Working paper

International Growth Centre

Using machine learning and remote sensing to value property in Rwanda

Paul Brimble
Patrick McSharry
Felix Bachofer
Jonathan Bower
Andreas Braun

February 2020

When citing this paper, please use the title and the following reference number:
C-38315-RWA-1

DIRECTED BY

FUNDED BY

Using machine learning and remote sensing to value property in Kigali

Paul Brimble[1,2]
Patrick McSharry[3,4,5]
Felix Bachofer[6]
Jonathan Bower[7]
Andreas Braun[8]

[1] Ministry of Finance & Economic Planning, Rwanda (MINECOFIN)
[2] Blavatnik School of Government, University of Oxford, UK
[3] Carnegie Mellon University Africa (CMU-Africa), Kigali, Rwanda
[4] African Centre of Excellence in Data Science, University of Rwanda, Kigali, Rwanda
[5] Oxford-Man Institute of Quantitative Finance, University of Oxford, Oxford, UK
[6] German Aerospace Center (DLR)
[7] International Growth Centre (IGC)
[8] Eberhard Karls University of Tübingen

This paper is partly funded by IGC and is a collaboration between all institutions listed above. Thanks to Rwanda Land Management and Use Authority, Rwanda's Ministry of Finance and Economic Planning, Rwanda Revenue Authority, National Institute of Statistics Rwanda, and Rwanda's Ministry of Infrastructure for data and ongoing support

Abstract

Property valuation models can achieve mass valuation transparently and cheaply. This paper develops a number of property valuation models for Kigali, Rwanda, and tests them on a unique dataset combining remote sensing data and infrastructure and amenities data for properties in Kigali, with sales transaction data for 2015. We use a machine learning approach, Minimum Redundancy Maximum Relevance, to select from 511 features those that minimise ten-fold cross validated Mean Absolute Error. Cross validated diagnostics are used to eliminate overfitting given that our goal is to generate a model that can be used to extrapolate value estimates out of sample. The performance of Ordinary Least Squares (OLS) is compared to that of a range of spatial models. Our best model covering all taxed parcels, achieves a cross validated $\mathbf{R}^2$ of 0.600 and a cross-validated Mean Absolute Error of 0.541. We find that locational variables relating to connectivity are most consistently important for overall property value across different models. We also attempt to develop the most accurate method of calculating building values. Our recommendations for future property valuation in Rwanda are: i) Given that the goal is extrapolation of the model to estimate the value of all properties outside of the sample of transacted properties, it is essential to eliminate overfitting as far as possible. This can be done by optimising cross validated diagnostics such as Mean Absolute Error and R^2. ii) The use of spatial models is desirable, in terms of out-of-sample accuracy, if and only if extensive testing of various spatial models alongside OLS on the basis of cross validated diagnostics, is possible; such models often overfit in sample but do not always outperform OLS out of sample. iii) Ideally a Computer Assisted Mass Appraisal would help determine full property taxes or land taxes, and not only building taxes, given that it is more accurate at estimating full property values or land values than it is at finding building values. Building values are also not directly observable and thus it is impossible to assess the accuracy of our imputed building value estimates. iv) Additional structural building data on variables such as building materials and numbers of rooms, which the Government of Rwanda plans to collect, would improve model accuracy.

1. Introduction

Property valuation is a valuable tool for effective tax revenue collection, but traditional valuation methods are expensive, time consuming, hard to independently verify, and vulnerable to corruption. This paper develops a number of property valuation models for Kigali, Rwanda, and tests them on a unique dataset combining remote sensing data for buildings in Kigali, with sales transaction data for 2015. We use machine learning techniques to select a model from among a large number of variables and ensure that the identified model is capable of generalising to out of sample properties. Specifically we employ Minimum Redundancy Maximum Relevance (mRMR) to select the variables that best predict property price data using Ordinary Least Squares (OLS) for parameter estimation and an out-of-sample evaluation. A key innovation is using cross-validation both to avoid overfitting and to obtain accurate estimates of performance. We employ this approach to estimate a model for the whole of Kigali province. We then estimate a range of spatial models based on the variables selected by mRMR and compare their performance with the benchmark OLS model.

Rwanda is one of the most densely populated countries in Africa and is urbanising rapidly from a low base of 16.5% (National Institute of Statistics Rwanda, 2012). The capital, Kigali has a fast-growing population which was 1.1 million in 2012 (National Institute of Statistics Rwanda, 2012) but is likely to grow to 2.5 million by 2032 (Bower and Murray 2019). Land and building values are increasing fast, and capturing some of the growth in these values will be vital to fund infrastructure to support the growing population.

Rwanda is of interest for two reasons. As Ali et al. (2018) note, it was the first African country to complete a nationwide land registration programme, helping to establish a complete and fully digitised legal cadaster. Secondly, Rwanda recently introduced a property tax law which came into effect on 1st January 2019, which moves from a "flat" rate per square metre to a rate based on building values, although land is still taxed at a flat rate decided at the local level. Computer Assisted Mass Appraisal (CAMA) is mentioned as a possibility for detecting inappropriately low building values in the new property tax law, with the intention to introduce it in January 2020. There is an opportunity to use the complete cadastral data to develop a CAMA that can support the implementation of the new property tax law either by validating self declarations of building value to detect under-valuation, or by classifying buildings into tax bands. The models in this paper could provide a prototype for a CAMA model which uses more up to date independent variable data. The model in this paper, or an updated version, could also be used to estimate the revenue potential of the new property tax law.

2. Literature Review

Hedonic pricing theory, which posits that price reflects certain internal and external utility-bearing characteristics of a product, was introduced in a seminal paper by Rosen (1974). Spatial variables were then increasingly incorporated into various hedonic regression analyses on land and property valuation. Harrison and Rubinfeld (1978) estimated the impact on housing prices of spatial variables such as weighted distance to employment centres and accessibility to radial highways, as well as air pollution and other variables. Shonkwiler and Reynolds (1986) used hedonic price models on distance and land use variables to analyse land prices in the urban fringe. Heikkila et al. (1989) found that simple central business district (CBD) gradient models are not sufficient to predict these values, especially in polycentric cities. Urban structures matter and spatial patterns have a distinct influence on accessibility, and economic and social interactions (Anas et al. 1998). In 1996, Wyatt published his results on property valuation using a Geographic Information System (GIS) for the geospatial analysis on accessibility to the road network.

The literature on modelling land values, housing values, or both, contains a range of characteristics, especially locational characteristics, of properties, that turn out to be statistically significant predictors: land-use, topography and environment (Heikkila et al. 1989, Srour et al. 2002, Demetriou 2016, Kim and Kim 2016, Ai 2005, Ali et al 2018, Jayyousi et al 2014); traffic connectivity (Wyatt 1996, Orford 2002, Yomralioglu and Nisanci 2004, Song and Sohn 2007, Cellmer 2014, Demetriou 2016, Kim and Kim 2016, Sasaki and Yamamoto 2018, Lan et al. 2018, Ai 2005); distance to amenities, services and CBD (Srour et al. 2002, Orford 2002, Yomralioglu and Nisanci 2004, Brasington and Hite 2005, Baroussa et al. 2007, Song and Sohn 2007, Kim and Kim 2016, Zainora et al. 2016, Lan et al. 2018, Ai 2005, Ali et al 2018, Jayyousi et al 2014); zoning and regulations (Glaeser and Gyourko 2002, Glaeser and Ward 2009, Ai 2005); socioeconomic variables (Heikkila et al. 1989, Orford 2002, Brasington and Hite 2005, Baroussa et al. 2007, Song and Sohn 2007, Cellmer 2014, Jiang et al. 2015, Zainora et al. 2016, Ali et al 2018); natural hazard risks (Brasington and Hite 2005, Sasaki and Yamamoto 2018) and housing characteristics (Sirmans et al. 2006, Bourassa et al. 2007, Song and Sohn 2007, Chrostek and Kopczewska 2013, Ali et al. 2018). A review of modelling approaches and spatial variables in the land and house valuation literature is available in Xiao (2017).

In addition, remote sensing-derived variables on buildings, infrastructure and land-use find their way into land valuation analyses (Dabrowski and Latos 2015, Chew at al. 2018). It allows the derivation and updating of urban information in a cost-effective manner (Dean and Owen 2019).

Our paper makes three contributions to the literature. First, we draw from a particularly wide range of data on property characteristics and data that we found was common in the literature, which we then compiled and processed. These include land-use, environment, road connectivity, distances to amenities and services, zoning and regulations, economic variables, and housing footprint area and volume. In particular, we use remote sensing-derived building footprint and height data for Kigali in 2015 published in Bachofer et al (2019).

Second, our paper contributes to a small and growing literature that models property values in Africa. Of the sample of 26 papers we reviewed that model property values, we counted 18 that cover cities in Europe or the US, five that cover Asia and the Middle East, one that covers New Zealand, and four that cover cities in Africa including Kumasi, Lagos, Cape Town, Nairobi and Kigali. As Ali et al (2018) note, Rwanda is of particular interest because it was the first African country to complete a nationwide land registration programme, helping to establish a complete and fully digitised legal cadaster.

Third, this paper contributes methodologically by i) comparing the results and accuracy from a range of spatial modelling techniques - for instance, we respond to the claim noted in Bidanset and Lombard (2014) that whilst they have found that geographically weighted regression performs better than OLS, further research is needed to evaluate and understand the performance of locally weighted regression and geographically weighted regression models; ii) unlike much of the literature, we use machine learning in the form of Minimum Redundancy Maximum Relevance (mRMR) to find the most accurate model; and iii) unlike some of the literature we take the need for out-of-sample accuracy seriously - because the goal of our model is to extrapolate property values from those with sales values, to the entire city - by using cross-validation to eliminate overfitting.

Ali et al (2018) was the first paper to model property values for Kigali; given that our paper does the same, it is worth pointing out a number of differences: Ali et al (2018) focus more on making the case for the potential of property valuation models, or Computer Assisted Mass Appraisal, to enable efficient property tax collection, and the benefits given the policy context of the introduction of a new property tax in Rwanda. Their policy discussion in particular is useful to read alongside this paper. However, our paper has a more purely methodological focus as well as a number of methodological differences. For instance, Ali et al (2018) use sales values from 2013 to 2016 to maximise the number of observations whereas our paper uses 2015 sale prices only, on the grounds that much of the data for the independent variables is from 2015; we compile and draw on a wider range of data, we test a greater range of spatial models, we use the machine learning approach mRMR, and we cross-validate our model to eliminate overfitting.

3. Data

To generate a comprehensive list of potential independent variables that can be mined to find an optimal property valuation model, we compile and process two types of data and merge them into a single dataset of characteristics on 367,667 parcels in Kigali Province. The first type of data is a national dataset of parcels, which includes parcel boundary shapes and unique parcel identifiers (UPIs) on the basis of which we match to the other two types of data; a subset of parcels that were transacted, also have sales values. The second data type includes a large range of parcel characteristics including building height and footprint data and a range of urban amenities, and is sourced from a variety of other, smaller datasets.

3.1 Parcel Boundaries and Sales Transaction Values

The first data type mentioned above is a dataset of parcel boundaries extracted from the Land Administration Information System (LAIS) which is hosted by the Rwanda Land Use and Management Authority. This is a comprehensive record of all the parcel boundary shapes for the entire country and provides the basis for linking the different types of data as each parcel is assigned a UPI.

Of these parcels, the subset for which there was a transfer of freehold title or emphyteutic lease, also have a market transaction value and date of transaction. The information on these transactions comes from Rwanda's LAIS, with the sales values recorded starting in 2015. The number of sales per year are listed in the first row of Table 1.

Table 1: **Sales Transactions for Kigali Province**

Year	**2015**	**2016**	**2017**	**2018**
All Sales	10,246	13,991	16,352	15,155
All Parcels	7,445	9,383	10,329	9,045
Taxed Parcels	4,726	5,608	5,521	4,634
Taxed Land Parcels	998	1,199	1,066	865
Taxed Building Parcels	3,728	4,409	4,455	3,769

Source: Land Administration Information System.

For the purpose of our analysis, the parcels are sorted into four additional main groups, which we refer to as "parcel data groupings" throughout the rest of the paper. The first, "all parcels" group, refers to the set of usable parcel data that results at the end of a four step filtering procedure described below; for 2015 this contains 7,445 parcels out of the original total of 10,246. The second "taxed parcels" group is a subset of the "all parcels" group, and includes only the parcels on which buildings will be taxed; for 2015 this contains 4,726 parcels. This includes parcels with the following official land use types: residential, commercial, industrial, economic, scientific, social and tourism; agricultural land is omitted as it is not taxed under the new law. Any buildings on these taxed parcels are subject to the new property tax law, but this group of data includes both built and unbuilt parcels. The third and fourth groups of data are two mutually exclusive subsets of the second group: the third group of data comprises taxed unbuilt parcels - land only - and includes 998 parcels in 2015; and the fourth group comprises taxed land that has a building on it and includes 3,738 parcels.

To get to the sales values we use as our final dependent variable, the natural log of price per square metre in 2015, we filter the sales using a four step procedure. In the first step we eliminate parcels that cannot be matched between the sales transactions data and the parcel characteristics data. When pairing the sales transaction data to parcel characteristics based on parcel boundary, matches do not always occur. This happens when parcel boundaries are not the same for the sales transaction data as they are for the parcel characteristics data. Over time, the number of parcels increases as parcels are merged, split and undergo boundary changes and completely new parcels are added to the system. Whenever there is a change, new UPIs are created and old UPIs are retired, except in the case of the amendment of the boundary between two parcels that otherwise remain intact. It is necessary to omit parcels that cannot be matched between sales transaction data and parcel characteristics data from the analysis.

The second step involves removing parcels that underwent a boundary change and for which there are discrepancies between the parcel areas for the same UPI; these are removed because the parcel characteristics will be incorrect. In practice, to avoid discrepancies due to precision, we remove the transaction from our dataset if the areas differ by more than 1%.

In the third step we remove duplicate sales transactions data for parcels that have been sold more than once. For the sales in 2015, we keep the sale which is nearest the date of the satellite image from which the building footprint and height data are derived. For sales in the following years, we keep the sale which occurred first, although we do not use this data in this paper.

In the final step we filter the data for any outliers. This includes removing parcels with a price less than 100 RWF per square metre, or above 300,000 RWF per square metre as well as parcels with areas less than 50 meters squared or greater than 300,000 meters squared. We identify the outliers through a combination of expert opinion and visualisation techniques by plotting the rank against the logarithm of sales value per meter squared and parcel size.

3.2 Parcel Characteristics

The second category of data is a broad range of characteristics assigned to each parcel which are extracted and processed from a variety of sources. These characteristics are used as the independent variables for our regression model; here we describe the various sub-types of data within this. We also include log and squared transformations where appropriate which allows us to capture non-linearities while still using linear models for estimation. In total, we end up with 511 potential independent variables in addition to a constant term.

3.2.1 Remote Sensing-Derived Building Characteristics Data

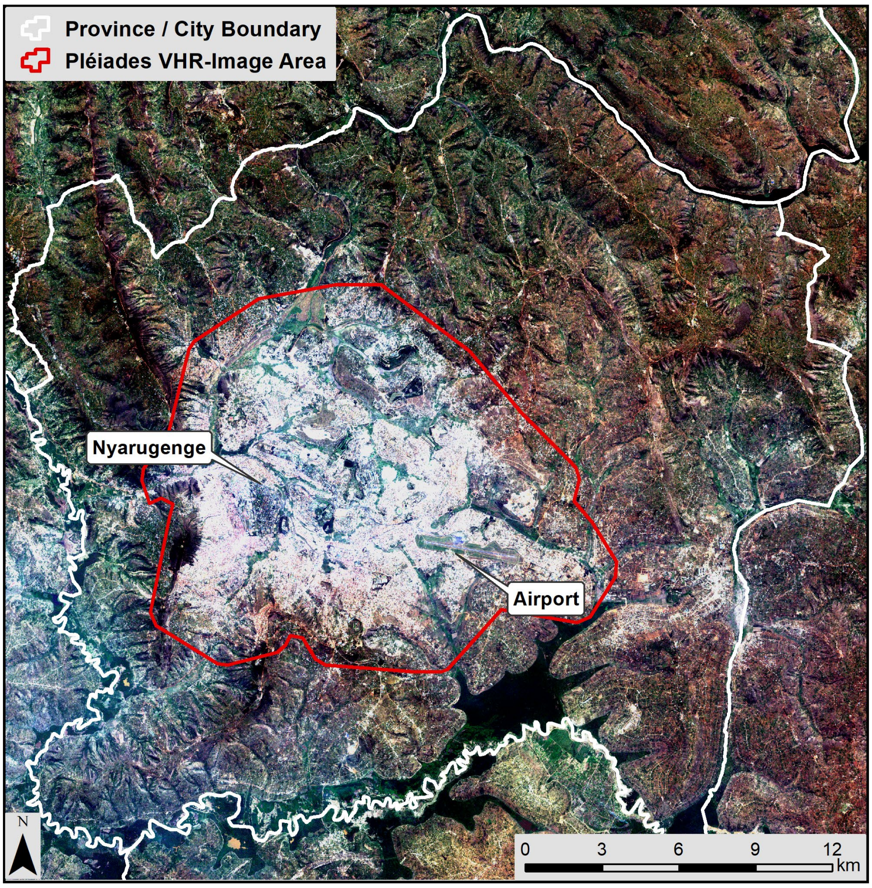
Building footprint, building height and characteristics data for Kigali are derived from two satellite images. The process by which these are generated is described in Bachofer and Murray (2018), which provides a comprehensive dataset to the City of Kigali (and collaborating researchers) on the stock of buildings in the city, and changes in the building stock, over time between 2008-2009 and 2015. To produce this data, images of Kigali were obtained from the very high resolution (VHR) Pléiades stereoscopic satellite and the high resolution (HR) RapidEye satellite mission for 2015, as well as aerial images for 2008-2009. The data covers the whole of Kigali Province, with the VHR images available inside the red boundary shown in Figure 1, which represents the densely built up area. Outside the red boundary, we obtain RapidEye satellite images which were sufficient to detect buildings, but not the respective building type.

For the 2015 satellite image data, a semi-automated process using an Object-Based Image Analysis (OBIA) approach (Blaschke 2010), followed by manual correction, is applied to identify building footprints and classify these into building typologies. The data and analysis for 2015 has been made available within the "Rapid Planning" project by the University of Tübingen. Image quality of the 2009 aerial images make a fully automated process impossible, so building footprints and building typology have been generated manually (Bachofer and

Murray 2018). The building objects of both datasets refer to each other, which makes a quantitative and qualitative change analysis possible. The entire dataset was revised in 2019 and published in an open access data repository and contains further information on building heights (Bachofer et al. 2019).

For the purposes of this paper, we focus on the building characteristics from 2015. We also calculate estimated building volume by combining data on building footprints and heights. Various combinations such as building volume per parcel area were also included. Furthermore, this dataset is essential for determining the taxed land and taxed built parcel categories.

Figure 1: **Kigali Province and High Resolution Satellite Image**



Source: Bachofer and Murray, 2018 (modified).

3.2.2 Amenities and Infrastructure Data

We compile amenities data on the location of roads, bus stops, bus routes, markets, schools, hospitals and other amenities in addition to zoning areas; this data was kindly made available by Rwanda's Ministry of Infrastructure, City of Kigali One Stop Centre, National Institute of Statistics or were retrieved from Open Street Map (OSM). This data is available for the years between 2012 and 2015.

Data on the location of amenities is used to calculate minimum distances between parcels and the amenities in addition to counts for different administrative levels including village, cell and sector; we also generate a non-administrative "block" level from the building data which is smaller than the village level. Data on zoning for permitted land use enables us to determine the area and share of each geographic level which is subject to a specific zoning regulation.

3.2.3 Economic Data

We also include economic variables which might capture labour market opportunities in certain geographic areas. Unfortunately, household survey data is only available at the district level which is insufficiently disaggregated for our purposes. Therefore, we rely on the 2014 establishment census for which the data is available at the sector level and include firm sizes and employment.

3.2.4 Grouping of Parcel Characteristics by Type

Furthermore, we can break down the parcel characteristics into three broad groups following Xiao (2017): i) structural, ii) locational, and iii) neighbourhood-level variables. Structural variables are internal attributes which describe the physical characteristics of the parcel. We further divide these variables into structural land and structural building variables. The structural land variables are any physical characteristics that are unrelated to any buildings such as slope, area and perimeter of the parcel. The structural building variables describe the buildings on the parcel and include footprint area, height and volume. Locational variables include the distance of the parcel to amenities and infrastructure. Neighbourhood variables include the social and economic characteristics of the neighbourhood at block level or other geographic level above the parcel level, and includes proportions and areas of zoning variables, counts of amenities and sector-level economic variables. Grouping the characteristics in this manner is a useful way to think about the variables which intuitively should affect land values (structural land, locational and neighbour variables) and those which should affect building values (structural building).

4. Model Types

In this section we describe a suite of econometric models, most of which incorporate spatial components, that we use to try to improve the accuracy of our property valuation models as applied to the four parcel data groupings described in section 3.1. Our modelling process has two phases: first, for each parcel data grouping we run a Minimum Redundancy Maximum Relevance process that incorporates k-fold cross-validation, which generates two "benchmark" OLS models, the parsimonious model and full model, defined in section 5, each containing a set of variables. Second, we use these eight sets of variables to generate three main groups of models described in this section: i) coordinate adjustment models, ii) spatial autoregressive models, and iii) local linear models. In total, we run a combined 97 models for each set of variables, which implies 194 models for each parcel data grouping. The majority of the spatial econometric models described below are reproduced from LeSage and Pace (2009).

4.1 Ordinary Least Squares

The benchmark model that is used for all comparisons is ordinary least squares. This is also the model that is used for the mRMR variable selection procedure. Let y be an $n \times 1$ vector of observations on the dependent variables, X be an $n \times k$ matrix of the observations on the k independent variables, β be a $k \times 1$ vector of coefficients and u be an $n \times 1$ vector of error terms. Then we can express the general linear model as:

$$y = X\beta + u$$

We then define the OLS estimate of the vector of parameters β as $\hat{\beta}$ and define the vector of predicted dependent variables as $\hat{y} = X\hat{\beta}$ and the vector of residuals is $\hat{u} = y - \hat{y}$.

4.2 Coordinate Adjustment Models

These types of models are simple modifications to OLS, that incorporate coordinate variables. Spatial components are modelled using linear, quadratic or cubic polynomial expansions of the latitude and longitude data of each parcel. These higher order polynomials help to capture the complexities of location. We define the three polynomial expansions with C_k where k is the order of the polynomial expansion.

4.2.1 OLS with Coordinates

The most direct way to incorporate these coordinate expansions is to supplement the OLS model with these variables directly:

$$y = X\beta + C\gamma + u$$

For the purposes of this paper, we use linear and quadratic coordinate expansions for this type of model.

4.2.2 Trend Surface Correction

An alternative is to model the error terms of the OLS model and create a locational variable that enters as a supplementary variable to the initial OLS model. It is a simple two step model. First, run the OLS regression $y = X\beta + u$ and obtain estimates of the error term $\hat{u}$. Then run a regression of $\hat{u}$ on a constant and the coordinates:

$$\hat{u} = \alpha + C\gamma + e$$

Then, obtain predictions of the error term and refer to this new variable as a locational variable, $l = \hat{\alpha} + C\hat{\gamma}$. Finally, add this location variable as a supplementary variable in the original OLS model:

$$y = X\beta + l\delta + \epsilon$$

For the purposes of this paper, we use the cubic coordinate expansion. For extrapolation, the trend surface correction models requires the auxiliary regression coefficients to calculate the location variable l and then final regression coefficients to obtain $\hat{y}$.

4.3 Spatial Autoregressive Models

Spatial autoregressive models quantify location through spatial weighting matrices W which are $n \times n$ matrices where the element $W_{i,j}$ captures the spatial weight of parcel j on parcel i. In its most general form the model is:

$$\begin{aligned} y &= \rho W_1 y + X\beta + u \\ u &= \lambda W_2 u + e \end{aligned}$$

In this model, there is a spatial component in the dependent variable and the error term. In general, W_1 does not have to equal W_2 but for our purposes, we impose the restriction that $W_1 = W_2 = W$. Given certain restrictions on the coefficients ρ and λ, we define a subgroup of spatial autoregressive models.

4.3.1 Spatial Autoregressive Error Model

When $\rho = 0$ for the above two equations, then location only enters the model through the error term. We refer to this model as the spatial autoregressive error model which reduces to:

$$y = X\beta + (I - \lambda W)e$$

4.3.2 Mixed Autoregressive Model

When $\lambda = 0$ and $\rho \neq 0$ then location only enters the model as a linear combination of neighbouring units. We refer to this model as the mixed autoregressive model which reduces to:

$$y = (I - \rho W)X\beta + (I - \rho W)e$$

4.3.3 General Spatial Model

In cases when there are no restrictions on both ρ and λ, we refer to this model as the general spatial model which reduces to:

$$y = (I - \rho W)X\beta + (I - \rho W)(I - \lambda W)e$$

4.3.4 Weighting Matrices

The type of weighting matrix, W, used is a key input for these types of models. For parcels i and j, $W_{i,j}$ is the weight that parcel j has on parcel i. By convention, the weight of a parcel with respect to itself is zero such that $W_{i,i} = 0$. These types of weighting matrices can be derived from contiguity relations, distance functions or a combination of both. For this paper, we focus on weighting matrices defined by distance.

To create this type of weighting matrix, we begin by creating a distance matrix D where $D_{i,j} = D_{j,i}$ measures the distance between parcels i and j. Note that this matrix is symmetric with all diagonal elements equal to zero as $D_{i,i} = 0$. With this distance matrix, there are three main characteristics that are needed to describe each weighting matrix: i) weighting function, ii) truncation, and iii) normalisation.

The first step is to define a weighting function f which converts distance into a weight. Suppose that U is the weighting matrix prior to truncation and normalisation. Then $U_{i,j} = f(D_{i,j})$ for some weighting function f. This does not apply to the main diagonal which is set to zero by definition. For this paper, we focus on two simple weighting functions: i) inverse distance, and ii) binary distance. The inverse distance function simply means that the weight of a parcel on another parcel is inversely proportional to the distance. The binary distance function takes on values of 0 or 1 and means that either a parcel has constant weight or no weight at all. This function only works if the model is truncated by some distance otherwise all parcels will have equal weight on all other parcels.

The next step is to truncate the model by some distance d. This means that parcels which are more than d units apart do not affect each other. Suppose that V is the truncated weighting matrix, then we have that:

$$V_{i,j} = 0 \qquad if \quad D_{i,j} > d$$

$$V_{i,j} = U_{i,j} \quad if \quad D_{i,j} \leq d$$

For this paper, we focus on three truncation distances in which $d_1 = 1$, $d_2 = 0.5$ and $d_3 = 0.25$ where the units are in kilometres. For the inverse distance weighting function, we also include a weighting matrix with no truncation which is not possible for the binary weighting function.

Finally, we chose a normalisation method to create the final weighting matrix W. For this paper, we focus on row and spectral normalisation approaches. Row normalisation simply means that the sum of each row is equal to unity. This could potentially involve multiplying each row by a different scalar which will convert a symmetric matrix into an asymmetric one instead. This normalisation amounts to spreading the spatial effect of neighbours proportionally. Spectral normalisation divides each element in the matrix by the largest eigenvalue of the matrix. Therefore, this normalised matrix differs only by a single scalar and thus, symmetry is maintained. This implies that in total, we consider two weighting functions, three truncation distances and two normalisation methods resulting in 14 total weighting matrices (including the pair of non-truncated inverse distance matrices).

4.3.5 Estimation Methods

We use two estimation methods for all these spatial autoregressive models. The first is through maximum likelihood and the second is a general method of moments approach. As both estimation methods yield different results, we include the regression results from both.

Therefore, in total, we have three types of spatial autoregressive models, 14 weighting matrices and two estimation methods for a combined 84 separate estimations per model.

4.4 Local Linear Models

The idea of a local model is to offer a means of representing global nonlinear relationships using a local approximation. Essentially, a local neighbourhood is defined around the state vector and an appropriate method is then used to describe the dynamics around that point. The state space refers to the space spanned by the set of explanatory variables. The geographical nature of models for valuing parcels suggests that there is also the potential of thinking about the locational aspects of parcels and their proximity to each other. It is intuitive to expect that neighbouring geographical parcels will have similar valuations. It is also possible to consider additional variables about characteristics such as building type and height, in order to appropriately measure similarity; however, for the purposes of this paper, we focus on geographical similarity.

In comparison to the previous global models which attempt to capture spatial dependence, local linear models account for spatial heterogeneity. In these cases, we attempt to model location i with its own location-specific model with location-specific coefficients. In a simple linear context, we would have the following for each location i:

$$y_i = x_i'\beta_i + u_i$$

By estimating location specific coefficients for the $k \times 1$ vector β_i for all locations using all n data points, we would normally encounter a degrees of freedom problem as we simply do not have sufficient data to obtain the nk total coefficients. Therefore, to overcome this particular issue, we employ local linear models.

4.4.1 Local Average

The simplest version of this approach is known as local analogue which refers to finding the most similar state vector observed in the past, often referred to as the nearest neighbour and using this to generate a forecast. An obvious extension is to consider the k nearest neighbours and to take an average of their trajectories to determine the forecast. These approaches have been successfully used for forecasting physical systems that display nonlinear and potentially chaotic dynamics (McSharry & Smith., 2004). It is also possible to use a kernel to weight the influence of the neighbours and this was found to provide highly competitive point and probabilistic forecasts of economic output (Arora et al., 2013).

4.4.2 Spatial Expansion Model

In the spatial expansion model, we run a global model on all the data to obtain a set of core coefficient estimates β_0 which can be transformed using location information for parcel i to obtain the location specific coefficients β_i. To describe this model, we need to introduce slightly different notation. The $n \times 1$ vectors y and u remain the same but the explanatory variable matrix $\tilde{X}$ is now an $n \times nk$ matrix consisting of the $k \times 1$ vectors x_i in the form below. We also define $\tilde{\beta}$ as an $nk \times 1$ vector containing all the location specific coefficients β_i stacked on top of each other. These two new matrices are:

$$\tilde{X} = \begin{pmatrix} x_1' & 0 & \dots & 0 \\ 0 & x_2' & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & x_n' \end{pmatrix}, \qquad \tilde{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}$$

Given these definitions, the spatial expansion model can then be written in a similar way to that of OLS:

$$y = \tilde{X}\tilde{\beta} + u$$

The spatial expansion aspect comes from the fact that we can obtain the location specific coefficient β_i with data on the latitude and longitude of location i, denoted by lat_i and lon_i, in addition to a set of core coefficients β_0, which is a $2k \times 1$ vector consisting of a $k \times 1$ vector of latitude coefficients β_{lat} stacked on top of another $k \times 1$ vector of longitude coefficients β_{lon}. The equation for β_i is given by:

$$\beta_i = \begin{pmatrix} lat_i \, I_k & lon_i \, I_k \end{pmatrix} I_{2k} \, \beta_0, \qquad \beta_0 = \begin{pmatrix} \beta_{lat} \\ \beta_{lon} \end{pmatrix}$$

Therefore, for the spatial expansion model, we estimate the core coefficients β_0 and can then easily extrapolate the set of coefficients for any parcel j as long as we have the coordinates data.

4.4.3 Geographic Weighted Regression

For geographic weighted regressions, we require a spatial weighting matrix that is different from that in the spatial autoregressive models. We denote these weighting matrices as $\tilde{W}^i$ for which the off-diagonal elements are all zero. The diagonal elements $\tilde{W}^i_{j,j}$ is the spatial weight between parcel i and j. For this paper, we define three weighting functions: i) gaussian, ii) exponential, and iii) tricube. We define $d_{i,j}$ as the distance between parcels i and j. The three weighting functions are presented as follows in order:

$$\tilde{W}^i_{j,j} = \phi(d_{i,j}/\sigma_i\theta)$$

$$\tilde{W}^i_{j,j} = \sqrt{exp(-d_{i,j}/\theta)}$$

$$\tilde{W}^i_{j,j} = (1 - (d_{i,j}/d_{i,\bar{q}})^3)^3 \; I(d_{i,j} < d_{i,\bar{q}})$$

The parameter θ is a bandwidth parameter that describes how the weights decline with distance. ϕ is the standard normal density and σ_i is the standard deviation of the distance vector d_i. Parcel $\bar{q}$ is the q^{th} nearest neighbour of parcel i and $I()$ is an indicator function. For the three different weighting functions, we calibrate the parameter θ or identify the q^{th} nearest neighbour.

Once we have the weighting function $\tilde{W}^i$ for each parcel, we run n separate weighted regressions of the following form:

$$\sqrt{\tilde{W}^i}\, y = \sqrt{\tilde{W}^i}\, X\beta_i + \sqrt{\tilde{W}^i}\; u$$

To extrapolate with this model, suppose that we only have explanatory variables for parcel j. We introduce additional notation for expositional purposes. We define X_{-j} and y_{-j} as the training data which excludes information on parcel j. We then create the weighting matrix $\tilde{W}^j$ and estimate the parcel j specific coefficients:

$$\hat{\beta}_j = (X'_{-j}\, \tilde{W}^j\, X_{-j})^{-1}(X'_{-j}\, \tilde{W}^j\, y_{-j})$$

With $\hat{\beta}_j$ and explanatory variables for parcel j, we can then calculate the predicted value of that parcel $\hat{y}_j$. Therefore, extrapolation requires running a new regression for each out of sample location. The in-sample data is used to calibrate the necessary parameters and as the source of data to run each additional regression. Finally, given the three weighting functions, we obtain three geographic weighted regressions in total.

4.4.4 Geographic Non-Weighted Regression

We define geographic non-weighted regressions as geographic regressions for which the weights are binary. This is equivalent to running regressions on subsets of the data. For a given q, we run a regression for parcel i using the q nearest neighbours. Using the indicator function notation from the earlier subsection, and defining parcel $\bar{q}$ as the q^{th} nearest neighbour, the weighting function would be:

$$W^i_{j,j} = I(d_{i,j} < d_{i,\bar{q}})$$

We calibrate q by choosing the value of q which minimises the mean absolute error of the model's predictions. We limit q to have a minimum value of twice the number of parameters estimated. We then double the value of q until we exceed the number of possible neighbours. A model for which q is equal to the entire sample of neighbours would be equivalent to a standard OLS model for the whole dataset.

The simplicity of this weighting function and the calibration procedure allows for a range of possible functional forms. For this paper, we consider OLS, all the three coordinate models and the spatial expansion model. Therefore, we obtain five geographic binary-weighted regressions in total.

5. Methodology

The aim is to find the model that best predicts building values on taxed parcels. We begin this approach by developing one model for each of the four parcel data groupings, which comprise all parcels, taxed parcels, taxed built parcels and taxed land. Each model only uses sales data from that particular grouping and the number of sales are recorded each year are given in Table 1. The first step is to select the variables to be used to estimate the model. Our variable selection method creates two sets of variables comprising the parsimonious model and the full model. The parsimonious model is the most conservative in that a variable is included only if machine learning has selected it as optimal for all ten folds of the cross-validation process described in section 5.3. The full model contains all variables selected for each of the ten folds in the cross-validation process. Afterwards, we estimate a complete suite of models - for each model type, a parsimonious and a full model for each of the four parcel data groupings. We then compare the models on the basis of R^2 and Mean Absolute Error, to find which combination of model type and set of variables gives the highest performing model as described in section 5.4. This results in a final four models for each of the parcel data groupings. We then find the optimal building model as described in section 5.5.

5.1 Machine Learning

Machine learning (ML) refers to a collection of techniques that are required to construct a model from data. The justification for the technique selected depends on the nature of the challenge, the type of data, the application domain and implementation. The task of identifying a model that can predict the value of parcels is termed "supervised learning" given that a set of historical transaction prices will be used to train the model. It is also necessary to choose a performance metric by which to compare models; we discuss performance metrics in section 5.2.

An overarching theme in ML is the challenge of overfitting, in which a model fits both the underlying data generating process and the noise in a given set of data, with the result that its accuracy metrics are spuriously high. In our case, the challenge for a property valuation model that will be accurate for a whole city is to find a model of property values for the parcels for which there is sales data, that can be most accurately extrapolated to the rest of the parcels. It is thus important that the underlying signal is extracted rather than fitting any noise that may be present in the particular set of observations that are available. It is relatively easy to overfit the training data when there are many variables relative to the amount of observations; we have 511 variables.

Arora et al. (2018) demonstrate that overly complex models are unlikely to be competitive for out-of-sample forecasting. The philosophical principle of Occam's Razor also provides motivation for parsimony, and for a property valuation model, transparency is a virtue. In the presence of so many potential variables, cross-validation is a useful way to eliminate them to leave only those that are likely to be predictive out-of-sample. The cross-validation process uses a different set of data for training and for testing so that all estimates are out-of-sample and never in-sample. It is important to use cross-validation techniques when constructing predictive models for which the out-of-sample predictive performance that is likely to be achieved in practice needs to be accurately assessed.

The ML approach usually involves three stages: (1) feature construction; (2) feature selection; and (3) model construction. Each of these generic stages is first explained before providing additional details of the chosen approaches for performance evaluation, cross-validation and variable selection. We describe them briefly here.

5.1.1 Feature Construction

In ML terminology, a feature is an explanatory variable; feature construction is the first stage of ML and involves the generation of features from available data and also new features from existing ones. For example, variables may be normalised with respect to others such as dividing by the number of square metres; mathematical transformations may be applied, for example, taking natural logarithms or quadratics; and for remotely sensed data, neighbourhood or block level variables may be constructed from geographically finer grained data such as percentage vegetation coverage. In what follows, we will refer to features as variables for consistency with the rest of the paper.

5.1.2 Feature Selection

In ML there are a large number of feature selection techniques available. A common approach to variable selection is to first select those variables that are most relevant. For example, relevance could be established by selecting those variables that are most strongly correlated to the dependent variable. There are a number of heuristic algorithms available, such as the sequential forward and backward selection, also known as stepwise regression. Stepwise approaches consider the addition of a new feature and removal of existing features at each step. A substantial challenge results from the fact that some variables are likely to be strongly correlated with each other and therefore including two of these variables in the model would lead to redundancy in the subset of selected variables. Ideally it is best to select features that are independent from each other while still containing predictive information about the dependent variable. It is also possible to start with all variables included and to eliminate one variable at a time. Both stepwise approaches will terminate once there is nothing left to add or remove. When faced with many variables, there is no guarantee that these two stepwise approaches will agree.

A more appropriate approach would be to add a penalty to the cost function that determines which variables are selected in order to focus on a sparse set of variables and therefore a more parsimonious model. Ridge regression, least absolute shrinkage and selection operator (LASSO) and elastic net are all variable selection techniques that rely on penalty terms (Hastie et al., 2009). In the case of our specific challenge, we have a large number of collinear variables and require a selection technique that can account for redundancy when considering candidate variables with high levels of correlation. Indeed the high levels of correlation between variables is exacerbated by including transformations of variables, such as logs and squared terms. One approach known as Minimum Redundancy Maximum Relevance (mRMR) selection was introduced to avoid this redundancy issue (Peng et al., 2005) and has been found to be more powerful than the maximum relevance selection in a number of empirical comparisons (Tsanas et al, 2012). mRMR is able to identify relevant variables but also takes account of the variables that have already been selected; we utilise this and describe it further in section 5.4.

5.1.3 Model Construction

Linear regression offers a means of understanding the importance of different explanatory variables. While there are many more complicated nonlinear models that could be considered, a linear model structure offers transparency and best facilitates communication of the meaning and relevance of the selected variables. This is an important consideration given that the resulting model will be used by policymakers and for decision-making. Nevertheless while linear regression is linear in the parameters, it can also include nonlinear terms. Indeed this has been the approach taken here and log terms have been selected. Therefore we have essentially constructed a nonlinear model while carefully monitoring the contribution of each individual variable.

There are many other model structures that could be considered: decision trees, random forest (RF), k nearest neighbours (KNN), support vector machine (SVM) and artificial neural networks (ANN). Unfortunately, many of these models have the disadvantage of being more complex and therefore much more difficult to understand. The beauty of the linear model is the ability to construct a scorecard that could be simply calculated. Our primary goal here is to achieve a high level of predictive accuracy, that we are confident will hold in the future and where we can easily explain the role of each variable.

5.2 Performance Evaluation

A range of metrics could be used to evaluate the performance of the models that we estimate. These include standard statistical and forecasting metrics such as the coefficient of determination, R^2, root mean squared error (RMSE) and the mean absolute error (MAE). Metrics exist that are specific to property valuation such as the coefficient of dispersion and coefficient of variation. Another metric that may be useful to policy makers is the proportion of valuations that lie within a certain percentage range of actual sales values. Where property values are self-declared for tax purposes, the property valuation model could both provide the benchmark model-estimated property value against which self-declarations are evaluated, and show a reasonable percentage range below which they should be investigated for under-valuation. This reasonable minimum percentage of the model-estimated value would be based on the actual dispersion of actual sales values around predicted values.

However, we choose to focus specifically on the MAE as the main metric for establishing success when comparing and contrasting different models. Given that the predictions are in natural logs, the mean absolute error can be interpreted as the mean percentage deviation of the model's predictions from the true values. This is particularly helpful as it can be easily conveyed to policymakers and the tax administrators.

It is worth distinguishing between deterministic and non-deterministic model diagnostics to compare models. Non-deterministic model diagnostics are the mean of multiple metrics, obtained for each data fold for the tenfold cross validation explained in the following section. In the context of a k-fold approach, for all k subsamples used as training data, we would obtain k independent model diagnostics. Therefore, the cross-validated nondeterministic model diagnostic would be the mean of these k subsample model diagnostics. The benefit of this approach is that it also provides standard errors for these diagnostics.

However, for this paper all cross-validated model diagnostics we report will be deterministic. This approach is only valid in cases where the cross-validation method yields a unique prediction for each observation in the entire dataset and all the predictions are used to obtain a single model diagnostic which is referred to as a deterministic model diagnostic. For example, instead of taking the mean cross-validated MAE for all k regressions generated by the cross validation process, we take the predictions from all the folds to generate a single model diagnostic. This approach has the benefit that it is perfectly comparable to any in-sample diagnostics.

Metrics such as the MAE are virtually the same if they are calculated in a deterministic or non-deterministic way, because they are linear; the only difference comes from the fact that not all of the k subsamples will be of equal size. The same cannot be said of other metrics such as the R^2. It is also worth noting that as our emphasis is on cross-validated model diagnostics, we do not use the adjusted R^2 as additional variables do not necessarily improve the cross-validated fit in the same way that it weakly improves the in-sample fit.

5.3 Variable Selection Using Minimum Redundancy Maximum Relevance and Cross Validation

A key contribution of this paper to the literature on property valuation is the consistent implementation of cross-validation techniques to ensure that we are optimising a model's accuracy for out-of-sample predictiveness. The rationale for cross-validation to avoid overfitting and overstating a model's out-of-sample accuracy is covered in section 5.1. This paper focuses on non-exhaustive cross-validation methods. For all of our models, we use a k-fold approach. This involves partitioning the data into k equally sized subsamples or "folds" of data. For the purposes of this paper, we set $k = 10$. We then run k estimations where each subsample is used as the validation data while the remaining $k - 1$ subsamples are used as training data. Given that $k = 10$, we thus use 90% of the data to generate a model, then test it on the remaining 10% of the data; we repeat this process for all folds of the data.

The choice of mRMR was explained in section 5.1.2. For each parcel data grouping (all parcels, all taxed parcels, taxed built parcels, taxed land parcels), mRMR is utilised to select optimal variables. mRMR assesses the performance of models with specifications consisting of different numbers of variables and a constant. As we rely on 10-fold cross-validation, for each n number of variables we obtain $k = 10$ separate regressions consisting of n variables.

We evaluate the cross-validated diagnostics across these k separate regressions to obtain the optimal number of variables n^* which minimise the mean cross-validated mean absolute error. An alternative approach could be to choose n^* as the number of variables which maximise the mean cross-validated R^2. In almost all cases, both definitions provide the same n^*. From here, we note that in these 10 separate regressions, there can either be a maximum of $10n^*$unique variables (in the unlikely case that each regression contains a different set of variables) or a minimum of n^* unique variables (in the case that each regression contains an identical set of variables). As noted in section 5 we define the *full* set of variables as all the unique variables that appear in all 10 separate regressions. We also define the *parsimonious* set of variables which contain only the variables which occur in all 10 of the regressions.

5.4 Model Selection for each Parcel Data Grouping

For each parcel data grouping (all parcels, taxed parcels, taxed built parcels, taxed land parcels), the mRMR variable selection method produces two sets of variables, the full set and the parsimonious set. For both sets of variables, we generate all types of model described in the earlier section. In order to determine the best model for each data grouping, we find the model from the list of all models, both full and parsimonious, that minimises the cross-validated MAE. In cases where the difference in MAE between the two best models is minimal, we evaluate the cross-validated R^2 to determine the best model. In cases where the difference in R^2 is minimal, the model with fewer variables is preferred. We then assign that model type and variable set as the best model for its parcel data grouping.

5.5 Model Selection to Obtain the Most Accurate Building Values

Whilst our broader goal is to find the best model for all parcel data groupings, the main goal of this paper is to obtain accurate building values, upon which the level of taxation in Rwanda is directly based. However, the building value is not directly observable, only the land value for unbuilt properties and the property value, which combines the value of land and of any buildings on it. For any built property, to predict the building value, we need to first obtain a prediction of land value. Thus, we use the models developed for the four parcel data groupings and evaluate which is the most predictive model for taxed land parcels - somewhat counterintuitively whilst we have used machine learning to find the optimal model for taxed land parcels, models generated using larger sample sizes may perform better. In the same way we evaluate which model is the most predictive for taxed built parcels (including land and buildings). Finally, to obtain the best predicted building values, we subtract the predicted values of taxed built parcels from the predicted land values of those parcels, to get the predicted building model.

6. Results

This section presents our results. Section 6.1 describes the types of variables chosen by our feature selection process for each parcel data grouping. Section 6.2 compares the model results for each parcel data grouping. Section 6.3 outlines the additional step to find the best model to use to impute building values. Finally, Section 6.4 illustrates the predictive performance of the final models arrived at in Sections 6.2 and 6.3, using a broader range of model diagnostics.

6.1 Variable Selection

For each parcel data grouping, the mRMR feature selection process generated two sets of variables. These variables are listed in Table 2, for a parsimonious and a full model as defined at the start of section 5. Furthermore, these variables are organised by the parcel characteristic groups described in Section 3.2.4.

The parcel perimeter is the only structural land variable that appears, and does so in all models except for the taxed built parcel category model. For three parcel data groupings containing built parcels, the structural building variables frequently chosen were, building volume and footprint area variables; often these variables are normalised by parcel area which is consistent with our choice of dependent variable, value per square metre. The building count variable is only selected for the model for all taxed parcels.

Locational variables proved to be consistently important determinants of value. Amongst the large list of distance variables available, distances to roads, bus stops and bus routes, or quadratic or logarithmic transformations of these three variables, were selected for all the parcel grouping models; this implies an unsurprising link between urban connectivity and property values. Additionally, a binary variable for routing distance to roads under 500m away and distance to primary school were selected for the taxed parcel and taxed land parcel groupings respectively.

Table 2: **Selected Variables**

All Parcels	Taxed Parcels	Taxed Land Parcels	Taxed Built Parcels
Parsimonious Model Variables			
Structural (Land)			
Perimeter^{1,2,3}	-	Perimeter^{1}	-
Structural (Building)			
Building Volume to Area^{1,2}	Building Volume to Area^{1,2}	-	Building Volume to Area^{1,2}
Building Footprint to Area	-	-	-
Building Count^{1}	-	-	-
Locational			
Distance to Road^{1,2,3}	Distance to Road^{1}	Distance to Road^{1}	Distance to Road^{1}
Distance to Bus Stop^{1,2,3}	Distance to Bus Stop^{3}	Distance to Bus Stop^{1}	Distance to Bus Stop^{1}
Distance to Bus Route^{1}	Distance to Bus Route^{1}	Distance to Bus Route^{1}	Distance to Bus Route^{1}
-	-	Distance to Primary School^{1}	-
Neighbourhood			
Block Agricultural Share^{1,2}	Block Agricultural Share^{1}	-	Block Agricultural Share^{1,2}
Block Vegetation Share^{1,2,3}	-	Block Vegetation Share^{1}	-
Block Agricultural Area^{1}	-	-	-
Block Nature Share^{1}	-	-	-
Cell Single Family Area^{1}	Cell Single Family Area^{1}	-	Cell Single Family Area^{1}
Cell Vacant Share^{1}	Cell Nature Share^{1}	Cell Nature Share^{1}	Cell Nature Area^{1}
Additional Full Model Variables			
Structural (Land)			
-	Perimeter^{1,2}	Perimeter^{2,3}	-
Structural (Building)			
Building Volume^{1}	Building Volume^{1}	-	-
Building Footprint to Area	Building Count^{1}	-	-
Locational			
Distance to Bus Route^{2,3}	Distance to Road^{2}	Distance to Road^{3}	
-	Routing Distance Under 500m to Road^{1}	Distance to Bus Stop^{3}	-
Neighbourhood			
-	Block Vegetation Area^{1}	Block Vegetation Area^{1}	-
-	Block Vegetation Share^{1}	Block Vegetation Share^{1,2}	-
-	Block Agricultural Share^{1,2}	Block Open Space Share^{1}	-
Cell Education Share^{1}	-	Cell Vacant Area^{1}	-
Cell Vegetation Share^{1}	-	Cell Vacant Share^{1}	-
Sector Education Share^{1}	-	Sector Defence Area^{1}	-
-	-	Sector Plantation Area^{1}	-

Notes: Superscripts 1, 2 and 3 refer to linear, quadratic and logarithmic transformations respectively.

For the neighbourhood variables, there is some variety in the types of variables selected across the different parcel grouping models. Furthermore, these variables appear at the block, cell and sector level of aggregation. The main variables that are consistent across all parcel grouping models are agricultural zoning and vegetation cover or block share, mostly at the block level. The other informative variables are single family residential zones, nature zones and vacant zones, often at the cell level.

Finally, it is worth highlighting some of the differences between the parsimonious and full set of variables for each parcel data grouping. For the taxed built parcel grouping, there is only a parsimonious set of variables as all nine of the selected variables were chosen by all ten folds in the mRMR procedure. This is not the case for the other three data groupings and the number of additional variables selected is often fairly large.

6.2 Model Selection for each Parcel Data Grouping

Given the set of full and parsimonious variables selected for each parcel data grouping, we run the entire set of models described in Section 4. In Table 3 we present the benchmark OLS regression model which is used in the mRMR feature selection procedure, and two other models; one that minimises the *in-sample* MAE and one that minimises the *cross-validated* MAE. For each model, we also present the in-sample and cross-validated MAE values for comparison and to highlight the benefits of our cross-validation approach.

A key benefit of the simple OLS model is that the loss in performance between in-sample and cross-validated MAE is very marginal. This is true for all eight OLS results presented in Table 3. This fact weighs in favour of the use of OLS models for the purpose of property valuation when cross-validation is not possible, because some complex spatial models that may perform extremely well in-sample, often suffer from weak cross-validated performance.

This point is best illustrated by evaluating the model diagnostics of the best in-sample model type, which in our case is always the tricube geographic weighted regression (GWR) which significantly outperforms the other model types according to in-sample MAE. However, despite large in-sample improvements, the cross-validated performance of this model type is worse than the benchmark OLS model across all groupings. This further highlights the pitfall of focusing on in-sample metrics in model type selection which would actually have lowered the model's predictive accuracy when compared to the benchmark.

The best cross-validated model type is less consistent, and is split three ways between the spatial expansion and trend surface correction geographic binary weighted regression (GBWR) model types, as well as the Gaussian GWR model type. All of these model types improve upon the benchmark OLS in terms of both the in-sample *and* cross-validated MAE. With the exception of the trend surface correction GBWR model type, the other two model

Table 3: **Best Models for Each Parcel Data Grouping**

			Models' Mean Absolute Error		
Grouping	**Variables**	**Mean Absolute Error Type**	**Benchmark**	**Best In-Sample**	**Best Cross-Validated**
All Parcels	**Parsimonious**		*OLS*	*GWR (Tricube)*	*GBWR (CAS)*
		In-Sample	0.623	0.478	0.538
		Cross-Validated	0.625	0.682	0.585
	Full		*OLS*	*GWR (Tricube)*	*GBWR (TSC)*
		In-Sample	0.617	0.502	0.581
		Cross-Validated	0.620	0.682	0.585
Taxed Parcels	**Parsimonious**		*OLS*	*GWR (Tricube)*	*GWR (Gaussian)*
		In-Sample	0.602	0.404	0.504
		Cross-Validated	0.604	0.637	0.551
	Full		*OLS*	*GWR (Tricube)*	*GBWR (TSC)*
		In-Sample	0.580	0.430	0.540
		Cross-Validated	0.583	0.674	0.541
Taxed Land Parcels	**Parsimonious**	Model	*OLS*	*GWR (Tricube)*	*GWR (Gaussian)*
		In-Sample	0.609	0.332	0.520
		Cross-Validated	0.614	0.673	0.576
	Full	Model	*OLS*	*GWR (Tricube)*	*GBWR (CAS)*
		In-Sample	0.573	0.390	0.517
		Cross-Validated	0.591	0.689	0.579
Taxed Built Parcels	**Full**	Model	*OLS*	*GWR (Tricube)*	*GBWR (CAS)*
		In-Sample	0.577	0.391	0.514
		Cross-Validated	0.579	0.617	0.550

Notes: The in-sample and cross-validated models presented here were the best models for each data grouping. GWR refers to Geographic Weighted Regression with the weighting function in parentheses. GBWR refers to the Geographic Binary-Weighted Regression with the estimation model in parentheses. CAS refers to the spatial expansion model and TSC refers to the trend surface correction model.

types experience a sizable deterioration in model accuracy from in-sample to cross-validated. Whilst we have shown that improvements over OLS in cross validated MAE are possible, it seems difficult to know in advance which model type may outperform OLS; again this supports the use of OLS when it is not possible to test multiple types of spatial models, and the significant additional effort of testing multiple model types may be unlikely to yield a sizeable accuracy benefit, especially when a range of locational variables are included.

We must finally choose a model that is "best" for each parcel data grouping, in terms of cross-validated MAE, from the full or parsimonious sets of variables. For the all parcel and taxed land parcel groupings, the improvements in performance from the full set of variables is very minor while this difference is slightly larger for the taxed parcel grouping. Therefore, we select as the "best" model for each parcel data grouping, the parsimonious set of variables for the all parcel and taxed land parcel groupings, and the full set of variables for the taxed parcel grouping. As the set of variables for the taxed building parsimonious and full models are the same, there is no distinction between the two.

6.3 Model Selection to Obtain the Most Accurate Building Values

To obtain building values, we first obtain predicted land values for all taxed parcels using the best taxed parcel model with the structural building values set to zero. Secondly, we obtain predicted property values using the best taxed parcel model for all taxed parcels. The difference between these two predicted values is the best possible prediction for the building value, provided that the best taxed parcel model (with both built and unbuilt parcels) is not inconsistent with the best taxed land model.

Another step is possible in our model selection and will clarify whether the best taxed parcel model underperforms the best taxed land model, for taxed land parcels. We thus evaluate the four models selected in the previous subsection 6.2 on different parcel groupings. The results of these comparisons are presented in Table 4. Surprisingly, and conveniently, the taxed parcel model performs the best for taxed land parcel data while it is marginally outperformed by the all parcel model for the taxed built parcel data. However, as these differences are small, we select the more parsimonious model of the two which in this case is the taxed parcel model.

***Table 4:* Model Comparison**

		Mean Absolute Error when the Best Parcel Data Grouping Model is Applied			
Grouping	**Mean Absolute Error Type**	**All Parcels**	**Taxed Parcels**	**Taxed Land**	**Taxed Built**
Taxed Parcels	In-Sample	0.506	0.540	0.737	0.554
	Cross-Validated	0.540	0.541	0.748	0.573
Taxed Land Parcels	In-Sample	0.540	0.576	0.520	0.657
	Cross-Validated	0.570	0.568	0.576	0.657
Taxed Built Parcels	In-Sample	0.497	0.530	0.795	0.526
	Cross-Validated	0.532	0.534	0.795	0.550

Notes: For the best taxed parcel models performance on the taxed land parcel grouping data, the cross-validated MAE is better than the in-sample MAE. This is due to the fact that the 10-fold cross-validation technique was performed over the taxed parcel grouping data combined with the fact that the in-sample method is run with the entire set of taxed parcel grouping data.

The fact that the all parcel and taxed parcel models fairly consistently outperform the taxed land and taxed building models for their respective data grouping is most likely driven by the fact that the former models draw on more data to avoid overfitting. Whilst the taxed land model outperforms the other models in its own grouping in terms of in-sample MAE, the taxed parcel model, followed by the all parcel model, outperforms it in terms of the cross-validated MAE.

Therefore, in an upgrade on the "best" models found in section 6.2, we will now use the taxed parcel model to predict both land and property values for all three taxed groupings - given that any differences between the performance of this model and that of the all parcel model are minimal. Using the taxed parcel model to predict both land and property values is equivalent to using the taxed parcel model to predict combined property values for all taxed parcels and then setting all structural building variable values to zero to impute the

land value. Choosing a single model for both is beneficial because using two different models for land and property values could lead to some building values being imputed as negative.

6.4 Final Model Performance

Here we present a more diverse range of model diagnostics to evaluate the performance of the best taxed parcel model in Table 5. In addition to the MAE on which the models were evaluated, we present R², root mean squared error (RMSE), and a statistic we label as ±20% which captures the fraction of predictions which fall within 20% of the true value.

Table 5: **Best Taxed Parcel Model**

		Model Diagnostics			
Category	**Sample**	**MAE**	**R²**	**RMSE**	**±20%**
Taxed Parcels	In-Sample	0.540	0.618	0.729	26.3%
	Cross-Validated	0.541	0.600	0.746	26.6%
Taxed Land Parcels	In-Sample	0.576	0.475	0.757	22.8%
	Cross-Validated	0.568	0.458	0.771	24.5%
Taxed Built Parcels	In-Sample	0.530	0.545	0.721	27.2%
	Cross-Validated	0.534	0.522	0.740	27.1%

Our best taxed parcels model achieves a cross-validated MAE of 0.541, an R² of 0.600, a RMSE of 0.746 and 26.6% of actual values within 20% of the predicted value. Furthermore, there is a noticeable difference in the model's predictiveness between taxed land and taxed built parcels. The model can more reliably predict the combined property values of taxed built parcels than just pure taxed land. This is possibly explained by the inclusion of additional structural building values that are used to predict the property values.

To obtain the most accurate building values possible, as noted, we subtract the predicted land values for all taxed parcels but with the structural building variables set to zero, from the predicted property values with the structural building variables included in the model. Whilst this is the most theoretically accurate building variable possible, it is not possible to directly observe building values and thus to measure the accuracy of our predictions. This underlines the technical difficulty of implementing a policy that stipulates the use of a CAMA for building values only.

7. Conclusions and Recommendations

We construct a set of models of property prices for the province of Kigali, Rwanda, for 2015, aiming to be as accurate as possible, in order to extract methodological lessons for a possible future CAMA to be implemented for Kigali in accordance with the property tax law that passed in 2018. To construct the models, we use digitised sales transaction data from the Land Administration Information System, for 7,445 parcels, divided by the area of the plot in square metres, to construct the dependent variable which was the logarithm of sales value per square metre. We then use a large set of explanatory variables, some of which are extracted from satellite images and processed, and some of which are extracted and processed from Government sources. The explanatory variables include structural land and building variables, locational variables such as distances to amenities and neighbourhood-level variables on land use, land zoning and land cover, land use-related variables and relied on different datasets from the Government of Rwanda, satellite images and GIS mapping tools. This comprises a total of 235 variables, which become 511 after various mathematical transformations (logarithms and squared terms).

Many of these variables are strongly correlated with each other and this represents a challenge for variable and model selection. We use a variable selection technique known as Minimum Redundancy Maximum Relevance, along with a deterministic cross-validation technique to eliminate overfitting. This is a vital step given that our goal is to use the 7,445 parcels for which there is sales data, to generate a model that can most accurately predict property values for all 367,000 parcels in Kigali in 2015 (explanatory variables are available for all parcels), the vast majority of which are out of sample. The critical diagnostic on which we compare models is the cross-validated Mean Absolute Error. We test various model types including Ordinary Least Squares and a wide range of spatial model types, and identify the best performing models for each of four parcel data groupings: (1) all parcels; (2) taxed parcels; (3) taxed land and (4) taxed buildings.

One goal is to find a way to predict building values as accurately as possible, because the Rwanda property tax law mandates the possible use of a Computer Assisted Mass Appraisal to generate building values on the basis of which building tax will be calculated. We find that surprisingly, and conveniently, the best model trained on taxed parcel data outperformed the best model trained on taxed land data only, for taxed land parcels. This means that the best way to find building values is to use the taxed parcel model with the structural building variables set to zero to find imputed land values, and by subtracting these predicted land values from the predicted property values when the structural building variables are included.

Our findings and reflections are as follows:

- Our best taxed parcel model has a cross validated R^2 of 0.600, a cross-validated MAE of 0.541, a cross-validated Root Mean Squared Error (RMSE) of 0.746 and 26.6% of actual values per square metre were within 20% of the cross-validated predicted value.
- When using the same model to predict taxed unbuilt land values, the accuracy reduces slightly, to a cross validated R^2 of 0.458, a cross validated MAE of 0.568, an RMSE of 0.771 and 24.5% of actual values per square metre are within 20% of the cross-validated predicted value. Whilst this accuracy is not ideal, we are confident that it is the best possible on the basis of the 998 available data points for taxed unbuilt land.
- The famous adage "location, location, location" is confirmed: Locational variables, especially distance to a road, a bus stop and a bus route, or logs and squared transformations of these variables, are consistently important, which underlines the importance of the interplay between urban connectivity and property prices. Other land use, land cover and land zoning variables, especially relating to agriculture, nature or vegetation cover, were consistently significant. Structural building variables - especially the building volume per unit land area and its squared term - are consistently important. Finally, a structural variable relating to land - specifically, the parcel perimeter - is present in the taxed parcel model and taxed land model, but not prominent.
- Whilst spatial models generally outperform OLS in terms of in-sample diagnostics, when applied to our data they tend to underperform relative to OLS in terms of cross-validated diagnostics, which implies that the added complexity of the spatial models tend to overfit the data. It follows that such excessively complex models with even the best-looking in-sample diagnostics, perform much worse than OLS for the purpose of out-of-sample property valuation unless they optimise cross-validated diagnostics. Whilst our extensive search for the best spatial model has indeed resulted in a set of models that outperform OLS in terms of cross-validated MAE, it is not possible to predict in advance which spatial model types will outperform OLS for any given dataset or model. We therefore conclude that use of spatial models is desirable if and only if extensive comparisons of different spatial models on the basis of cross validated diagnostics, is possible.

- Any property valuation model loses accuracy over time, as the explanatory variables change, and as inflation takes place; our analysis also shows this. It will be necessary to recalibrate the property valuation model behind a CAMA with up-to-date data on explanatory variables, every three to five years - at a rate that balances accuracy with cost.

- However, the characteristics of properties change even from year to year, so value estimates for a property can be updated more regularly by adjusting the values of variables - for example if a road is built which decreases the distance to the nearest road for a number of properties and thus increases their value. In Rwanda, the property tax law specifies that a CAMA can only be used to tax buildings, and thus the Government would only need to be concerned with updating the building values. Where the single model approach to calculating building values described in section 6.3 is used, the only data it would be necessary to update between full model re-calculations would be the structural building variables, which in our model are the built volume per unit area and its squared term. This is because the other variables make no difference: building values can be calculated by simply multiplying the structural building variables - namely built volume per unit area and its squared term - by their model coefficients.[9]

- Property tax law would ideally account for the technical limits of property valuation models in terms of accuracy. A property valuation model can most accurately predict total property values, and can also predict land values less accurately; however, it is hardest to predict building values accurately, as is necessary for Rwanda. Moreover, it is impossible to know how accurate the model is, given that buildings are not sold separately from the land on which they are built. In future, the Government of Rwanda might consider applying the CAMA either to all property, or to land, not solely to buildings.

- The data used in this paper is almost exclusively of a "top down" nature. Additional "on the ground" building data would probably benefit model accuracy, including building materials, numbers of bedrooms, and similar details. Rwanda Revenue Authority plans to collect this data, which should be included in a future property valuation model if it is of sufficient quality and coverage.

[9] The coefficients are constant in the case of OLS but vary across space for spatial models; in the latter case the coefficients would be already known from when the model was originally estimated.

References

Ali, Daniel Ayalew; Deininger, Klaus W.; Wild, Michael. 2018. "Using satellite imagery to revolutionize creation of tax maps and local revenue collection" (English). Policy Research working paper; no. WPS 8437. Washington, D.C. : World Bank Group. http://documents.worldbank.org/curated/en/347231526042692012/Using-satellite-imagery-to-revolutionize-creation-of-tax-maps-and-local-revenue-collection

Anas, A., Arnott, R., & Small, K. A. 1998. "Urban Spatial Structure". *Journal of Economic Literature, 36*(3), 1426-1464.

Arora, S., Little, M. A. and McSharry, P. E. 2013. "Nonlinear and Nonparametric Modelling Approaches for Probabilistic Forecasting of the US Gross National Product." Studies in Nonlinear Dynamics and Econometrics, 17(4), 395-420.

Bachofer, F., and Murray S., 2018. "Remote sensing for measuring housing supply in Kigali." Policy paper, International Growth Centre, London.

Bachofer, Felix & Braun, Andreas & Adamietz, Florian & Murray, Sally & d'Angelo, Pablo & Kyazze, Edward & Mumuhire, Abias & Bower, Jonathan. 2019. "Building Stock and Building Typology of Kigali, Rwanda". Data. 4. 105. 10.3390/data4030105.

Bidanset, P. E., & Lombard, J. R. 2014. "The effect of kernel and bandwidth specification in geographically weighted regression models on the accuracy and uniformity of mass real estate appraisal". Journal of Property Tax Assessment & Administration, 10(3), 5-14.

Blaschke, T. 2000. "Object based image analysis for remote sensing", ISPRS Journal of Photogrammetry and Remote Sensing, 65. 10.1016/j.isprsjprs.2009.06.004.

Bourassa, S. C., Cantoni, E., & Hoesli, M. 2007. "Spatial Dependence, Housing Submarkets, and House Price Prediction". *The Journal of Real Estate Finance and Economics, 35*(2), 143-160. doi:10.1007/s11146-007-9036-8

Bower, J., & Murray, S., 2019. "Housing need in Kigali". Policy paper, International Growth Centre.

Brasington, D. M., & Hite, D. 2005. "Demand for environmental quality: a spatial hedonic analysis." *Regional Science and Urban Economics, 35*(1), 57-82. doi:10.1016/j.regsciurbeco.2003.09.001

Cellmer, R. 2014. "The Possibilities and Limitations of Geostatistical Methods in Real Estate Market Analyses." In *Real Estate Management and Valuation* (Vol. 22, pp. 54).

Chew, R., Jones, K., Unangst, J., Cajka, J., Allpress, J., Amer, S., & Krotki, K. 2018. "Toward Model-Generated Household Listing in Low- and Middle-Income Countries Using Deep Learning". *ISPRS International Journal of Geo-Information, 7*(11). doi:10.3390/ijgi7110448

Chrostek, K. & Kopczewska, K., 2013 - “Spatial Prediction Models for Real Estate Market Analysis”. Ekonomia journal, Vol: 35: 35.

Dąbrowski, R., & Latos, D. 2015. “Possibilities of the Practical Application of Remote Sensing Data in Real Property Appraisal”. In *Real Estate Management and Valuation* (Vol. 23, pp. 68).

Deane, G., & Owen, R. 2019. “Cost-Effectiveness Analysis of a Satellite-Based Approach to Maintaining a Property Database”. Paper presented at the 2019 World Bank Conference on Land and Poverty, Washington DC, USA.

Demetriou, D. 2016. “The assessment of land valuation in land consolidation schemes: The need for a new land valuation framework”. *Land Use Policy, 54*, 487-498. doi:10.1016/j.landusepol.2016.03.008

Glaeser, E. L., & Gyourko, J. 2002. “Zoning's Steep Price”. *Regulation, 25*(3), 24-30.

Glaeser, E. L., & Ward, B. A. 2009. “The causes and consequences of land use regulation: Evidence from Greater Boston”. *Journal of Urban Economics, 65*(3), 265-278. doi:10.1016/j.jue.2008.06.003

Hanchuan Peng, Fuhui Long, and Chris Ding. 2005. “Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp.1226-1238, 2005.

Harrison Jr, D. & Rubinfeld, D. L., 1978. “Hedonic housing prices and the demand for clean air”. Journal of Environmental Economics and Management, 1978, vol. 5, issue 1, 81-102.

Hastie, T. Tibshirani, R. and Friedman, J. H. 2009. “The Elements of Statistical Learning: Data Mining, Inference, and Prediction”. 2nd Edition. Springer, London.

Heikkila, E., Gordon, P., Kim, J. I., Peiser, R. B., Richardson, H. W., & Dale-Johnson, D. 1989. “What Happened to the CBD-Distance Gradient?: Land Values in a Polycentric City”. *Environment and Planning A, 21*(2), 221-232. doi:10.1068/a210221

Jayyousi, M., Estima, J. & Ghedira, H., 2014 - “Spatial Multi Criteria Decision Analysis Based Assessment of Land Value in Abu Dhabi, UAE”. - In: “Group Decision and Negotiation. A Process-Oriented View”: Joint INFORMS-GDN and EWG-DSS International Conference, GDN 2014, Toulouse, France, June 10-13, 2014. Proceedings, Zaraté, P. et al. (Eds.); Springer International Publishing: Cham, 120-127.

Jiang, S., Alves, A., Rodrigues, F., Ferreira, J., & Pereira, F. C. 2015. “Mining point-of-interest data from social networks for urban land use classification and disaggregation”. *Computers, Environment and Urban Systems, 53*, 36-46. doi:10.1016/j.compenvurbsys.2014.12.001

Kim, B., & Kim, T. 2016. “A Study on Estimation of Land Value Using Spatial Statistics: Focusing on Real Transaction Land Prices in Korea”. *Sustainability, 8*(3), 203. doi:10.3390/su8030203

Lan, F., Wu, Q., Zhou, T., & Da, H. 2018. "Spatial Effects of Public Service Facilities Accessibility on Housing Prices: A Case Study of Xi'an, China". *Sustainability, 10*(12). doi:10.3390/su10124503

McSharry, P. E. and Smith, L. A. 2004. "Consistent Nonlinear Dynamics: identifying model inadequacy". Physica D 192: 1-22.

Ndegwa, James. 2018. "Determinants of Apartment Prices within Housing Estates of Nairobi Metropolitan Area". International Journal of Economics and Finance. 10. 104. 10.5539/ijef.v10n6p104.

Orford, S. 2002. "Valuing Locational Externalities: A GIS and Multilevel Modelling Approach". *Environment and Planning B: Planning and Design, 29*(1), 105-127. doi:10.1068/b2780

Owusu-Ansah, Anthony. 2012. "Examination of the Determinants of Housing Values in Urban Ghana and Implications for Policy Makers". Paper presented at the African Real Estate Society Conference in Accra, Ghana from 24 -27 October 2012.

Rotimi Boluwatife Abidoye & Albert P. C. Chan. 2017. "Modelling property values in Nigeria using artificial neural network", Journal of Property Research, 34:1, 36-53, DOI: 10.1080/09599916.2017.1286366

Rosen, S. 1974. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition". *The Journal of Political Economy, 82*(1), 34-55. doi:10.1086/260169

Sasaki, M., & Yamamoto, K. 2018. "Hedonic Price Function for Residential Area Focusing on the Reasons for Residential Preferences in Japanese Metropolitan Areas". *Journal of Risk and Financial Management, 11*(3). doi:10.3390/jrfm11030039

Shonkwiler, J. S., & Reynolds, J. E. 1986. "A Note on the Use of Hedonic Price Models in the Analysis of Land Prices at the Urban Fringe". *Land Economics, 62*(1), 58-63.

Sirmans, G. S., MacDonald, L., Macpherson, D. A. & Zietz, E. N., 2006 - "The Value of Housing Characteristics: A Meta Analysis". The Journal of Real Estate Finance and Economics, Vol: 33 (3): 215-240

Song, Y., & Sohn, J. 2007. "Valuing spatial accessibility to retailing: A case study of the single family housing market in Hillsboro, Oregon". *Journal of Retailing and Consumer Services, 14*(4), 279-288. doi:10.1016/j.jretconser.2006.07.002

Srour, I., Kockelman, K., & Dunn, T. 2002. "Accessibility Indices: Connection to Residential Land Prices and Location Choices". *Transportation Research Record: Journal of the Transportation Research Board, 1805*, 25-34. doi:10.3141/1805-04

Tsanas, A., M.A. Little, P.E. McSharry. 2013. "A methodology for the analysis of medical data", Chapter 7 in Handbook of Systems and Complexity in Health, pp. 113-125, Eds. J.P. Sturmberg, and C.M. Martin, Springer, 2013

Whittle, Jennifer & Barry, Michael. 2004. “Fiscal Cadastral Reform and the Implementation of CAMA in Cape Town”. Paper presented to 3rd FIG Regional Conference Jakarta, Indonesia, October 3-7, 2004.

Wyatt, P. 1996. “Using a geographical information system for property valuation”. *Journal of Property Valuation and Investment, 14*(1), 67-79. doi:10.1108/14635789610107507

Xiao, Y. 2017. “Hedonic Housing Price Theory Review”. In *Urban Morphology and Housing Market* (pp. 11-40): Springer Geography.

Yomralioglu, T., & Nisanci, R. 2004. “Nominal Asset Land Valuation Technique by GIS”. Paper presented at the FIG Working Week, Athens, Greece.

Zainora, A. M., Norzailawati, M. N., & Tuminah, P. 2016. “A Spatial Analysis on Gis-Hedonic Pricing Model on the Influence of Public Open Space and House Price in Klang Valley, Malaysia”. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLI-B8*, 829-836. doi:10.5194/isprsarchives-XLI-B8-829-2016

The International Growth Centre (IGC) aims to promote sustainable growth in developing countries by providing demand-led policy advice based on frontier research.

Find out more about
our work on our website
www.theigc.org

For media or communications
enquiries, please contact
mail@theigc.org

Subscribe to our newsletter
and topic updates
www.theigc.org/newsletter

Follow us on Twitter
@the_igc

Contact us
International Growth Centre,
London School of Economic
and Political Science,
Houghton Street,
London WC2A 2AE

IGC
International
Growth Centre

DIRECTED BY
LSE
UNIVERSITY OF
OXFORD

FUNDED BY
UKaid
from the British people

Designed by soapbox.co.uk