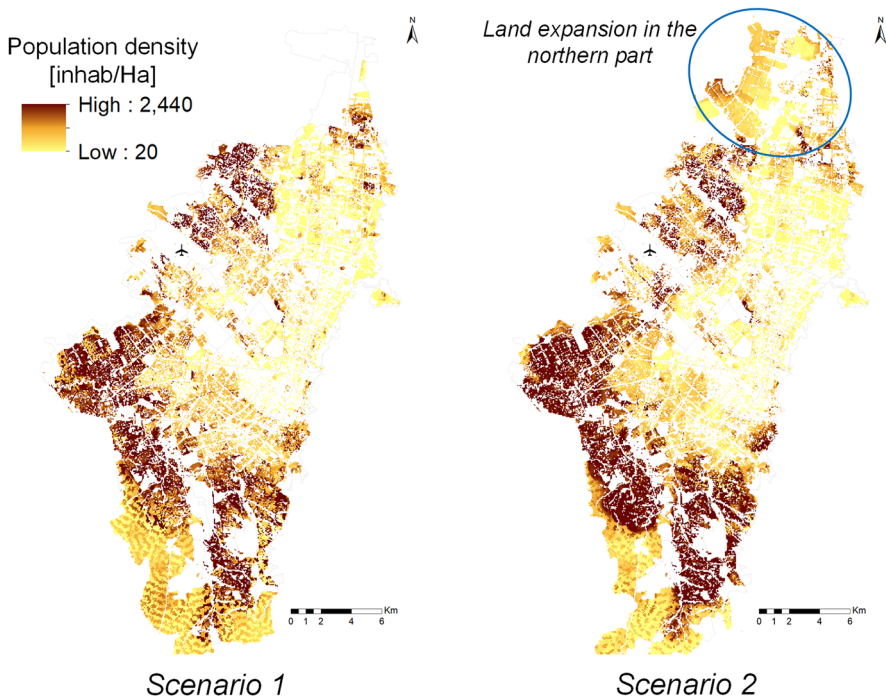

Scenario 1   *Scenario 2*

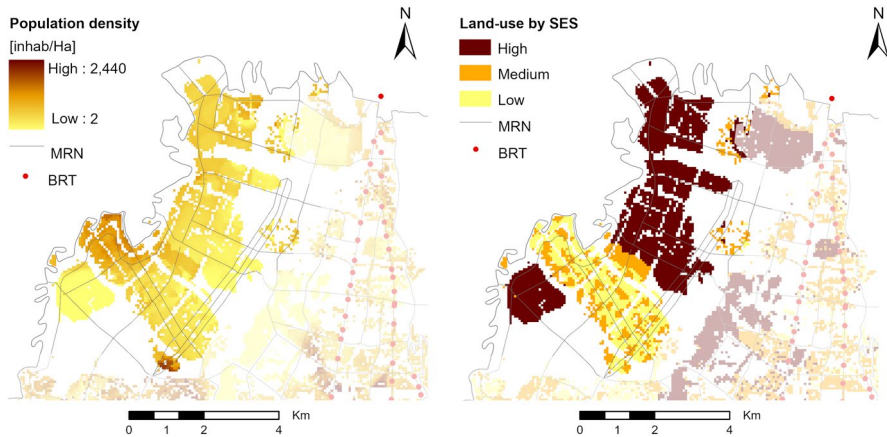**Fig. 5** Resulting population density by scenario 2050

Fig. 6 Population density and SES distribution results for Scenario 2

According to the CA model results, we must note the number of resulting pixels for both scenarios for each SES inside Bogotá. We can see that Scenario 2, which allows intervention in specific large areas in the northern part of the city, spreads into a larger area than Scenario 1. The basis of the second scenario, urban expansion in the north, is that the city just ended the Land-Use Master Plan and currently was issued without discussion. One possibility consists of modifying the current regulation that restricts the development of that area of Bogotá. This is currently a 1,396-ha zone with high-income low-density residential dwellings and sparse agricultural, entertainment, and educational uses (Guzman et al., 2020a). Then, the population density was estimated according to the RF results. It is essential to highlight that neither scenario grows into a very similar area. However, even though the areas are similar in size, the mean population density is higher in Scenario 1 for low-SES.

The population density distribution for the urban expansion scenario is shown in Fig. 6 left). These results show a pronounced development of residential land-use in a compact configuration in the north, with high densities around major roads in the west. Medium- and low-SES occupation tends to be located in the corridor that connects Bogotá with neighboring municipalities to the west. These uses are allocated as a cluster in proximity to the current urban area of Bogotá, where medium- and low-income areas predominate. These results suggest a series of new low-density high-SES neighborhoods in the urban expansion scenario and a reduction of low- and medium-SES urban spread in the southern part of the city.

The new urban expansion in the northern part of the city, simulated in Scenario 2, presents new possibilities and new challenges for the local administration, mainly because it is currently used primarily for agriculture use and is a protected area. When we consider that, by our calculations, around 720,000 inhabitants will settle in this area by 2050, public utility availability and critical infrastructure accessibility could become an issue for the local administration and the new settlers. The population density predictions are summarized in Table 4. These results are presented in a pixel-level prediction (60mx60m) in inhabitants

Table 4 Population density
results for both scenarios in
2050

Population density results			
SES	**Statistic**	**Scenario 1**	**Scenario 2**
High	Mean	334.3	359.1
	Sum	8,670,663.7	7,532,279.1
	Std Dev	123.3	131.9
Medium	Mean	587.3	574.3
	Sum	38,379,985.4	34,470,426.4
	Std Dev	184.6	156.7
Low	Mean	540.2	596.4
	Sum	311,654,000.8	32,786,044.4
	Std Dev	219.2	240.1

per hectare. This population density model can help estimate the infrastructure and utility needs for this type of development.

We built an interactive data visualization webpage where the original and predicted data in both proposed scenarios are placed. The page uses ArcGIS web map services hosted in the Universidad de Los Andes servers and the free hosting provided by GitHub. You can visit the page at this link: https://zibramax.github.io/RF-DENSPOB/.

The trips produced by each pixel were also estimated with the parameters and the projected variables resulting from the population density model. Table 5 presents the results of the travel demand estimated under each scenario by SES. In Scenario 2, there is an increase of 4% in travel demand compared to the base year. For Scenario 1, this increase is about 1.6%. Analyzing the trip growth disaggregated by SES, Table 5 also shows a sharp increase for the high SES, particularly in Scenario 2. Conversely, the low SES contraction of about 17% for Scenario 1 and 12% for Scenario 2.

The increases in travel demand are consistent with the explanatory variables' predictions from the RF model, the relationship between these variables, and the travel demand model. The low-SES gains participation in the urban area, but its average density falls in both scenarios. On the other hand, the medium- and high-SES lose participation in the study area. However, they become denser on average in Scenario 2, and only the high-SES lose average density under Scenario 1. As for the mean distance to the CBD, it increases in both scenarios due to urban footprint growth.

Table 5 Total projected trips
for present and both future
scenarios 2050

SES	Scenario 1	Scenario 2
High	738,939	1,018,854
Medium	4,689,094	4,425,358
Low	3,666,203	3,860,977
Total	**9,094,235**	**9,305,189**

Because the future is intimately linked to the past, the population and travel projections provide reasonably accurate predictions of how the city will change in the future. Although predicting the future is impossible, our imperfect estimations could be extremely useful tools for planning and analysis if they are constructed and interpreted properly.

Discussion and conclusions

Population density is one of the most critical indicators, creating consistent and objective population density maps for large, and growing urban areas remains challenging. Usually, the population is not distributed uniformly across the urban territory, causing a divergence between planning, high densities, and quality of life. For example, residents may be a small proportion of the population density in a mixed-use zone. This is why working with a small-scale land-use change model is essential. Studying and quantifying population growth and change at the disaggregated level has become more important for urban planning.

The rapid expansion of the urban footprint and changes in transport infrastructure has an unknown and differential influence on new settlements. Population density estimations can serve broader needs; they can be inputs into models to estimate population distribution, public utility provision, and travel demand estimation. In a particular social context, knowing the size and spatial distribution of the population is essential for planning and public policies. With few exceptions, previous work on CA-based land-use modeling has used 2D representations of land-use types. In our case, we previously developed a 2D model that lacked the third dimension for population density. We complemented it using the RF model. This research has demonstrated the potential of measuring one of the characteristic population features with simple explicative variables and using those features to obtain urban planning implications.

Creative analyses can effectively use existing resources in cities where information is scarce. They can add value to the land-use models commonly used to analyze and simulate urban growth by reproducing complex dynamics and adding valuable indicators such as population density. The Random Forest algorithm allowed us to simultaneously work with both continuous and categorical variables. We tested nine explicative variables of population density calculated from two simulated scenarios with different land regulations and transport infrastructure implementations across the study area. Furthermore, when these variables have a geographical component, it boosts the algorithm's capabilities. The RF algorithm in urban planning is a suitable way to show that many of the past analyses that we deemed incomplete or undoable can be easily complimented. Also, our approach can be replicated anywhere that can build a land-use model and has small-scale population density information.

This research uses a methodology that can complement land-use models with a population density estimation using an RF-based classification and regression, a machine learning method used in classification and prediction. This is the same approach that other models such as WorldPop use (Stevens et al., 2015). The ability of RF models to include both continuous and categorical variables allows the

introduction of relationships within variables where OLS and GWR fall short. A similar situation happens with the low randomness of the geographic data distribution where neither OLS nor GWR models work correctly.

This approach allows us to use a widely available dataset, expanding predictive modeling by allowing it to become accessible for planners and decision-makers. Population density is an essential issue in terms of economic efficiency and public infrastructure provisions for essential utilities. We also developed a travel demand model to estimate the produced trips regarding population densities and socioeconomic levels. This analysis could be a good push for the local administrations towards better planning and design of transport systems.

Finally, we are exploring ways to improve this model by acquiring more data or using more advanced algorithms like deep learning or neural networks that may not present the same limitations as the RF model. Besides extrapolating to the adjacent municipalities, we believe that using an advanced algorithm would also allow us to include demographic data like gender or age distribution into the model. This is important for understanding how residential densities can influence travel demand based on location and socioeconomic level, expanding interdisciplinary research on links between population, mobility, and the urban structure.

Declarations

Conflict of interest The authors declare no competing interests.

References

Archila Bustos, M. F., Hall, O., Niedomysl, T., & Ernstson, U. (2020). A pixel level evaluation of five multitemporal global gridded population datasets: A case study in Sweden, 1990–2015. *Population and Environment, 42*(2), 255–277. https://doi.org/10.1007/s11111-020-00360-8

Bassolas, A., Barbosa-Filho, H., Dickinson, B., Dotiwalla, X., Eastham, P., Gallotti, R., et al. (2019). Hierarchical organization of urban mobility and its connection with city livability. *Nature Communications, 10*(1), 4817. https://doi.org/10.1038/s41467-019-12809-y

Batty, M., & Longley, P. (1994). *Fractal cities: A geometry of form and function*. Academic Press.

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing, 114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Bettencourt, L. M. A., Lobo, J., Helbing, D., Kuhnert, C., & West, G. B. (2007). Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences, 104*(17), 7301–7306. https://doi.org/10.1073/pnas.0610172104

Boulange, C., Gunn, L., Giles-Corti, B., Mavoa, S., Pettit, C., & Badland, H. (2017). Examining associations between urban design attributes and transport mode choice for walking, cycling, public transport and private motor vehicle trips. *Journal of Transport and Health, 6*, 155–166. https://doi.org/10.1016/j.jth.2017.07.007

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324.

Cantillo-García, V., Guzman, L. A., & Arellana, J. (2019). Socioeconomic strata as proxy variable for household income in transportation research. Evaluation for Bogotá, Medellín, Cali and Barranquilla. *DYNA, 86*(211), 258–267. https://doi.org/10.15446/dyna.v86n211.81821.

Carlino, G. A., Chatterjee, S., & Hunt, R. M. (2007). Urban density and the rate of invention. *Journal of Urban Economics, 61*(3), 389–419. https://doi.org/10.1016/j.jue.2006.08.003

Cervero, R. B. (2013). Linking urban transport and land use in developing countries. *Journal of Transport and Land Use, 6*(1), 7. https://doi.org/10.5198/jtlu.v6i1.425

Chen, Y. (2010). A new model of urban population density indicating latent fractal structure. *International Journal of Urban Sustainable Development, 1*(1–2), 89–110. https://doi.org/10.1080/19463131003661025

Chi, G. (2009). Can knowledge improve population forecasts at subcounty levels? *Demography, 46*(2), 405–427. https://doi.org/10.1353/dem.0.0059

Chi, G., & Voss, P. R. (2011). Small-area population forecasting: Borrowing strength across space and time. *Population, Space and Place, 17*(5), 505–520. https://doi.org/10.1002/psp.617

Chi, G., Zhou, X., & Voss, P. R. (2011). Small-area population forecasting in an urban setting: A spatial regression approach. *Journal of Population Research, 28*(2–3), 185–201. https://doi.org/10.1007/s12546-011-9053-6

Fernández-Aracil, P., & Ortuño-Padilla, A. (2016). Costs of providing local public services and compact population in Spanish urbanised areas. *Land Use Policy, 58*, 234–240. https://doi.org/10.1016/j.landusepol.2016.07.032

Gao, J., & O'Neill, B. C. (2020). Mapping global urban land for the 21st century with data-driven simulations and Shared Socioeconomic Pathways. *Nature Communications, 11*(1), 2302. https://doi.org/10.1038/s41467-020-15788-7

Gaughan, A. E., Stevens, F. R., Linard, C., Jia, P., & Tatem, A. J. (2013). High Resolution Population Distribution Maps for Southeast Asia in 2010 and 2015. *PLoS ONE, 8*(2), e55882. https://doi.org/10.1371/journal.pone.0055882

Gerland, P., Raftery, A. E., Ševčíková, H., Li, N., Gu, D., Spoorenberg, T., et al. (2014). World population stabilization unlikely this century. *Science, 346*(6206), 234–237. https://doi.org/10.1126/science.1257469

Güneralp, B., Reba, M., Hales, B. U., Wentz, E. A., & Seto, K. C. (2020). Trends in urban land expansion, density, and land transitions from 1970 to 2010: A global synthesis. *Environmental Research Letters, 15*(4), 044015. https://doi.org/10.1088/1748-9326/ab6669

Guzman, L. A. (2019). A strategic and dynamic land-use transport interaction model for Bogotá and its region. *Transportmetrica b: Transport Dynamics, 7*(1), 707–725. https://doi.org/10.1080/21680566.2018.1477636

Guzman, L. A., Arellana, J., Cantillo-García, V., & de Dios Ortúzar, J. (2021). Revisiting the benefits of combining data of a different nature: Strategic forecasting of new mode alternatives. *Journal of Advanced Transportation, 2021*, 1–15. https://doi.org/10.1155/2021/6672961.

Guzman, L. A., & Bocarejo, J. P. (2017). Urban form and spatial urban equity in Bogota, Colombia. *Transportation Research Procedia, 25*, 4491–4506. https://doi.org/10.1016/j.trpro.2017.05.345

Guzman, L. A., Escobar, F., Peña, J., & Cardona, R. (2020a). A cellular automata-based land-use model as an integrated spatial decision support system for urban planning in developing cities: The case of the Bogotá region. *Land Use Policy, 92*(104445), 104445. https://doi.org/10.1016/j.landusepol.2019.104445

Guzman, L. A., Oviedo, D., & Bocarejo, J. P. (2017a). City profile: The Bogotá Metropolitan Area that never was. *Cities, 60*, 202–215. https://doi.org/10.1016/j.cities.2016.09.004

Guzman, L. A., Oviedo, D., & Rivera, C. (2017b). Assessing equity in transport accessibility to work and study: The Bogotá region. *Journal of Transport Geography, 58*, 236–246. https://doi.org/10.1016/j.jtrangeo.2016.12.016

Guzman, L. A., Peña, J., & Carrasco, J. A. (2020b). Assessing the role of the built environment and sociodemographic characteristics on walking travel distances in Bogotá. *Journal of Transport Geography, 88*, 102844. https://doi.org/10.1016/j.jtrangeo.2020.102844

H2O.ai. (2021). H2O.ai. *Distributed Random Forest (DRF)*. http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/drf.html

Jiao, J., Chen, Y., & Azimian, A. (2021). Exploring temporal varying demographic and economic disparities in COVID-19 infections in four U.S. areas: Based on OLS, GWR, and random forest models. *Computational Urban Science, 1*(1), 27. https://doi.org/10.1007/s43762-021-00028-5.

Jochem, W. C., Bird, T. J., & Tatem, A. J. (2018). Identifying residential neighbourhood types from settlement points in a machine learning approach. *Computers, Environment and Urban Systems, 69*, 104–113. https://doi.org/10.1016/j.compenvurbsys.2018.01.004

Kajimoto, M., & Susaki, J. (2013). Urban density estimation from polarimetric SAR images based on a POA correction method. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 6*(3), 1418–1429. https://doi.org/10.1109/JSTARS.2013.2255584

Kamusoko, C., & Gamba, J. (2015). Simulating urban growth using a random forest-cellular automata (RF-CA) model. *ISPRS International Journal of Geo-Information, 4*(2), 447–470. https://doi.org/10.3390/ijgi4020447

Khosravi, K., Pham, B. T., Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., et al. (2018). A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. *Science of the Total Environment, 627*, 744–755. https://doi.org/10.1016/j.scitotenv.2018.01.266

Lewis, S. (2018). Neighborhood density and travel mode: New survey findings for high densities. *International Journal of Sustainable Development & World Ecology, 25*(2), 152–165. https://doi.org/10.1080/13504509.2017.1321052

Liu, X. H., Kyriakidis, P. C., & Goodchild, M. F. (2008). Population-density estimation using regression and area-to-point residual kriging. *International Journal of Geographical Information Science, 22*(4), 431–447. https://doi.org/10.1080/13658810701492225

Onda, K., Sinha, P., Gaughan, A. E., Stevens, F. R., & Kaza, N. (2019). Missing millions: Undercounting urbanization in India. *Population and Environment, 41*(2), 126–150. https://doi.org/10.1007/s11111-019-00329-2

Ortúzar J, & Willumsen, L. G. (2011). *Modelling transport* (4th Editio.). Chichester: Wiley.

Puissant, A., Rougier, S., & Stumpf, A. (2014). Object-oriented mapping of urban trees using Random Forest classifiers. *International Journal of Applied Earth Observation and Geoinformation, 26*, 235–245. https://doi.org/10.1016/j.jag.2013.07.002

Qiang, Y., Xu, J., & Zhang, G. (2020). The shapes of US cities: Revisiting the classic population density functions using crowdsourced geospatial data. *Urban Studies, 57*(10), 2147–2162. https://doi.org/10.1177/0042098019871191

Rahmati, O., Pourghasemi, H. R., & Melesse, A. M. (2016). Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: A case study at Mehran Region, Iran. *CATENA, 137*, 360–372. https://doi.org/10.1016/j.catena.2015.10.010

Raymer, J., Abel, G. J., & Rogers, A. (2012). Does specification matter? Experiments with simple multiregional probabilistic population projections. *Environment and Planning a: Economy and Space, 44*(11), 2664–2686. https://doi.org/10.1068/a4533

Sarmiento, O. L., Useche, A. F., Guzman, L. A., Rodriguez, D. A., Dronova, I., Guaje, O., et al. (2021). Built environment profiles for Latin American urban settings: The SALURBAL study. *PLoS ONE, 16*(10), e0257528. https://doi.org/10.1371/journal.pone.0257528

SDP - Secretaría Distrital de Planeación de. (2018). *Análisis demográfico y proyecciones poblacionales de Bogotá*. Bogotá. http://www.sdp.gov.co/sites/default/files/demografia_proyecciones_2017_0.pdf

Sinha, P., Gaughan, A. E., Stevens, F. R., Nieves, J. J., Sorichetta, A., & Tatem, A. J. (2019). Assessing the spatial sensitivity of a random forest model: Application in gridded population modeling. *Computers, Environment and Urban Systems, 75*, 132–145. https://doi.org/10.1016/j.compenvurbsys.2019.01.006

Smith, S. K., Tayman, J., & Swanson, D. A. (2013). *A Practitioner's Guide to State and Local Population Projections*. (S. K. Smith, J. Tayman, & D. A. Swanson, Eds.) (Vol. 37). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-7551-0

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLoS ONE, 10*(2), e0107042. https://doi.org/10.1371/journal.pone.0107042

Stevens, F. R., Gaughan, A. E., Nieves, J. J., King, A., Sorichetta, A., Linard, C., & Tatem, A. J. (2020). Comparisons of two global built area land cover datasets in methods to disaggregate human population in eleven countries from the global South. *International Journal of Digital Earth, 13*(1), 78–100. https://doi.org/10.1080/17538947.2019.1633424

UAECD. (2019). UAECD. *Unidad Administrativa Especial de Catastro Digital*. https://datosabiertos.bogota.gov.co/dataset?q=UAECD

van Delden, H., & Vanhout, R. (2018). A short presentation of Metronamica. In *Geomatic Approaches for Modeling Land Change Scenarios* (pp. 511–519). https://doi.org/10.1007/978-3-319-60801-3_38

van Vliet, J., Hurkens, J., White, R., & van Delden, H. (2012). An activity-based cellular automaton model to simulate land-use dynamics. *Environment and Planning b: Planning and Design, 39*(2), 198–212. https://doi.org/10.1068/b36015

Wu, C., & Murray, A. T. (2005). A cokriging method for estimating population density in urban areas. *Computers, Environment and Urban Systems, 29*(5), 558–579. https://doi.org/10.1016/j.compenvurbsys.2005.01.006

Zheng, H. W., Shen, G. Q., Wang, H., & Hong, J. (2015). Simulating land use change in urban renewal areas: A case study in Hong Kong. *Habitat International, 46*, 23–34. https://doi.org/10.1016/j.habitatint.2014.10.008

Authors and Affiliations

Luis A. Guzman[1] · **Ricardo Camacho[2]** · **Arturo Rodriguez Herrera[3]** · **Carlos Beltrán[1]**

Ricardo Camacho
rcamachocast@student.unimelb.edu.au

Arturo Rodriguez Herrera
da.rodriguezh@uniandes.edu.co

Carlos Beltrán
ca.beltran957@uniandes.edu.co

[1] Departamento de Ingeniería Civil Y Ambiental, SUR, Grupo de Sostenibilidad Urbana Y Regional, Universidad de los Andes, Bogotá, Colombia

[2] Centre for Spatial Data Infrastructures and Land Administration (CSDILA), Department of Infrastructure Engineering, The University of Melbourne, Parkville, VIC 3010, Australia

[3] Departamento de Ingeniería Civil Y Ambiental, Universidad de los Andes, Bogotá, Colombia