Cape Town Airbnb Price Prediction: An exploration of spatial statistic and machine learning methods

Courtney Williams

A Minor Dissertation Submitted for the Degree of Master of Science in Data Science

University of Cape Town
Department of Statistical Sciences

February 2023

Supervisors: Mr Sulaiman Salau
Dr Sebnem Er

Abstract

This thesis predicts the prices of Airbnb listings in Cape Town, South Africa and in doing so, investigates the price determinants in the market. Using data from InsideAirbnb, traditional, spatial and machine learning models are compared and contrasted. The Cape Town Airbnb market has significant spatial correlation and heterogeneity, and traditional models such as OLS regression do not account for this spatial dependence, however, it is addressed by spatial models. By accounting for spatial effects, model predictive performance does improve, but not so much as to outperform non-spatial, non-linear machine learning model predictions. While Airbnb is a new and unique platform, the most important price determinants are consistent with those of traditional housing and accommodation markets such as property type, location and amenities.

Declaration

This thesis is an original work of my research and contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Signature: _____

Print Name: Courtney Williams

Date: _____

Acknowledgements

I would like to thank my supervisors, Mr Sulaiman Salau and Dr Sebnem Er, for their consistent guidance and encouragement which has made this thesis possible. Additionally I would like to thank the larger UCT statistics department for its contribution to my undergraduate and postgraduate education which I value endlessly. Lastly, I would like to thank my family and friends without whose support I would not be here today.

Contents

List of Figures

List of Tables

Abbreviations

OLS	**O**rdinary **L**east **S**quare
GWR	**G**eographically **W**eighted **R**egression
SLM	**S**patial **L**ag **M**odel
GLM	**G**eneralised **L**inear **M**odel
SEM	**S**patial **E**rror **M**odel
SDM	**S**patial **D**urbin **M**odel
SVM	**S**upport **V**ector **M**achine
GBM	**G**radient **B**oosting **M**odel
NLP	**N**atural **L**anguage **P**rocessing
COD	**C**oefficient **O**f **D**ispersion
LM	**L**agrangian **M**ultiplier
PVO	**P**redicted **V**erse **O**bserved

Chapter 1

Introduction

1.1 Airbnb

Founded in 2008, Airbnb is an American peer-to-peer online platform where hosts can create rental listings and guests can book short term accommodation options around the world. After creating a profile, anyone can request to list or book a property on the platform. A host's listing gets approved by Airbnb and then published publicly. Guests can browse the available listings on the Airbnb website, request to book them for a specified period of time, and then rent them for the host-approved period (Airbnb, 2018). The accessibility of the platform's offerings has meant that like other peer-to-peer networks such as Uber, it has grown rapidly into a global market disruptor (Anwar, 2018). The company listed on the NASDAQ stock market in 2020 and in 2022 had a 3rd quarter revenue of $2.9 billion, up 29% year on year, and 99.7 million bookings, up 25% year-on-year (Airbnb, 2022). By 2019, nine years after being introduced in South Africa, there were approximately 45 500 local hosts, and the platform contributed R11 billion to the country's GDP, up 29% from 2018 (Short et al., 2021).

Due to Airbnb's growing prevalence and popularity, a large body of research has been conducted around its platform. The research varies from Airbnb's impact on hotels and traditional accommodation (Zervas, Proserpio and Byers, 2017 ; Oskam and Boswijk, 2016) to the potential need for its regulation (Wegmann and Jiao, 2017) to its overall impact on local tourism and communities (Guttentag, 2015). One advantage of Airbnb that is frequently highlighted is its potential to generate revenue for host communities (Henama, 2018; Basuroy, Kim and Proserpio, 2020; Short et al., 2021). Consequently a variety of research is being conducted around

price prediction and price determinants of Airbnb listings in order to better understand the market dynamics and optimise potential revenue.

1.2 Problem statement

There is large diversity in Airbnb offerings and the market dynamics vary greatly by geography (Adamiak, 2022). While there are many studies focused on predicting Airbnb listing prices in various cities across the world, to the best of the author's knowledge, Airbnb price prediction has not been conducted in South Africa. This thesis aims to address the gap in the literature by predicting prices of listings in the South African city of Cape Town using an open source data set from InsideAirbnb (2023). Cape Town is chosen as it is South Africa's premier tourist destination and it is the city with the highest number of Airbnb listings nationally (Short et al., 2021). In addition, Cape Town's diverse geography with long coastlines, prominent mountains and formal and informal settlements (Pirie, 2017) make for an interesting case study.

In addition to predicting listing prices, this thesis aims to investigate price determinants in Cape Town's Airbnb market through the use of supervised learning models. The listing price is treated as the target variable whose value changes based on listing-related information that form the predictor variables. In addition to traditional continuous and categorical predictor variables such as number of bedrooms and type of listing, the Airbnb data also has location information, such as geographic coordinates and neighbourhoods, that can be incorporated into the model. Given the complex and geospatial nature of the data, numerous methods such as those that account for non linearity and spatial effects are explored. Their outputs and performance are compared and contrasted.

To set the scene and give insight into the research that has been conducted in the space to date, a literature review is compiled in Chapter Two. The review also informs the choice of models to be used and explored in Chapter Three. Thereafter, the InsideAirbnb (2023) data set used in the prediction is explored in Chapter Four, before the results are presented in Chapter Five, with conclusions detailed in Chapter Six.

Chapter 2

Literature Review

This chapter contextualises the Airbnb price prediction problem by presenting an overview of the research that has been conducted in the space to date. In addition to research directly relating to Airbnb, research centred around international and local traditional real estate and tourism is provided in order to offer a broadened perspective on the problem.

2.1 Price prediction in real estate and travel accommodation markets

While academic research into the Airbnb platform and its price prediction problem are fairly new, the Airbnb rental concept is closely related to traditional real estate and travel accommodation and their associated pricing problems which have been around for a long time.

Rosen (1974) formulated the hedonic model; a model which states that the price of a marketed product, for example a house, can be modelled as a function of its measurable characteristics and that the effect of the characteristics on the product's price can be quantified. Since its introduction, the hedonic model framing has been widely used in real estate research to determine property prices based on property characteristics or similarly to assess the contribution of various characteristics to the price. Early, and frequently cited, examples include those by Harrison Jr and Rubinfeld (1978) and Li and Brown (1980) who both apply hedonic models to the Boston, USA housing data in their attempts to provide a base for predicting price effects. They highlight that in addition physical property attributes like property size, number of rooms

and its distance to non-residential land areas, variables which are more complex to measure such as noise pollution and air quality also affect price.

Hedonic models have also been used in the context of hotel room accommodation to identify and quantify the variables that affect the price paid by guests for hotel rooms in Spain (Espinet et al., 2003), Taiwan (Chen and Rothschild, 2010) and Portugal (Soler et al., 2019). Studies in coastal holiday regions by Espinet et al. (2003) and Soler et al. (2019) find that the type of hotel and its star rating, as well as the distance to the beach and availability of a golf course and parking are important factors in determining room price. Contrasted to the city setting of Taipei, Taiwan where TV, internet and conference room access are all important (Chen and Rothschild, 2010).

Given its previous applications, it is unsurprising that hedonic models are a natural starting point to determine Airbnb prices and factors affecting price. While the relationship between price and price determinants in hedonic models can be mapped in numerous ways, a common way to is to infer a linear relationship through the use of ordinary least squares (OLS) regression where price is modeled as a linear combination of price determinants. A study by Gibbs et al. (2018) uses Airbnb data from five Canadian cities and applies a hedonic OLS regression model with price as the dependent variable. The independent variables in the study not only include traditional real estate variables like number of bedrooms, presence of amenities such as a pool or gym, but also include Airbnb specific variables like whether the host is a 'superhost' or whether the listing has 'instant booking' available. A 'superhost' is a host that consistently receives high reviews and a listing that has 'instant booking' does not require the host to approve a guest's booking request. Gibbs et al. (2018) find that while the Airbnb platform promotes a unique spin on short term rental accommodation, traditional characteristics such as size and location are still the most important determinants of listing price. A similar study by Dogru and Pekin (2017) using OLS on AirDNA data, a third party source of Airbnb data (AirDNA, 2017), from Boston, USA finds that space and privacy are the most important factors - where privacy refers to renting an entire place as opposed to just a room or a shared room.

Across all of the housing, hotel and Airbnb studies reviewed, locational variables are identified as some of the most important factors affecting price, where the locational variables include distance to a point of interest such as city centre or categorical variables indicating neighbourhood or suburb. Conceptualising location in this framework does not account for two of the spatial features of the Airbnb price problem, namely spatial correlation and spatial non-stationarity.

Spatial correlation refers to the dependence or association of objects on or with each other based on their spatial proximity to each other. Spatial non-stationarity refers to heterogeneity based on location, in other words, the non-uniform variation of object properties across space (Anselin, 2005). Spatial models have been developed to account for these phenomena.

2.1.1 Spatial models

The first law of geography proposed by Tobler (1970) states that "Everything is related to everything else, but near things are more related than distant things". In real estate and accommodation research, this translates to house, hotel or Airbnb prices being more dependent on the listings located in close proximity than those further apart. It is likely that in some areas, certain housing characteristics are more important and carry more weight than in other areas due to localised supply and demand (Helbich et al., 2014). For example, a pool might be more sought after, and therefore more important in the suburbs than in the city centre.

In an attempt to address spatial heterogeneity Bitter, Mulligan and Dall'erba (2007) explore the use of a geographically weighted regression (GWR) model and a spatial expansion model to predict housing prices in Arizona, USA. Both models allow parameters to vary over space; the expansion model does so by interacting housing characteristics with location characteristics and the GWR specifies separate OLS regression equations for each housing observation. GWRs incorporate a kernel function and bandwidths to optimise weights. Bitter, Mulligan and Dall'erba (2007) use a Gaussian adaptive spatial kernel. The Gaussian kernel allows for distance decay where the magnitude of weightings, and therefore their effect, reduces the further away the listings are. The adaptive kernel allows the bandwidth to vary depending on the density of house listings around each point. This means a smaller geographical area is considered when there are more houses; and a larger geographical area when there are fewer houses.. To address spatial correlation, they also test the inclusion of a spatial lag variable in the expansion model which is calculated as the distance weighted average price of each listing's nearest neighbors. The results show that the inclusion of the spatial lag variable improves the performance of the expansion model. The GWR, however, outperforms the expansion model in terms of explanatory power and predictive accuracy, even when the lag term is included in the expansion model.

A similar study comparing the performance of GWRs to spatial lag models (SLMs) on a mass real estate appraisal in Virginia, USA is conducted by Bidanset and Lombard (2014). The authors were concerned with predictive performance in relation to uniformity and equity of the

pricing. They were interested in the minimisation of variation over the entire spatial region and measure this using a coefficient of dispersion (COD). They found that a GWR with a Gaussian kernel function achieves more uniform results than a SLM. Notably, while the GWR achieves the lowest COD at the aggregate level for the entire region, it is outperformed by the SLM in certain neighbourhoods when considering these neighbourhoods in isolation. The conclusion from this study is for modelers to explore different models in different locations in order to optimise results.

Using Airbnb data from Tennessee, USA a study by Zhang et al. (2017) compares a generalised linear model (GLM) to its GWR counterpart. Also opting for a Gaussian kernel function, they found the GWR to be a better technique for investigating price determinants with variation in variable coefficients displaying spatial heterogeneity and improving overall explained variation. For both models, distance to a convention centre had a significant negative impact on price. In the GWR model, the effect is stronger in central Nashville, Tennessee than for the more remote areas of Nashville, Tennessee.

Related to price determinants, Xu et al. (2020) uses OLS and GWR models to investigate the spatial distribution features of Airbnb listings in London, England and their relationship with neighbourhood environments. Using data sourced from InsideAirbnb (an open source data platform), Openstreetmap.com and Tripadvisor, they use a kernel density estimation to get a reading of density of Airbnbs in wards, and then create lists of attractions or points of interest within the wards. They group the points of interest by type and weight them by popularity. Using OLS and GWRs, the density is then regressed on points of interest for each ward. The results show that London Airbnbs are mainly located in the city centre and around tourist attractions and that the GWR has a better model fit than OLS. The difference in the coefficients in regions shows the existence of significant spatial non stationarity. It also implies that GWR may better explain the impact of environment factors on the spatial distribution of Airbnb listings than OLS (Xu et al., 2020).

Based on the findings by Bailey, Muth and Nourse (1963) that including repeat sales improves efficiency and accuracy of real estate pricing regression models, a study by Deboosere et al. (2019) that used AirDNA data for New York, USA only considers listings that generated revenue in at least two separate months. With a focus on price and revenue prediction, they used three categories of independent variables: structural, host, and location/ neighbourhood. To account for the variation within and among neighbourhoods, a multilevel regression model with three

levels (listing level, census tract and borough) is used. They found that locational factors, above all transit accessibility to jobs, and neighbourhood variation have a large impact on both price per night and monthly revenue and, additionally, that seasonality has a significant impact on the price per night (Deboosere et al., 2019).

Spatial models are used to investigate the distribution and factors affecting the distribution of Airbnbs in Rome, Italy (Crisci et al., 2022) and Suzhou, China (Sun, Wang and Hu, 2022). The Italian study used an OLS, Spatial Error Model (SEM) and a spatial lag model (SLM) to investigate the Rome Airbnb market. The authors found that the concentration of Airbnbs to be correlated with tourist attractions and that both the SLM and SEM models are significant, and have better fits than the OLS model. The SLM has a slightly better fit than the SEM (Crisci et al., 2022). In the Chinese study, through the use of a Spatial Durbin model (SDM) incorporating signifcant SEM and SLM effects, traditional folk dwellings, restaurants and shopping malls are found to affect Airbnb distributions (Sun, Wang and Hu, 2022).

From the literature, it is evident that spatial models can be used in a variety of ways to explore the spatial aspects of Airbnb data. Spatial methods appear to frequently fit better than their OLS equivalents and methods such as GWR appear to outperform OLS in terms of prediction. All of the models discussed in this section assume a linear relationship between dependent and independant variables. Machine learning models model non linear relationships and when applied to Airbnb data they are frequently found to perform well.

2.1.2 Machine learning models

Due to their impressive predictive power, machine learning models are a natural choice for inclusion to the Airbnb supervised learning price prediction problem. Tang and Sangani (2015) used support vector machines (SVMs) to predict neighbourhood price ranges for Airbnb listings in San Francisco, USA. Cai, Han and Wu (2019) used Melbourne, Australia Airbnb data to predict listing prices using various methods such as OLS, ridge regression, SVMs, random forests, gradient boosting machines (GBMs) as well as neural networks to predict price. They found that GBMs perform the best but random forests and neural networks also perform competitively.

Due to the fact that Airbnb listings have text information in the form of listing descriptions and user reviews, some studies employ the use of natural language processing (NLP) to create features to feed into their supervised models. Tang and Sangani (2015) created bag of words,

word class and sentiment features to feed into their SVM while Cai, Han and Wu (2019) also created vector representations for words that is fed into their neural network. In addition, Tang and Sangani (2015) incorporate visual features into their SVMs which are created using listing photographs and computer vision packages to identify listing properties.

GBMs and neural networks were also applied to a Beijing InsideAirbnb data set by Yang (2021). They found about a third of the features to have a near-zero feature importance value in the GBM, and the number of people the listing accommodates as the most important feature. The inclusion of house price and distance to subway also improved both GBM and Neural Network models, but the GBM achieves a better test set mean square error.

Unsupervised learning was incorporated into the price prediction by hypothesising that the distance to the landmarks and the popularity of the landmarks usually are latent factors in house pricing. Multi-Scale Affinity Propagation (MSAP) were performed which involved first clustering the landmarks and then clustering houses based on their infrastructure.

In a study by Li et al. (2016), unsupervised learning is incorporated into the price prediction by hypothesising that distance to the landmarks and the popularity of the landmarks usually are latent factors in house pricing. Multiscale clustering is used by way of Multi-Scale Affinity Propagation (MSAP) (and Kmeans and DBSCAN) which involves initially clustering landmarks and then clustering houses (within the landmark clusters) based on their infrastructure. Thereafter, within clusters, OLS is used to predict pricing. Regression inputs are facilty factors, distance to nearest landmark and popularity of landmark. Li et al. (2016) conducted the investigation for Boston and Los Angeles, USA, London, England and Tokyo, Japan, using data from Airbnb and renting information sites and Tripadvisor (for landmark clustering) and found that the pricing varies between clusters showing strong non-stationarity in all cities.

There are numerous machine learning models that can be incorporated into the Airbnb prediction problem and they have all shown demonstrable value. As with the spatial and non-spatial Airbnb prediction literature, the bulk of the research has been conducted outside of Africa.

2.2 Real estate price prediction and the Airbnb and tourism market in Cape Town

While Airbnb prediction research has mostly been conducted on data sets pertaining to first world cities, closely related research has been conducted into property price evaluation in Cape Town, South Africa as well as into Cape Town tourism and Airbnb market dynamics.

In housing market evaluations, Yacim and Boshoff (2019) explored the predictive performance of fixed and adapative Gaussian and Bi-square kernel functions using GWRs and compared them to their OLS counterparts. They found that OLS coefficients do not reflect the true relationship within the housing data sets, as the omission of the spatial variables meant the spatial effects were unaddressed causing bias in the coefficient estimates. They found the adaptive Gaussian kernel to perform the best however, while it was able to correct for spatial heterogeneity, it was unable to achieve low CODs such as those achieved by Bidanset and Lombard (2014).

In a separate study, the same authors explored the predictive accuracy of OLS, neural networks, M5Ptrees, SVMs and additive non-parametric regression on mass appraisal property valuations and found that neural networks outperformed the rest of the models followed closely by M5Ptrees (Yacim and Boshoff, 2016). In another investigation comparing neural networks to linear, semi-log and log-log models, they found Levenberg-Marquardt trained neural networks outperformed in-prediction accuracy and reliability ranking order, but cautioned that a semi-log linear model should be preferred in mass appraisals over black box neural networks due to its explainability (Yacim and Boshoff, 2018).

Subroyen, Turpin and de Waal (2021) performed topic modelling on Cape Town InsideAirbnb data. The authors distinguished between marketer-generated content (MGC) and user generated content (UGC) and highlighted that MGC data is largely related to property specifics while UGC, such as user reviews, tend to be centred more around experience and sentiment. Subroyen, Turpin and de Waal (2021) pointed out the potential for hosts to improve their offers, and therefore ability to increase listing price, by addressing the themes in UGC.

Adamiak (2022) highlighted that the diversity of Airbnb activity and offers is often underrated. The platform caters for a diverse audience, listing everything from spare rooms in a small houses to entire luxury mansions. The platform attracts local and international travelers (Short et al., 2021) and competes with hostels, bed and breakfasts, hotels, traditional vacation rentals and the like (Adamiak, 2022).

The tourism sector in Cape Town is very diverse and there is a large contrast between the activities on the Atlanic Seaboard and in the City Bowl to those in the previously disadvantaged areas of Khayelitsha, Langa and Gugulethu to name a few (Pirie, 2017). This diversity has filtered into the Airbnb market where there are upmarket offerings on the Atlantic Seaboard but also 'township experience' offerings in Khayelitsha (Henama, 2018). There have been campaigns and drives by Airbnb to promote township tourism in Cape Town and, while they have seen some success and encouraged local entrepreneurship, the effects of the geographic boundaries laid out during apartheid are still felt today meaning that activities are still largely segregated and distinct for different parts of the city (Hofäcker and Gebauer, 2021; Henama and Mathole, 2022).

Both the predictive and qualitative local studies point to the evidence of spatial dependence and heterogeneity in the housing, tourism and Airbnb market in Cape Town.

2.3 Summary

The hedonic model pricing framing is frequently used in the housing and property rental market and, more recently, also in the Airbnb market. Breaking down the price of traditional housing and short term rental into measurable factors, physical factors such as property size are often the most important price determinants, however, other factors such as relative location has been consistently found to be important. In the Airbnb rental price prediction literature, the same holds true with physical factors, amenities and location also being the most important price determinants, even in the presence of unique Airbnb-specific determinants like being a 'superhost'. For hotels and Airbnbs, the importance of the factors also vary based on location and also by time of year.

Traditional OLS models do not account for the spatial dependence inherent in locational Airbnb data. A variety of spatial models have been deployed in the literature such as SLM, SER and GWR. Spatial models are used in conjunction with a spatial weights matrix and, in the reviewed studies, the Gaussian kernel function is the most frequently used and adaptive kernels are preferred to fixed kernels. In the reviewed literature, GWR is frequently found to outperform not only OLS but also other spatial models.

Machine learning models are also used to predict on spatial data and they have been found to perform well on Airbnb and housing data. Neural networks and tree-based methods often perform best, however Neural networks have the disadvantage of being less interpretable. Natural language processing and computer vision have also been incorporated into the Airbnb prediction space. Given that there is textual data as well as photographs of listings, this has been used to create additional information around sentiment and experience. Incorporating the additional information into price prediction has been found to improve prediction.

While, to the best of the author's knowledge, Airbnb price prediction in South Africa has not been conducted, there is related research focusing on Cape Town housing pricing and the Cape Town Airbnb and tourism market in general. The housing pricing prediction literature found significant spatial dependence and heterogeneity in Cape Town housing prices and found that model prediction performance was not consistent and varied by location. The research into the Cape Town Airbnb and tourism market in general highlights a similar spatial variability calling attention to the contrast between the tourism offering and settlement patterns in places such as the Atlantic Seaboard and the Cape Flats.

The next chapter provides further explanation of some the models discussed in the literature review and used in the analysis.

Chapter 3

Methods

This chapter provides an analysis of multiple models, examining their formulations, strengths, weaknesses, and unique characteristics through comparisons.

3.1 Ordinary least squares regression (OLS)

One of the most widely used supervised learning techniques is ordinary least squares (OLS) regression. OLS models a linear relationship between a single dependent variable and one or more independent variables.

For a single observation, i, the problem is formulated as in Equation (3.1) where y_i is the single dependent variable and x_{i1} to x_{ip} are the p independent, explanatory variables. β_0 is the intercept, β_1 to β_p are the independent variable coefficients and ε_i the random error.

The $\boldsymbol{\beta}$ parameter estimates, $\hat{\boldsymbol{\beta}}$, are found by minimising the square of the residuals, which are the differences between the predicted, $\hat{y}_i$, and observed, y_i, dependent variables (Equation 3.2).

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i. \tag{3.1}$$

$$\begin{aligned} \varepsilon_i &= y_i - \hat{y}_i. \\ \hat{\boldsymbol{\beta}} &= \min\{\varepsilon_i^2\}. \end{aligned} \tag{3.2}$$

For a sample of size n, the problem can be represented in matrix notation as in Equation (3.3) and condensed to Equation (3.4).

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{3.3}$$

$$\begin{aligned} \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \varepsilon_i \overset{iid}{\sim} (0, \sigma^2) \\ \hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}. \end{aligned} \tag{3.4}$$

OLS regression is widely used for its simple formulation, however it is limited in that it assumes a linear relationship between the dependent and independent variables. Additionally, it requires the n error terms to be independent and identically distributed with a zero mean and constant variance as in Equation (3.4). The errors are also assumed independent from the independent variables, and, as the $\boldsymbol{\beta}$ parameters are constant for all observations, it assumes the relationship between dependent and independent variables to be constant across space.

A non-linear regression model could be used to address the linearity assumption however it still requires the assumptions of independence. Spatial models are typically employed to address this issue as they allow for dependence in the variables as well as the error terms.

3.2 Geospatial data and spatial models

Geospatial data represents features or objects on the earth's surface. Latitude and longitude co-ordinates reference specific points on the earth's surface and can be considered point data, while polygon or lattice data represent areas. As stated in the first law of geography, nearer things are more related than distant things (Tobler, 1970). This results in spatial correlation and spatial heteroscedasticity. Spatial correlation, or spatial dependence, can occur as spatial error, where observation error terms are correlated with each other, or as spatial lag, where the dependent variables are affected not only affected by their own independent variables, but also the independent variables of other observations (Bivand et al., 2008). Spatial heteroscedasticity occurs when the importance of independent variables vary over space.

3.2.1 Spatial error and spatial lag models

Spatial error models (SEM) and Spatial lag models (SLM) and are auto-regressive extensions of the OLS regression. They are auto-regressive in that weighted measures of the error term or the response variable are included as regressors.

Spatial error model

In the case that spatial variation is not fully accounted for by the independent variables, the error terms may be correlated, violating the assumptions of OLS and affecting prediction. The SEM in Equation (3.5) addresses this spatial dependence in the error term, ε, through the use of an autocorrelation parameter, λ, and a weighting of other observation errors, $\mathbf{W}u$. The value of λ can be varied to control the size of the effect of the autoregressive $\mathbf{W}u$. Higher values of λ indicate a more significant autoregessive effect, while a value of zero simplifies the model to OLS regression (Fischer and Wang, 2011).

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \tag{3.5}$$

$$\mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}.$$

Spatial lag model

When dependent variables are affected not only by their own independent variables, but also those of other observations, the SLM presented in Equation (3.6) accounts for this form of spatial interaction. Given that dependent variables, $\mathbf{y}$, are functions of their independent variables, the interaction is accounted for through the addition of a weighted lag term, $\mathbf{W}\mathbf{y}$, to the regessors. Like λ, ρ is a spatial autocorrelation parameter that can be varied to control the size of the effect of the autoregressive $\mathbf{W}\mathbf{y}$ (Anselin et al., 2001).

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3.6}$$

The auto-regressive nature of SEMs and SLMs result in the violation of many OLS assumptions. In such cases, OLS parameter estimators are no longer unbiased and a more appropriate estimation technique, such as a maximum likelihood, is used to estimate the parameters (Anselin et al., 2001).

Various Lagrangian multiplier (LM) tests can be conducted to deduce the nature of the spatial dependence in a data set and therefore which model to use to best address it. The model choice decision process can be seen in Figure 3.1. LM error and LM lag tests assess the significance of the previously discussed λ and ρ respectively. Both may be significant, in which case, robust Lagrangian multiplier tests can be run which determine whether the two effects are significant, even in the presence of each other.

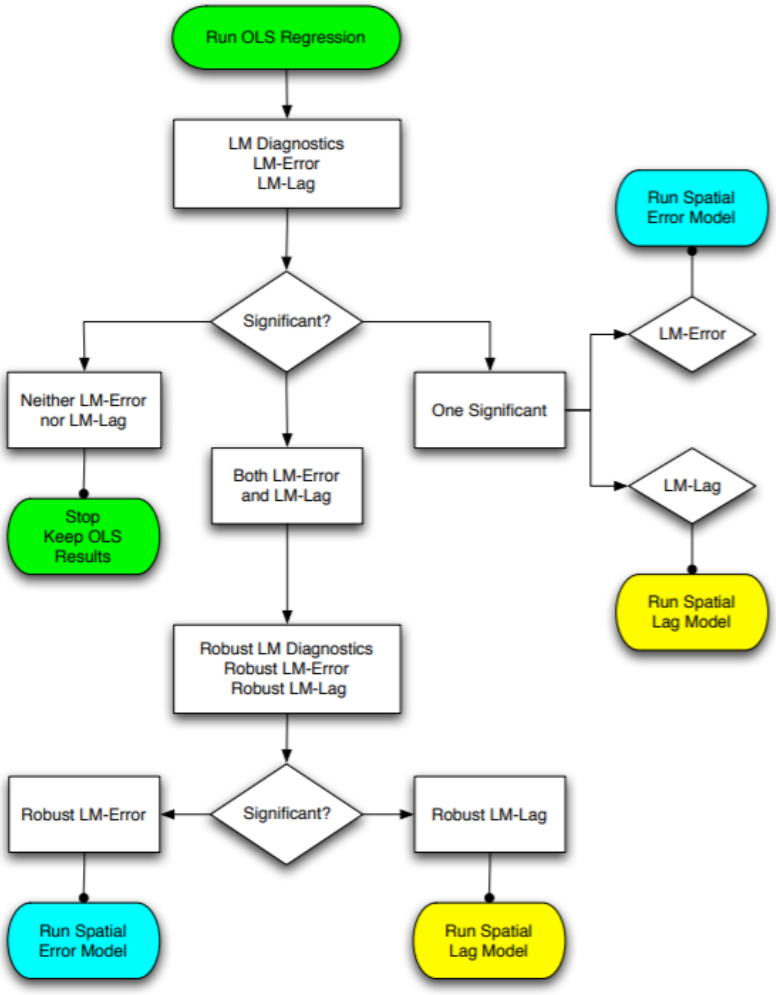


FIGURE 3.1: Regression model choice flow (Anselin, 2005)

While these regression models account for spatial dependence through the use of weight matrices and autoregressive variables, they do not necessarily correct for spatial heterogeneity.

3.2.2 Geographically weighted regression

Geographically weighted regression (GWR) is adapted from OLS to allow parameter estimates to vary over space in order to address spatial heterogeneity. Traditionally, OLS $\boldsymbol{\beta}$ parameters are

constant across all n observations, while in GWR, each observation i has its own β coefficients. Each observation, i, is formulated by Equation (3.7).

$$y_i = \beta_{i0} + \sum_{k=1}^{p} \beta_{ik} X_{ik} + \varepsilon_i. \tag{3.7}$$

In order to estimate parameters, an OLS regression is run for every observation, i, with each β_i estimated by $\hat{\beta}_i$ in Equation (3.8). $\mathbf{W}_i$ is a diagonal matrix with diagonal elements corresponding to the elements of the ith row of a spatial weights matrix (Brunsdon, Fotheringham and Charlton, 1998).

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{i}} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{W}_i\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{W}_i\mathbf{y_i}. \tag{3.8}$$

$$\mathbf{W}_i = \begin{bmatrix} w_{11} & 0 & \dots & 0 \\ 0 & w_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_{nn} \end{bmatrix} \tag{3.9}$$

GWR is susceptible to multicollinearity and so regularised LASSO or ridge regression may be used to address this. An additional limitation of GWR is that it still assumes a linear relationship between dependent and independent variables.

Spatial weights matrix

The spatial weights matrix, $\mathbf{W}$, presented in Equation (3.10) and referenced in Equations (3.5), (3.6) and (3.8), is central to the implementation of SEMs, SLMs and GWR models. $\mathbf{W}$ is a n x n matrix where elements w_{ij} are weights that are indications of relationship strength between observations i and j. Key concepts in defining a spatial weights matrix is that of a neighbourhood and a kernel function.

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \dots & w_{nn} \end{bmatrix} \tag{3.10}$$

Neighbourhood

Each observation is considered to have its own neighbourhood, where members of the neighbourhood (neighbours) are the 'nearest' or most 'similar' observations. Neighbourhoods can be defined through contiguity or distance methods. Continuity methods, which are only applicable to lattice data determine neighbours by shared borders or vertices. the distance based neighbourhood approach, used in this thesis, can be used for both point and lattice data, with the distance corresponding to the distance between points or between polygon centroids.

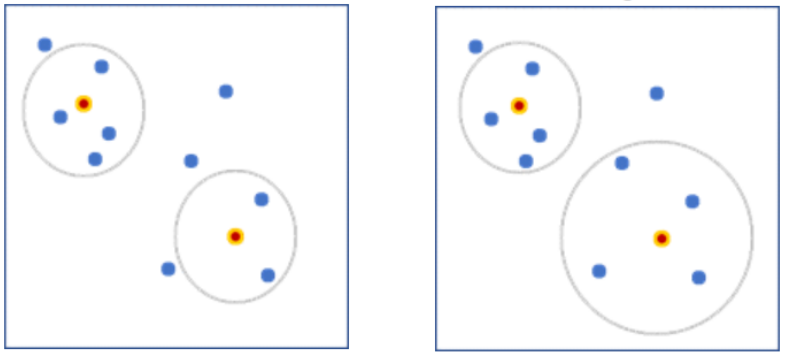
Neighbourhoods have a bandwidth, b, which is the distance from a observation to its furtherest neighbour, that is, the distance at which an observation has no effect on the central observation. Bandwidths can be fixed or adaptive, corresponding to neighbourhoods that are distance-band or number-of-neighbours respectively.

Distance-band neighbourhoods have a constant radius and if d_{ij}, the distance between observations i and j, is less than b, then observations i and j are neighbours. Various distance metrics can be used to calculate the distance between observations with coordinates (x, y). These include Euclidean distance, in Equation (3.11), or the Haversine distance, in Equation (3.12). Euclidean is straight line distance while Haversine is an 'as the crow flies' measurement that takes into account the curvature of the earth, where r in Equation (3.12) is the radius of the earth.

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}. \tag{3.11}$$

$$d_{ij} = r \times \arccos^{-1} \left[\cos |x_i - x_j| \cos y_i \cos y_j + \sin y_i \sin y_j \right]. \tag{3.12}$$

Neighbourhoods with adaptive bandwidths are determined by k nearest neighbours and so the neighbourhood radius changes based on density of observations, while the number of observations in a neighbourhood, k, remains constant.



(A) Fixed bandwidth

(B) Adaptive bandwidth

FIGURE 3.2: Distance based neighbourhoods (ESRI, 2021)

Kernel functions

The weight of each observation in a neighbourhood is determined through the use of a kernel function. The weight, w_{ij}, is a function of the distance, d_{ij}, between observations i and j. The bandwidth, b, represents the previously discussed neighbourhood bandwidth. As distance increases, the relationship, and therefore weight, decreases. The decay in weights is determined by a kernel function.

Kernel functions denote the way in which the feature space is enlarged. The most commonly used function is the Gaussian function which maps weight as a negative exponential function of distance and bandwidth as seen in Equation 3.13.

$$w_{ij} = \exp\left(-0.5 \times \left(\frac{d_{ij}}{b}\right)^2\right). \tag{3.13}$$

Each w_{ij} in Equation (3.10) is non zero when j is in i's neighbourhood and zero otherwise. $\mathbf{W}$ is not necessarily symmetric and observations are not considered a part of their own neighbourhood - reflected by the zeros in the $\mathbf{W}$ diagonal.

Transforming $\mathbf{W}$ using Equation (3.14) results in standardised matrix rows such that $\sum_j w_{ij} = 1$. This formulation is useful when, for example, multiplying with the $\mathbf{y}$ vector as $\mathbf{Wy}$ can then be interpreted as a weighted average of neighbours.

$$w^*_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}. \tag{3.14}$$

Through the use of a spatial weights matrix, spatial models are able to account for spatial dependence and spatial heterogeneity, unaccounted for by OLS. The models, however, still assume a linear a relationship between dependent and independent variables and so may be less effective in the presence of non-linearity.

3.3 Machine learning tree based models

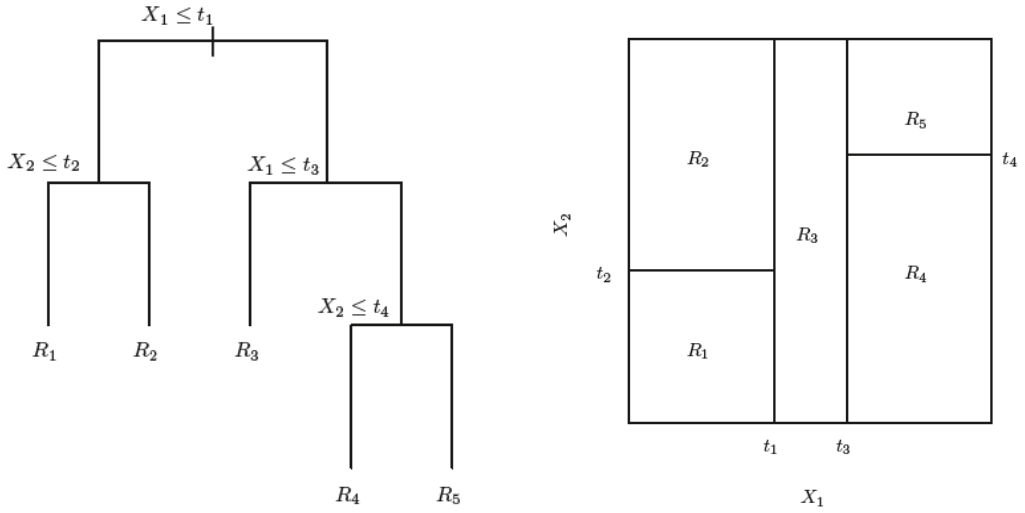
Machine learning models are able to capture and model complex relationships in data. Tree-based methods are part of the supervised learning, machine learning algorithms. They are able to model non-linear relationships between dependent and independent variables and can model both regression and classification problems where the target variables are continuous and discrete, respectively. Trees have an added advantage of being easily interpreted and visualised, while also being able to provide insight into variable importance, the respective weights that independent variables contribute to predicting the dependent variable (James et al., 2013).

Regression decision trees

Regression trees work by recursively dividing or segmenting the predictor space into regions. Each decision point or split in the tree is called a node and the resultant regions are called *terminal nodes* or *leaves*. Each split is a binary split, in that, the division of the region is into two parts. Observations in the ith region, R_i, are predicted as the mean of all the observations in the region, $\hat{y}R_i$. As a measure of how well the tree predicts the regression output, the tree's sum of square errors (the square of the difference between the actual value and the value predicted by the model) is calculated as in Equation (3.15), where J is the number of regions (Hastie, Tibshirani and Friedman, 2017).

$$\sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j}\right)^2 . \tag{3.15}$$

Visualising the recursive dividing in Figure 3.3 (A), the choice of predictor, X_j, and predictor value, t_i, on which to split is made such that a maximum reduction in the sum of square errors is achieved. In a predictor space with two independent variables X_1 and X_2 such as in Figure 3.3, the first binary split is made on a t_1 value of X_1, then on t_2 value of X_2, then t_3 value of X_3 and so on.



(A) Decision tree (B) Segmented predictor space

FIGURE 3.3: Decision tree and predictor space (James et al., 2013)

Trees can be split or *grown* to any depth until there are as many regions as observations, however in this form, they are likely to overfit the data. Overfitting occurs when the model is too complex. Representing the data, the tree is trained too closely meaning that, when it predicts unseen data, it will perform poorly as it has not been able to generalise.

Bagging

Decision trees suffer from high variance, meaning that their predictions vary greatly based on the data they are trained on. Bagging, or bootstrap aggregating, aims to address this (Hastie, Tibshirani and Friedman, 2017). Bagging aggregates trees that are grown from bootstrapped samples which are randoms samples of the data with replacement. The bagged predicted value $\hat{f}_{\text{bag}}(x)$ is calculated by averaging the predictions, $\hat{f}^b(x)$, from B trees grown from bootstrap samples (Equation 3.16).

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x). \tag{3.16}$$

Random Forests

Given that the bagged trees are grown from samples repeatedly bootstrapped from the same original sample, there is a risk of the bagged trees being correlated which would result in them making similar predictions. This results in the high variance, overfitting problems described above. Averaging correlated trees will not result in the desired reduction in variance and so

random forests (RF) aim to improve predictive performance by decorrelating trees (Hastie, Tibshirani and Friedman, 2017). This is done by only considering a random subset, m of the p explanatory variables at each split of the tree. Only considering a subset of variables at each split aids in decorrelating trees and reducing variance, as well as in identifying which variables are the most importanct based on the frequency on which they are split. Values for m can be varied but is typically chosen such that $m \approx \sqrt{p}$.

Gradient boosting machines

Gradient Boosting Machines (GBMs), like random forests, work by combining decision trees to make a prediction. While RF trees are grown independently and in parallel, GBMs are grown sequentially, where each tree depends on the one before it (Hastie, Tibshirani and Friedman, 2017).

Each tree in a GBM, 'learns' from the previous tree by modelling its prediction as a function of the previous tree's prediction error. In practice, this is done by fitting the new tree to the negative gradient of some loss function, measuring the current difference between the models predicted value compared to the actual value (Friedman, 2001). Trees are grown until some stopping criteria is met such as a certain number of trees are grown, or a certain value of the loss function is reached.

Given the ability to train on errors and iteratively improve, GBMs are frequently found to be very powerful in prediction problems. They do, however, have many parameters that need to be specified such as learning rate associated with the loss function, number of trees to be grown, and the choice of loss function to name a few. This means that while they do allow for complex modelling that often improves on RFs, they are sensitive to the choice of parameters.

A variety of spatial and non-spatial as well as linear and non-linear models have been described in this chapter. These models and their predictive performance on the Cape Town Airbnb dataset are compared and contrasted in Chapter Five, after first exploring the data in the Chapter Four.

Chapter 4

Exploratory Data Analysis

This chapter explores the Cape Town Airbnb dataset analysed in the prediction problem. A description of the data is followed by an exploration of the variables in order to decide which variables to include in the models.

4.1 Data description

Inside Airbnb is a project that aims to provide data that can be used to "understand, decide and control the role of renting residential homes to tourists" (InsideAirbnb, 2023). For cities across the world, including Cape Town, South Africa, Inside Airbnb scrapes data from the Airbnb website approximately every three months, and make the data publicly available on their website.

The COVID-19 pandemic, which started in early 2020, introduced international economic shock as well as travel restrictions which negatively affected global tourism and the Airbnb market (Kourtit et al., 2022). The Airbnb market pricing has been volatile since the start of the pandemic (Kourtit et al., 2022) and the long term effects are still being hypothesised (Dolnicar and Zare, 2020). The most recent website scrapes before the pandemic were in September and December 2019. December is in the high season for Cape Town tourism and Airbnb prices tend to be higher during this time and the number of listings vary (Ndaguba, 2021). The September 2019 data is chosen for modelling in this thesis as it is the latest pre-pandemic and pre-December data set and assumed to be the most likely to represent a stable market.

Three data files are available from Inside Airbnb; representing the listings data, review data and calendar data. The review and calendar files contain textual reviews and calendar dates from previous guests who stayed at the listing. This information could be incorporated into the analysis but extends the scope of this thesis and may be considered an exercise for further research. The listings data is the only file used in this thesis as it contains the most relevant information such as quantitative, qualitative and locational information about the Airbnb listed properties, or listings, in Cape Town. There is one entry per listing, and examples of available variables include number of bedrooms, listing description and zip code.

Text fields that are filled by hosts contain words and sentences that are subject to being misspelled and disorderly. Thorough text cleaning and processing is required to extract valuable information about the listing from them. This exceeds the scope of this thesis and so these variables have been excluded. A list of these variables as well as other initially excluded variables, that are deemed uninformative, due to having mostly nulls, little or no variation or being irrelevant to the prediction problem, are available in Table A.1 in the appendix. Their description and reason for exclusion is also provided.

In addition to the variables, listings that have a negative price per night were also removed as they are assumed to be data errors. The remaining listings data contains approximately 23 000 listings and 43 variables. Univariate and bivariate exploration of these variables are presented using descriptive statistics and graphical representations in the sections that follow.

4.2 Variable exploration and transformations

In this section, the available variables are grouped and explored by themes. The themes loosely represent those identified in the literature and are shown in Table 4.1. Variable transformations, aggregations and exclusions are detailed as part of the decision criteria for selecting the final set of variable used in modelling in Chapter Five.

4.2.1 Dependent variable

Price per night

Price is a continuous target variable for the Airbnb prediction problem. It represents the Rand value for a listing's nightly rent rate. The distribution of price is positively skewed, even post

TABLE 4.1: Grouped variables considered for modelling

Variable grouping	Variables
Price	Price
Additional costs	Security deposit, cleaning fee, guests included, extra people
Basic property variables	Property type, room type, accommodates, bathrooms, bedrooms, beds
Amenities	Amenities
Availability	Minimum days, maximum days, average minimum, average maximum, availability 30 days, availability 60 days, availability 90 days, availability 365 days, cancellation policy, instantly bookable
Host	Verified host identity, host listings count, superhost, house rules
Location	Zipcode, latitude, longitude, ward number, exact location
Reviews	Number of reviews, number of reviews last 12 months, first review, last review, reviews rating, reviews accuracy, reviews cleanliness, reviews check in, reviews communication, reviews location, reviews value, reviews per month

log transformation as seen in Figure 4.1. This skewness of price may affect the predictive performance of some of the chosen models.

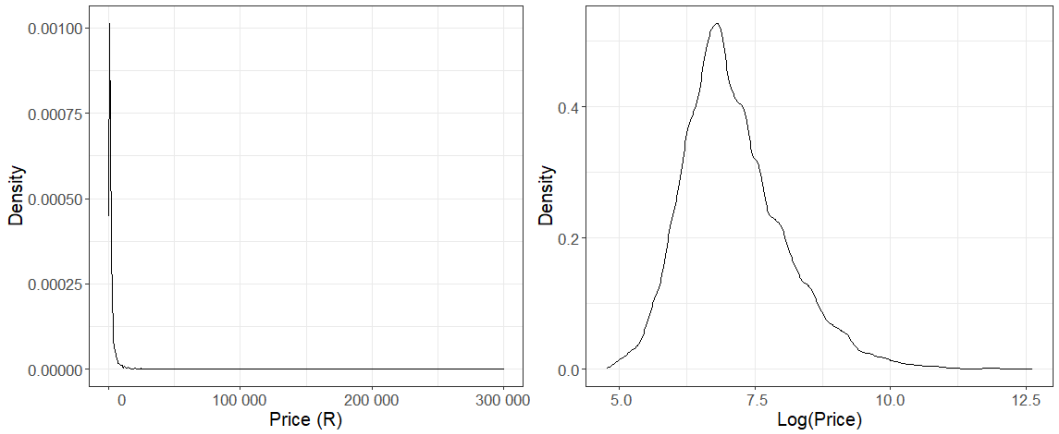


FIGURE 4.1: Distribution of Price and Log (Price)

Price has a minimum value of R 119, a maximum of R 299 998 an upper bound outlier value, using the calculation of $Upper\ quartile + (1.5 \times interquartile\ range)$, of R 4 004. Excluding listings with prices above R4 004 would result in a loss of 11.6% of listings. This is a material proportion of observations to lose and so after considering other cut off prices, a price of R 8 000 a night is chosen as it limits observations lost to only 4%, leaving approximately 22 000 listings, and maintains the composition of listing property types (Table 4.2). The price and log price distributions for listings of R 8 000 or less are given in Figure 4.2. While the price distribution appears to still be positively skewed, the tail is shorter and the log distribution is more symmetric and normal shaped. The log transform was chosen for this thesis however other scaling or normalisation methods could also have been applied.

TABLE 4.2: Top 5 proportions of property types pre and post outlier removal

	Property Type	Post removal Proportion	Pre removal Proportion
1	Apartment	39%	38%
2	House	30%	31%
3	Guest suite	6%	6%
4	Guesthouse	4%	4%
5	Serviced apartment	4%	4%

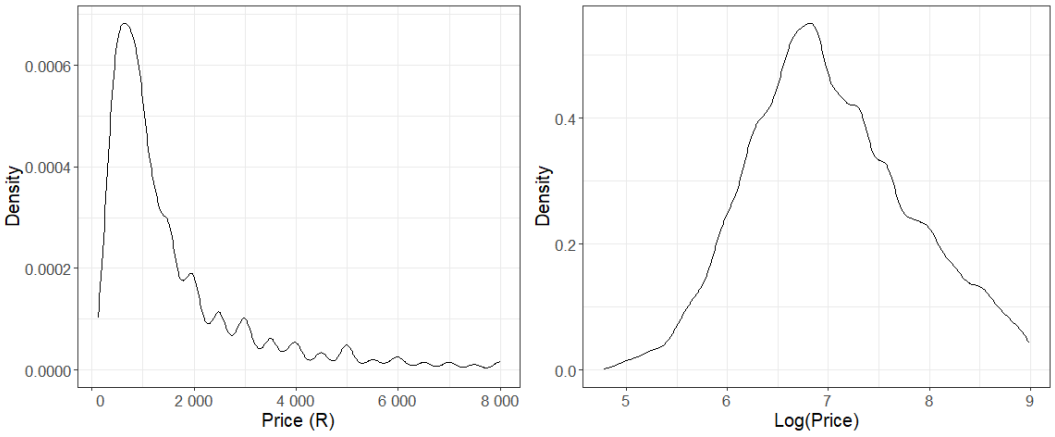


FIGURE 4.2: Distribution of Price and Log (Price) post outlier removal

4.2.2 Independent variables

Additional Cost Variables

A large proportion of listings have null values for security deposit (38% of listings) and cleaning fee (32% of listings). Further more, of the 62% of listings that have non null security deposits, only 33% have deposits greater than zero. Of the 68% of listings with non null cleaning fees, 54% have fees greater than zero. Given the larger number of listings with null or zero deposits and fees, both security deposit and cleaning fees are transformed into binary variables, where 1 indicates a deposit/fee of greater than 0.

A listing's capacity, the number of people it can accommodate, is represented by the accommodates variable. While some listings' stated nightly price include all the accommodated guests, others do not. That is, some listings charge more should a guest wish to have a number of people over a certain allowed threshold. The included guests variable indicates how many guests are included in the listing price, while the extra people variable indicates, in Rands, how much each additional guest will cost per night. The distributions of these two variables are provided in Figure 4.3.

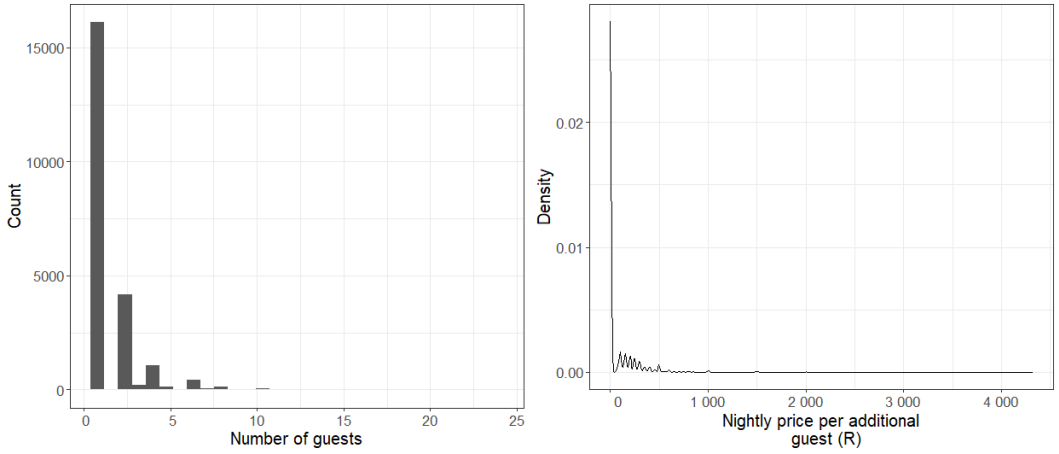


FIGURE 4.3: Distribution of number of guests included in nightly price and addition cost per night per guest

The distributions are very similar to that of price. They are highly positively skewed, with a large number of listings charging nothing for having an extra person. More value might be derived by using these two variables as binary indicators, indicating whether the listing price does not include capacity, and whether additional guests above capacity are charged. Two binary variables are therefore created to replace indicating i) when accommodates is greater than included guests and ii) when extra guest price is greater than zero. From Table 4.3, only 20% of listings do not have all guests included in the nightly price and also charge for additional guests.

TABLE 4.3: Proportion of listings that are all inclusive and that charge additionally over capacity

		Additional Charge	
		True	**False**
All inclusive	**True**	-	19%
	False	20%	61%

Basic property variables

Traditional property related variables include the type of property, the type of room, the number of people the listing accommodates, the number of bathrooms, the number of bedrooms and the number of beds.

There are over 40 different property types. While some property types are traditional such as house, there are also less common or 'unique' property types such as boats, tipis or tree houses. The variable suffers from high cardinality, with some property types having as little as one listing. An aggregation was therefore considered such as unique or not resulting in more than 99% of listings considered to be unique. Alternatively, a grouping such as houses, apartments and other

was considered, however, the classifications appear somewhat subjective and ambiguous. For example some hosts would classify a 'townhouse' as a 'house', while others would classify it as a 'townhouse'. Similarly some 'boutique hotels' are classified as 'hotels', others as 'boutique hotels'. Similarly there are 'serviced apartments', 'apartments' or 'apartment hotels'. One of the provided types, 'condominium', refers more to an ownership structure and could easily be an 'apartment', 'townhouse' or 'house'. Additionally there are other non descriptive types such as 'vacation home' or 'other'. Given the ambiguity and subjectivity in the variable, there is little confidence in the accuracy of its representation and it is left out of the modelling.

Room type is a categorical variable with four categories that represent the listings type. The room types and associated listing proportion are displayed in Table 4.4. Given that shared rooms account for less than 1% of listings, shared rooms are grouped together with private rooms as shared rooms have the most similar price distribution (Figure 4.4).

TABLE 4.4: Proportion of room types

Room Type	**Listing proportion**
Entire home/apartment	74%
Private room	22%
Hotel room	4%
Shared room	<1%

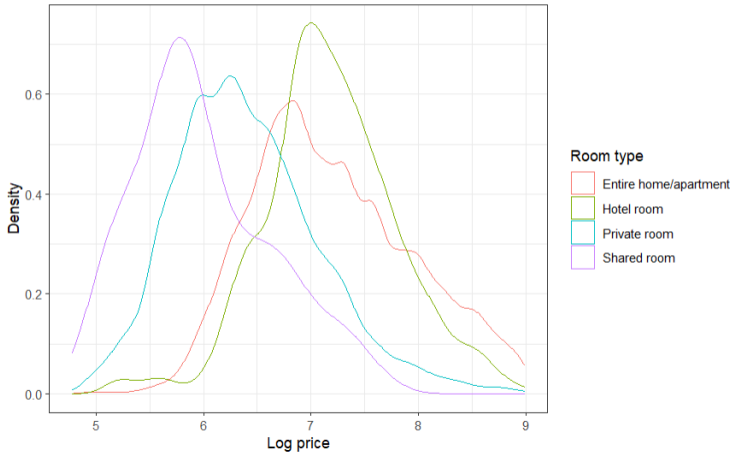


FIGURE 4.4: Distribution of log price by room type

Accommodates, bathrooms, beds and bedrooms are all discrete numeric variables that give an indication of the available space in the listing. Histograms and Spearmans correlations of these variables are given in Figure 4.5 and Table 4.5 respectively. All the variables have positively skewed distributions and are strongly positively correlated with each other.

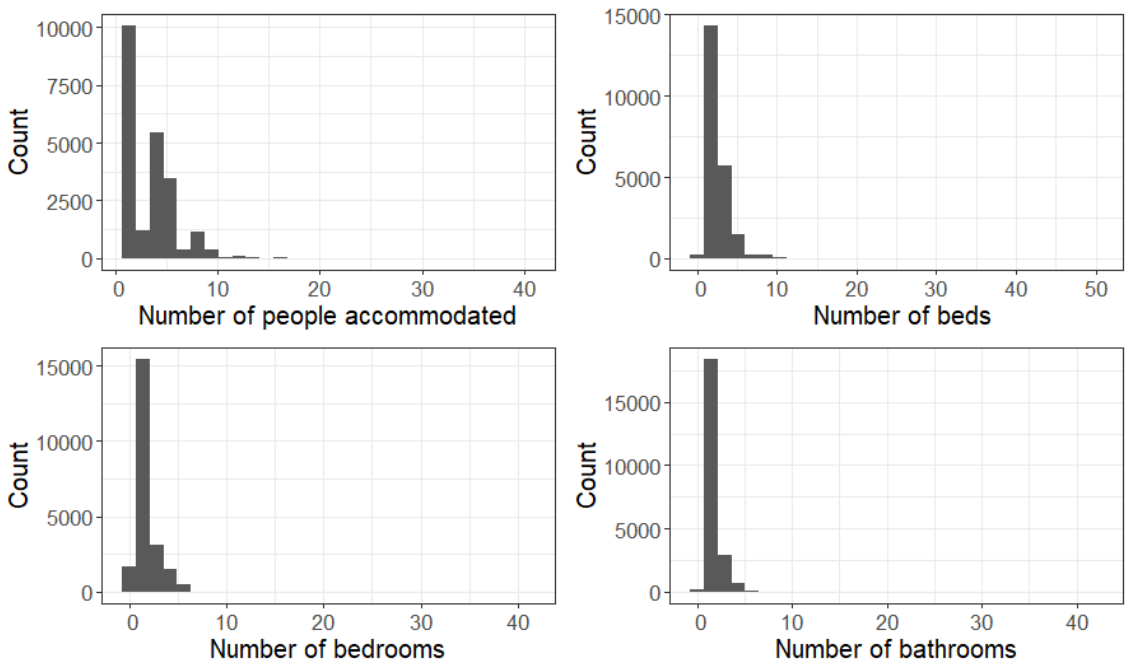


FIGURE 4.5: Histograms of people accommodated, number of beds, bedrooms and bathrooms

TABLE 4.5: Correlations of log price, accommodates, beds, bedrooms and bathrooms

	Log price	Accommodates	Beds	Bedrooms	Bathrooms
Log price	1.00	0.65	0.55	0.62	0.60
Accommodates		1.00	0.86	0.87	0.73
Beds			1.00	0.82	0.68
Bedrooms				1.00	0.77
Bathrooms					1.00

The inclusion of a set of such highly correlated variables may result in problems caused by multicollinearity. Accommodates has the highest correlation with log price (0.65) and, on average, it also has the highest correlations with the other three variables. Accommodates is therefore retained for modelling, while the rest of the variables are dropped.

Amenities

Amenities is a host-filled text variable that contains a long, comma separated string, detailing the available amenities at the listing. According to Airbnb, some of the top amenities that guests search for on the platform are a pool, wifi, kitchen, free parking, a hot tub/ jacuzzi, a kitchen, air conditioning and a washer (Airbnb, 2020). Given that Cape Town is a coastal city, beachfront/waterfront amenities are also of interest. Eight binary variables, one for each amenity, are created to indicate whether the listing has the amenity listed or not. From Figure 4.6, it can be seen that more listings have wifi, a kitchen and free parking, while fewer listings have a jacuzzi or hot tub, beach or water front as well as air conditioning.

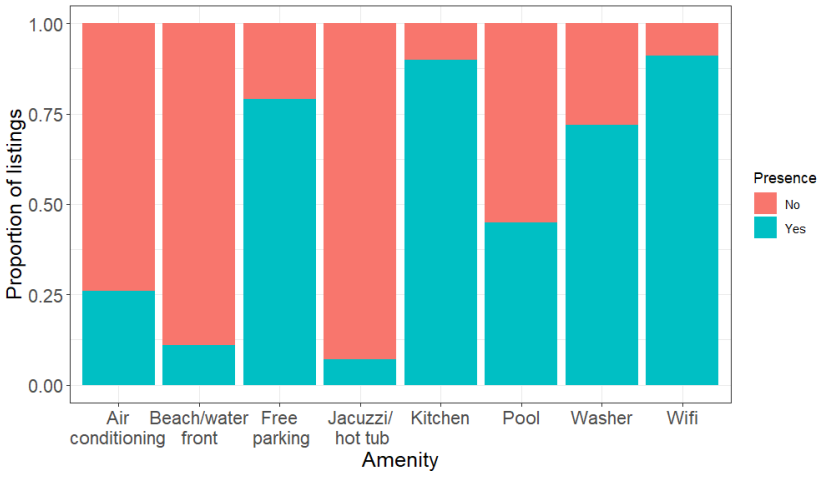


FIGURE 4.6: Proportion of listing amenities

Availability and booking duration related variables

Listing hosts are able to stipulate the minimum and maximum duration guests are allowed to stay. They are also able to change the stipulated duration throughout the year. The results are numeric variables indicating i) the current minimum number of nights the listing can be rented for ii) the current maximum number of nights the listing can be rented for iii) the average minimum number of nights the listing could be rented for in the past year and iv) the average maximum number of nights the listing could be rented for in the past year.

From the distributions in Figure 4.7, it is evident that for most listings, the minimum and average minimum nights are very small, with 82% of listings having minimum required nights between one and four. There are, however, some listings that require as many as 365 days minimum nights, potentially indicating that the listing is only available for long term rental. As with the minimum night variables, maximum nights and average maximum nights are very positively skewed. For maximum nights, 22% of listings allow a maximum stay of less than or equal to 30 nights, possibly indicating that the listing is solely available for short term rentals while 62% of listings are available for more than 365 days while the remainder have arbitrarily large numbers potentially indicating no enforced maximum.

From Table 4.6, the correlation value for minimum and maximum nights with average minimum and average maximum nights is 0.99. Minimum nights has a very low correlation of 0.05 with price and maximum nights effectively has a correlation of zero. Both variables are indications of availability for short and long term rental. Minimum nights of greater than 30 days are indications of only long term availability while maximum nights of less than 30 are indicators

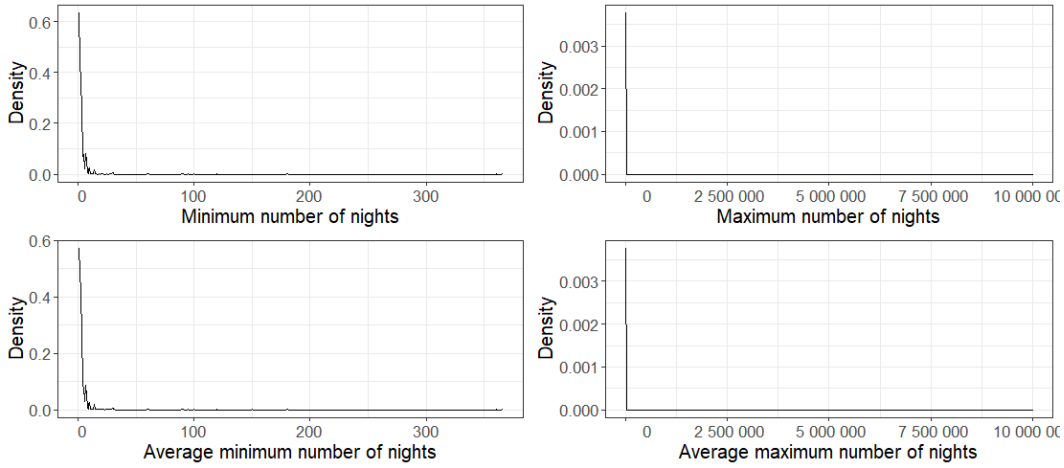


FIGURE 4.7: Distributions of minimum, maximum, average minimum and average maximum number of nights

of only short term availability. These listings do not overlap and so a binary variable is created to capture this, indicating whether the listing is only available for short term rental and the average variables are dropped.

TABLE 4.6: Correlations of log price, minimum, maximum, average minimum and average maximum nights

	Log price	**Minimum nights**	**Maximum nights**	**Average minimum nights**	**Average maximum nights**
Log price	1.00	0.04	0.00	0.04	0.00
Minimum nights		1.00	0.00	0.99	0.00
Maximum nights			1.00	0.00	0.99
Average minimum nights				1.00	0.00
Average maximum nights					1.00

Closely related to minimum and maximum nights are the availability variables. Four availability variables indicate the number of days the listing is available for in the next 30, 60, 90 and 365 days. Listings are unavailable if they are booked by a guest or if the host block out the listing in the booking/listing calendar. Distributions for the four variables are shown in Figure 4.8 and their corresponding correlations in Table 4.7.

Across the board, the distributions are largely resemble U-shaped distributions, peaking on the low and high end of the spectrum indicating that for the given periods, listings tend to be largely available for many of the days (high end of the spectrum) or tend to be largely unavailable for many of the days (low end of the spectrum). If hosts are unwilling or unable to rent out their listing over specific dates, they are able to 'block out' their calendar on those days such that no one can rent their listing over that period. Low availability could be an indication of a listing's popularity or of the host having blocked out the calendar. Regarding 365 day availability, there

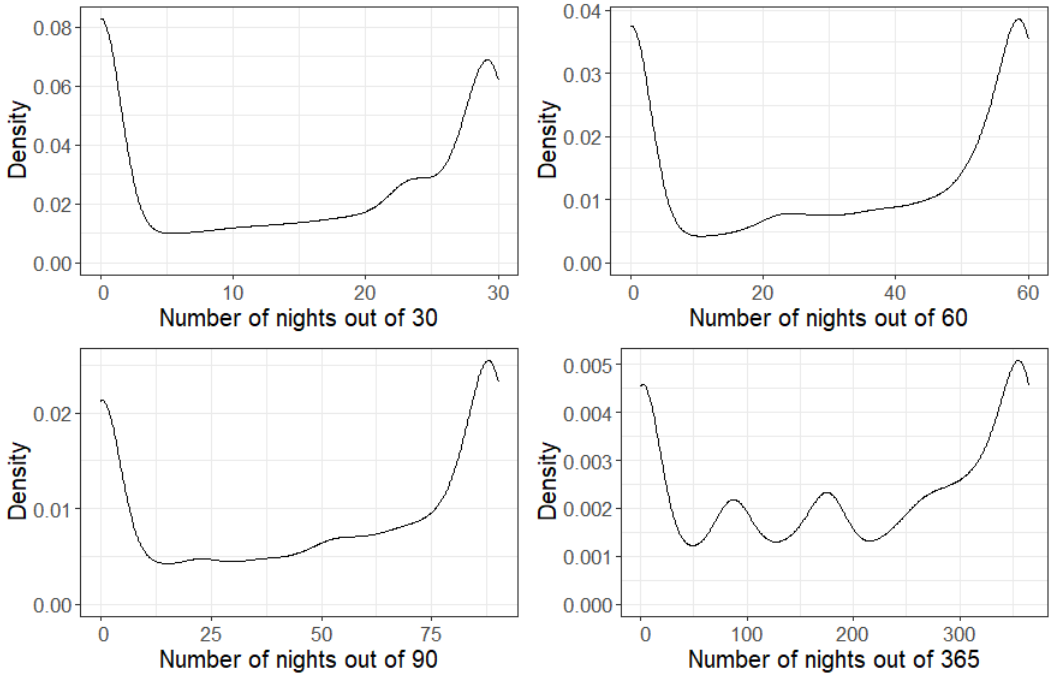


FIGURE 4.8: Distribution of 30, 60, 90 and 365 day availability

are some peaks in between 50 to 300 days potentially indicating listings that are only available in high seasons or available a certain number of months a year.

TABLE 4.7: Correlations of log price, 30, 60, 90 and 365 availability

	Log price	**30 day availability**	**60 day availability**	**90 day availabilty**	**365 day availability**
Log price	1.00	-0.04	-0.06	-0.05	-0.01
30 day availability		1.00	0.96	0.92	0.65
60 day availability			1.00	0.97	0.68
90 day availability				1.00	0.70
365 day availability					1.00

The 30, 60 and 90 day variables are highly positively correlated with each other while the 365 day variable is also positvely correlated with them, but slightly less so. Given this, only the 30 and 365 day variables are retained. These are both proportion variables but on different time scales and are therefore converted to percent variables indicating the percentage of available days, in the two time periods.

The cancellation policy variable is a factor variable indicating the degree of strictness around booking cancellation. Table 4.8 shows the proportion of listings with each type. With three of the cancellation types having less than 1% of observations, and another with only 2%, the variable is converted into a binary of flexible and moderate vs other (strict).

'Instantly bookable' is a binary variable denoting whether the guest can automatically book the listing, without the host requiring to accept their booking request. Just more than half,

TABLE 4.8: Cancellation Policy proportions

Policy	Proportion
Flexible	33%
Moderate	23%
Luxury moderate	<1%
Strict 14 with grace period	41%
Strict	<1%
Super strict 30	2%
Super strict 60	<1%

53%, of listings can be booked instantly and this variable is left unchanged and included in the modelling.

Host variables

There are numerous variables that relate to the listing's host. There is a binary variable indicating whether the host's identity has been verified, a numeric variable of how many properties the host has listed on the Cape Town Airbnb market, a binary indication of whether they are a 'superhost' as well as field a text variable containing the host's 'house rules'.

Superhost is an Airbnb concept. A host is deemed a superhost if they; i) have an average overall rating of 4.8/5 or more ii) have guests stay ten or more times per year or 100 nights over three stays iii) have a cancellation rate of less than 1% and iv) have a 90% response rate within 24 hours. Airbnb updates superhost status every three months.

Looking at the proportion of hosts that are superhosts in conjuction with whether the host has a verified identity in Table 4.9, most hosts are not superhosts and most do not have verified identities, however the ratio of verified vs not verified is much higher for superhosts.

TABLE 4.9: Superhost and verified identity proportions

		Superhost		
		True	False	Total
Identity verified	True	8%	19%	27%
	False	13%	60%	73%
	Total	21%	79%	

The distribution of how many Cape Town Airbnb listings hosts have listed is shown in Figure 4.9. While many of the hosts only have one listing, there is a long tail where the maximum number of listings is 373, potentially indicating that the host is a rental agency or a person running a business.

The last host variable, house rules, is a text variable in which hosts can detail their preferences for guest behaviour and impose restrictions on guest activity or access. The variable is empty

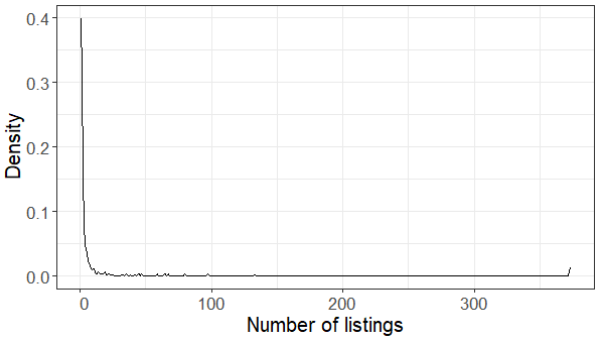


FIGURE 4.9: Distribution of host local listings count

for 50% of the listings, potentially indicating no rules. For the remaining 50%, as with the other initially excluded text variables, substantial work is needed to extract useful and valuable information. The variable is therefore converted to a binary variable indicating whether there are house rules or not.

Reviews variables

Review variables can loosely be grouped into ratings or review scores and number of reviews.

Ratings

Approximately 35% of listings have null values for accuracy, cleanliness, check in, communication, location and value, potentially indicating that these listings have never been reviewed. Review score rating is an overall score that loosely incorporates the other score variables. The correlations of these variables with each other and log price are presented in Table 4.10. In addition, the distribution of all the non null review score ratings is shown in Figure 4.10.

TABLE 4.10: Review score correlations

	Log price	**Review score rating**	**Review score accuracy**	**Review score cleanliness**	**Review score check in**	**Review score communication**	**Review score location**	**Review score value**
Log price	1.00	0.08	0.04	0.04	0.03	0.02	0.09	-0.04
Review score rating		1.00	0.72	0.67	0.59	0.62	0.50	0.72
Review score accuracy			1.00	0.62	0.58	0.62	0.51	0.68
Review score cleanliness				1.00	0.50	0.50	0.42	0.61
Review score check in					1.00	0.70	0.45	0.55
Review score communication						1.00	0.48	0.57
Review score location							1.00	0.52
Review score value								1.00

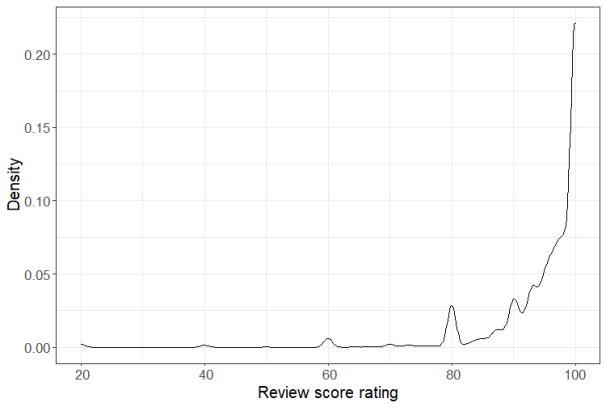


FIGURE 4.10: Distribution of reviews scores

Number of reviews

The same listings that have null values for ratings have null values for first review and last review and zero values for number of reviews, number of reviews in the last 12 months as well as reviews per month. Table 4.11 shows more reviews are negatively correlated with log price and that number of reviews, number of reviews in the last 12 months and reviews per month are highly correlated with each other.

TABLE 4.11: Review count correlations

	Log price	Number of reviews	Number of reviews last 12 months	Reviews per month
Log price	1.00	-0.08	-0.09	-0.14
Number of reviews		1.00	0.80	0.68
Number of reviews last 12 months			1.00	0.85
Reviews per month				1.00

Given that 35% of listings are missing review scores, the choice is made to only retain review score rating as it is largely correlated with the other score types. Given that it has such a negative skew (Figure 4.10), the variable is aggregated into the buckets shown in Table 4.12. Similarly, given the high correlation between the review numbers variables, only number of reviews is retained as it represents a longer historical view not captured by the averaged reviews per month, and contains more data than just the last 12 months.

TABLE 4.12: Review score listing proportion

	Proportion of listings
Rating scores: 0-93	18%
Rating scores: 93-99	24%
Rating scores: 100	24%
Rating scores: null	34%

Locational variables

Provided variables

Five locational variables are provided; zipcode, latitude, longitude, ward number and an indication of whether the provided coordinates represent the listing's exact location. The position of listings, according to their latitude and longitude coordinates is given in Figure 4.11 which when viewed simultaneously with Figure 4.12, shows that the highest densities of listings are found in the City Bowl, Atlantic Seaboard and Southern Suburbs while the lowest density and fewest number of listings is in the South East Suburbs. Each location is in a ward, a low level municipal boundary, and each ward is in a zipcode also known as a postal code area in South Africa.

Ward number is very granular and is intuitively correlated with latitude and longitude and zipcode. For modelling, ward number is aggregated into more commonly known 'suburbs' that are shown in Figure 4.12.

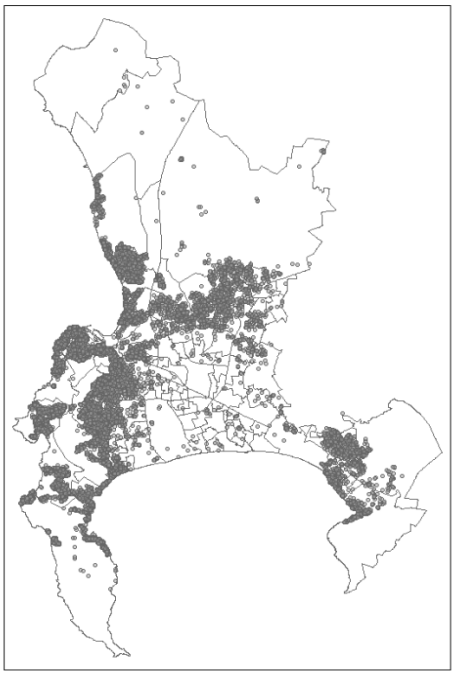


FIGURE 4.11: Position of listings

Exact location is a binary variable and 30% of listings do not provide exact location. This may be due to security concerns. For these listings, the provided coordinates will be used regardless.

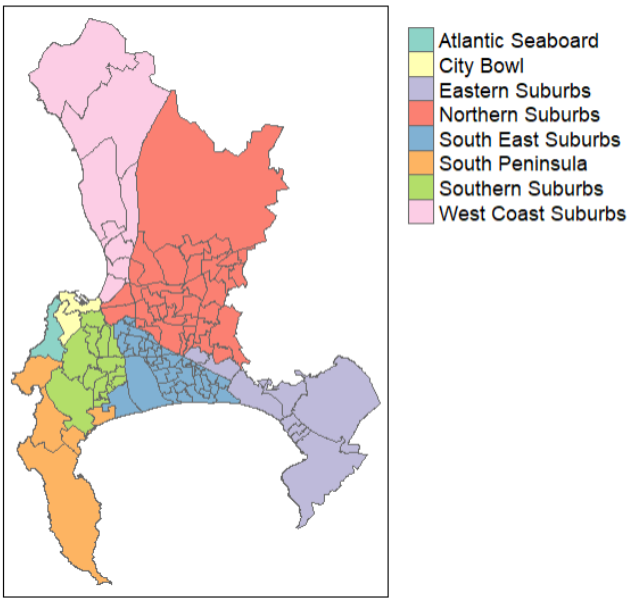


FIGURE 4.12: Outline of suburbs

Generated variables : Distance to nearest tourist attraction and distance to airport

Proximity to tourist attractions and transit facilities are often found to be important in the literature. These variables are consequently created for the analysis.

Tripadvisor, an internationally popular travel and tourism website, lists the top Cape Town attractions (Tripadvisor, 2021). Only 19 of the top 20 are located within the City of Cape Town's boundaries. These attractions' positions are displayed in Figure 4.13. For each listing, the distance to each attraction is measured using the Haversine distance, discussed in Chapter Three, which takes into account the curvature of the earth. The minimum of these is taken as the distance to the nearest attraction. The distance is measured in meters and the distribution is shown in Figure 4.14. The distribution is positively skewed with most listings being less than 10 000 meters / 10 km from the nearest attraction. To represent proximity to transit, the distance to the airport for each listings is calculated in the same manner.

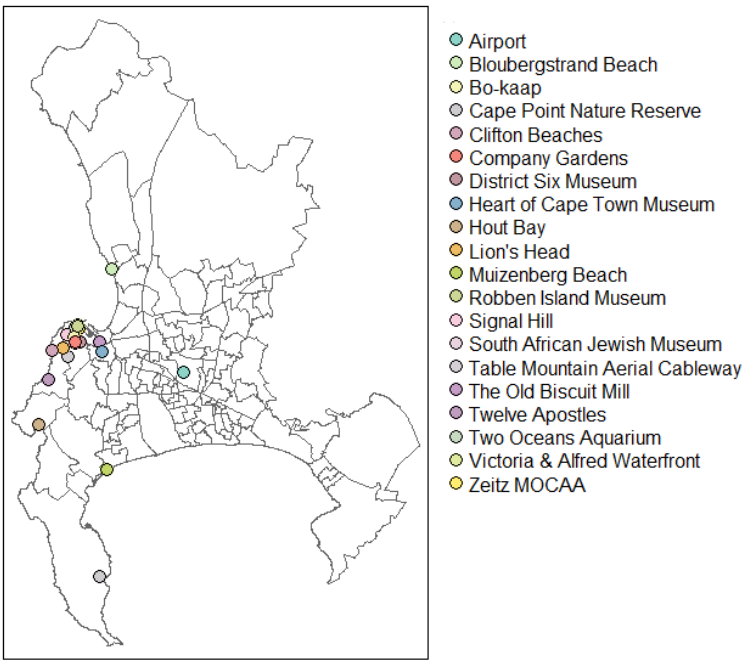


FIGURE 4.13: Position of Cape Town airport and tourist attractions

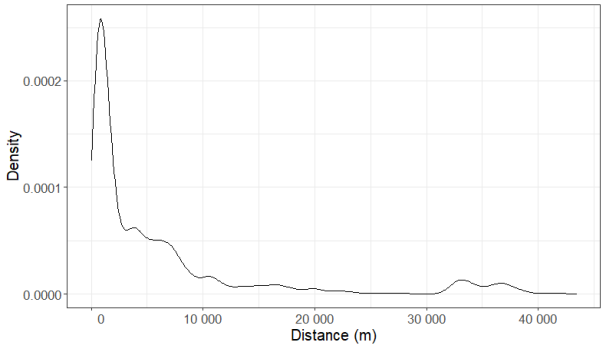


FIGURE 4.14: Distribution of distance to nearest attraction

Final variables

Having detailed the choices for inclusion, transformation or exclusion, the final variables used in modelling are shown in Table 4.13.

TABLE 4.13: Grouped final variables used for modelling

Variable grouping	Variables
Price	Log price
Additional costs	Security deposit, cleaning fee, all inclusive, additional charge
Basic property variables	Room type, accommodates
Amenities	Wifi, free parking, pool, hot tub, kitchen, aircon, washer, beach / water front
Availability	Short term only, availability 30 days, availability 365 days, cancellation policy, instantly bookable
Host	Verified host identity, host listings count, superhost, house rules
Location	latitude, longitude, suburb, exact location, distance to airport, distance to nearest attraction
Reviews	Number of reviews, reviews rating

In the next chapter, the transformed data set is modelled using various methods in an attempt to best predict the Cape Town Airbnb prices as well as study the price determinants and the spatial variability of the market.

Chapter 5

Results

The model training and testing process as well as the model results are described in this chapter. The data set is split into 80% training and 20% testing. In the training process, cross validation is performed on the 80% of data in order to decide on final model configuration. This model selection is decided based on the performance metrics which are described in this chapter. The same metrics are used to judge model performance on the unseen 20% test set.

5.1 Metrics

Root mean square error and root mean square log error are the two main metrics on which model performance is judged in this thesis. Adjusted R^2 and Akaike information criterion are also be considered briefly.

Root mean square error

Root mean square error (RMSE) is commonly used to evaluate continuous prediction problems. Models with lower values RMSE values calculated using Equation 5.1 are considered to perform better than those with higher values. In Equation 5.1, $\hat{y}_i$ represents the value predicted by the model and y_i the actual, observed value corresponding to that prediction.

$$RMSE = \sqrt{\frac{\sum_i (\hat{y}_i - y_i)^2}{n}} \tag{5.1}$$

Root mean square log error

Root mean square log error (RMSLE) is a less common metric introduced in recent years and its calculation is shown in Equation 5.2 where, as in Equation 5.1, $\hat{y}_i$ represents a the value predicted by the model and y_i the actual, observed value corresponding to that prediction. Models with lower values RMSLE values are also considered to perform better than those with higher values.

When compared to RMSE, RMSLE:

1. Is Less affected by outliers. The $(\hat{y}_i - y_i)^2$ error calculation in RMSE means that the size of y affects the calculation and so larger values of y and potential outliers can explode the overall RMSE. Whereas the $(log(\hat{y}_i + 1) - log(y_i + 1))^2$ error calculation in RMSLE is a relative measurement since $log(\hat{y}_i + 1) - log(y_i + 1) = \frac{log(\hat{y}_i+1)}{log(y_i+1)}$. RMSLE is therefore less affected by outliers and larger values of y.
2. Penalises underestimation more than over estimation. By looking at Figure 5.1 it is evident that underestimation and overestimate of the same absolute error size are treated the same in RMSE. This is not the case for RMSLE where, because of log properties, underestimations ie. values of $y - \hat{y}$ greater than 0 have higher RMSLE values than their overestimation equivalent.

$$RMSLE = \sqrt{\frac{\sum_i (log(\hat{y}_i + 1) - log(y_i + 1))^2}{n}} \tag{5.2}$$

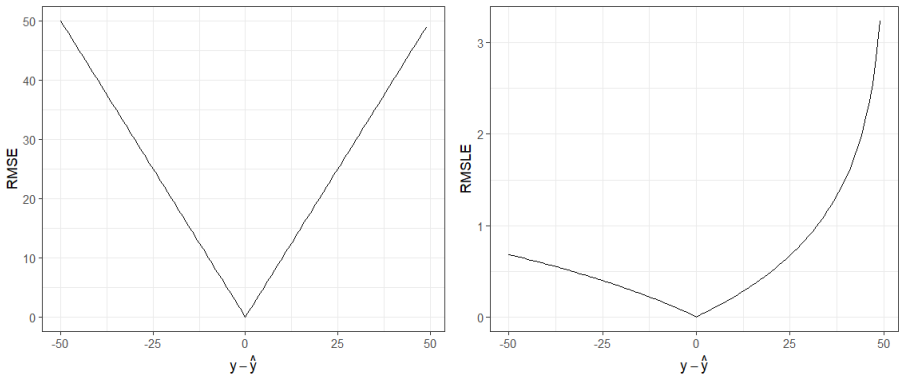


FIGURE 5.1: Illustrative example of RMSE vs RMSLE

In this thesis, the exponential function is pre-applied to y_i values in metric calculations such that price is represented instead of log price.

Adjusted R^2

Adjusted R^2 is a measure of model fit that indicates the degree to which the variation in the model's dependent variable is explained by the model's independent variables. Higher values adjusted R^2 indicate more explained variability. It is calculated as in Equation 5.3 and adjusts for the number of variables used in the model, k.

$$R^2_{adj} = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - k - 1}\right) \tag{5.3}$$

Akaike information criterion

Akaike information criterion (AIC) is often used in model selection. Calculated as in Equation 5.4, k is the number of parameters used in the model and $\hat{L}$ is the maximum value of the likelihood function for the model. Lower AIC values are an indication of better model quality, balancing both model complexity (k) and model fit ($\hat{L}$).

$$AIC \ = \ 2k - 2\ln(\hat{L}) \tag{5.4}$$

5.2 Training and validation

In this section, the training process, training decisions and training results of all the models are described.

OLS

Five different model specifications of the OLS regression model are considered and presented in Table 5.1. For each model, the training data is split into five equally sized folds and the model is trained on four out of the five folds and its predictive performance on the left-out fold (the fifth fold) is recorded. This process is repeated another four times, each time leaving out a different

fold of data. This process is called five-fold cross-validation. The values of the performance metrics of the five folds is averaged, creating a validation measure of RMSE and RMSLE. The model is rerun using the whole training set in order to obtain model adjusted R^2. The results are provided in Table 5.2.

TABLE 5.1: Linear model specifications

Model	Description	Variables
Model 1	Full model	All variables
Model 2	Step model	All variables from Model 1 except Superhost and Latitude
Model 3	Coefficient and t-statistic filtering	All variables from Model 2 except House rules, Availability 365 days, Distance to airport and Distance to nearest attraction
Model 4	Coefficient and t-statistic filtering	All variables from Model 3 except Verified host identity, Exact location and Hot tub
Model 5	Coefficient and t-statistic filtering	All variables from Model 4 except Short term availability and Beach / water front

Model 1 uses all variables followed by a Model 2 resulting from a step-wise regression, a process that sequentially adds and removes variables, each time assessing the effect of their addition/removal. Thereafter, variables are progressively removed based on the size of their t-statistics and coefficients in Models 3 to 5 with the criteria for removal being small t-statistics, large p-values and small coefficients. Variable coefficients and significance does not change drastically as variables are removed and the models perform similarly with metrics not varying greatly. That being said, performance decreases slightly from Model 3 to 5, evident by decreasing adjusted R^2 and increasing validation errors. Models 1 and 2 are very similar, however Model 2 performs marginally better with a lower RMSE. Model 2 is therefore chosen as the final linear regression specification and so the only omitted variables are superhost and latitude. The interpretations that follow focus on Model 2 results.

The base levels of all the categorical variables are provided in Table 5.3.

Basic variables

The accommodates t-statistic of ∼87 is the largest by far. This is unsurprising given that accommodates was highly correlated with log price and that it approximates the size of listing / number of bedrooms on so on. The positive coefficient and therefore positive relationship with log price is highly intuitive and in line with findings in the literature.

Private/ shared room type has the second highest absolute coefficient showing the second largest effect on log price. Both Private / share room and Hotel room have intuitive negative effects showing that they are likely to be cheaper than renting out an entire house or apartment (the

TABLE 5.2: OLS model summaries

		Model 1		Model 2		Model 3		Model 4		Model 5	
		Coeff	t-stat	Coeff	t-stat	Coeff	t-stat	Coeff	t-stat	Coeff	t-stat
Category	Variable	(Std err)	(pp-value)	(Std err)	(p-value)	(Std err)	(p-value)	(Std err)	(p-value)	(Std err)	(p-value)
	(Intercept)	38.10	8.88	40.57	15.53	40.55	18.13	40.51	18.08	40.59	18.10
		(4.29)	(0.00)	(2.61)	(0.00)	(2.24)	(0.00)	(2.24)	(0.00)	(2.24)	(0.00)
Basic	Accommodates	0.17	86.69	0.17	86.75	0.18	87.76	0.18	87.73	0.18	87.82
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	Room type: Hotel room	−0.06	−2.41	−0.06	−2.42	−0.05	−2.36	−0.05	−2.27	−0.05	−2.35
		(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	Room type: Private/ shared room	−0.42	−39.06	−0.42	−39.09	−0.42	−39.28	−0.42	−39.28	−0.42	−39.36
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
Amenities	Wifi	0.10	7.13	0.10	7.16	0.10	7.35	0.10	7.18	0.10	7.14
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Free parking	0.09	8.96	0.09	8.98	0.09	9.32	0.09	9.35	0.09	9.34
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Pool	0.18	22.84	0.18	22.85	0.18	22.70	0.18	22.79	0.18	22.78
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Hot tub	0.06	3.85	0.06	3.85	0.05	3.71				
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)				
	Kitchen	−0.12	−8.49	−0.12	−8.50	−0.13	−8.88	−0.12	−8.81	−0.12	−8.69
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Aircon	0.23	25.48	0.23	25.53	0.24	25.76	0.24	25.99	0.24	25.95
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Washer	0.11	11.77	0.11	11.78	0.11	11.23	0.11	11.85	0.11	12.20
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Beach / water front	0.07	5.61	0.07	5.69	0.07	5.93	0.07	5.99		
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)		
Host	Superhost	0.01	0.84								
		(0.01)	(0.40)								
	Verified host identity	0.04	4.60	0.04	4.62	0.04	4.42				
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)				
	Host listings count	0.00	6.96	0.00	6.93	0.00	7.04	0.00	6.44	0.00	5.93
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	House rules	−0.02	−2.09	−0.02	−2.08						
		(0.01)	(0.04)	(0.01)	(0.04)						
Location	Latitude	−0.07	−0.73								
		(0.10)	(0.46)								
	Longitude	−1.86	−13.16	−1.86	−13.15	−1.85	−15.22	−1.85	−15.19	−1.85	−15.20
		(0.14)	(0.00)	(0.14)	(0.00)	(0.12)	(0.00)	(0.12)	(0.00)	(0.12)	(0.00)
	Exact Location	−0.04	−4.98	−0.04	−4.93	−0.04	−5.09				
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)				
	Suburb: City Bowl	−0.16	−10.93	−0.16	−11.02	−0.18	−12.76	−0.18	−12.78	−0.19	−13.50
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Suburb: Eastern	0.15	2.27	0.15	2.35	0.32	5.51	0.32	5.46	0.32	5.47
		(0.06)	(0.02)	(0.06)	(0.02)	(0.06)	(0.00)	(0.06)	(0.00)	(0.06)	(0.00)
	Suburb: Northern	−0.18	−4.91	−0.19	−5.43	−0.20	−5.95	−0.20	−6.04	−0.21	−6.22
		(0.04)	(0.00)	(0.03)	(0.00)	(0.03)	(0.00)	(0.03)	(0.00)	(0.03)	(0.00)
	Suburb: South East	−0.27	−5.27	−0.26	−5.25	−0.31	−6.37	−0.31	−6.41	−0.31	−6.48
		(0.05)	(0.00)	(0.05)	(0.00)	(0.05)	(0.00)	(0.05)	(0.00)	(0.05)	(0.00)
	Suburb: South Peninsula	−0.33	−13.81	−0.31	−19.56	−0.28	−18.23	−0.28	−18.17	−0.28	−18.11
		(0.02)	(0.00)	(0.02)	(0.00)	(0.02)	(0.00)	(0.02)	(0.00)	(0.02)	(0.00)
	Suburb: Southern	−0.23	−11.66	−0.23	−11.70	−0.26	−14.80	−0.26	−15.02	−0.27	−15.64
		(0.02)	(0.00)	(0.02)	(0.00)	(0.02)	(0.00)	(0.02)	(0.00)	(0.02)	(0.00)
	Suburb: West Coast	−0.28	−12.21	−0.29	−14.73	−0.29	−14.86	−0.29	−14.72	−0.28	−14.43
		(0.02)	(0.00)	(0.02)	(0.00)	(0.02)	(0.00)	(0.02)	(0.00)	(0.02)	(0.00)
	Distance to airport	0.00	4.60	0.00	4.57						
		(0.00)	(0.00)	(0.00)	(0.00)						
	Distance to nearest attraction	0.00	2.24	0.00	2.42						
		(0.00)	(0.03)	(0.00)	(0.02)						
Reviews	Number of reviews	−0.00	−7.74	−0.00	−7.72	−0.00	−7.59	−0.00	−7.34	−0.00	−7.33
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	Reviews: 100	0.14	12.13	0.14	12.38	0.15	12.71	0.15	12.44	0.15	12.53
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Reviews: 93-99	0.02	1.60	0.02	1.95	0.03	2.14	0.02	2.01	0.03	2.06
		(0.01)	(0.11)	(0.01)	(0.05)	(0.01)	(0.03)	(0.01)	(0.04)	(0.01)	(0.04)
	Reviews: null rating	0.25	21.75	0.25	21.80	0.25	22.20	0.25	22.11	0.25	22.07
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
Costs	Cleaning fee	−0.06	−7.21	−0.06	−7.20	−0.07	−7.59	−0.07	−7.60	−0.07	−7.45
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Security deposit	0.08	8.26	0.08	8.27	0.07	8.14	0.08	8.36	0.08	8.56
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	All inclusive	0.08	7.51	0.08	7.51	0.08	7.57	0.08	7.78	0.08	7.79
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Additional charge	−0.21	−23.11	−0.21	−23.15	−0.21	−23.56	−0.21	−23.56	−0.21	−23.57
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
Availability	Short term only	0.04	4.61	0.04	4.56	0.04	4.09	0.03	3.79		
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)		
	Availability 30 days	0.08	7.02	0.08	7.02	0.12	12.22	0.11	12.05	0.11	11.94
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Availability 365 days	0.05	3.88	0.05	3.93						
		(0.01)	(0.00)	(0.01)	(0.00)						
	Instantly bookable	−0.05	−6.65	−0.05	−6.64	−0.05	−6.50	−0.05	−7.08	−0.05	−6.89
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Cancellation policy	−0.13	−16.17	−0.13	−16.20	−0.13	−16.46	−0.13	−16.82	−0.13	−16.63
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
Adj. R^2		0.618		0.618		0.617		0.615		0.614	
Validation RMSE		1020		1019		1022		1024		1027	
Validation RMSLE		0.493		0.493		0.494		0.495		0.495	

TABLE 5.3: Categorical variable base variables

Categorical variable	**Base level**	**Other levels**
Room type	Entire home or apartment	Hotel room, Private or shared room
Suburb	Atlantic Seaboard	City bowl, Eastern, Northern, South East, South Peninsula, Southern, West Coast
Reviews	0 - 93	93 - 100, 100, null rating

base room type level). This effect, however, is more significant for the Private / shared room evident by its larger coefficient and smaller p-value.

Amenities

Having a kitchen is the only amenity that reduces listing price in that it is the only amenity with a negative coefficient. This is counter intuitive as one would expect any amenity to add value. Air conditioning has the largest coefficient as well as t-statistic, followed by having a pool. This may be an indication of the preference for the ability to take part in 'summer activities' in the Airbnb market in Cape Town, such as swimming in a pool.

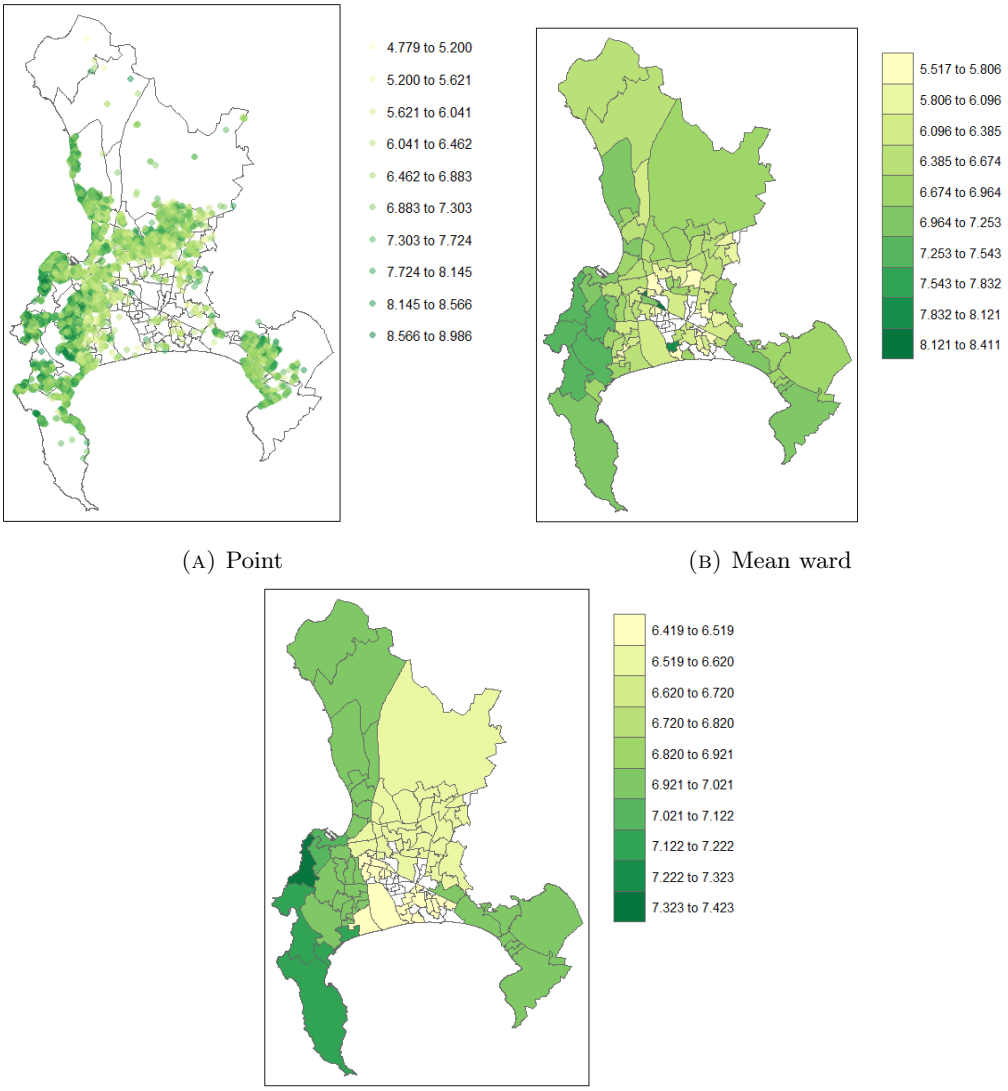
Host

The host variables on the whole have smaller effects compared to the other variables. The host being a superhost, verified as well as having more listings all marginally add to the price of the listing. This makes sense as they allude to quality and experience. The host having house rules is the only host variable with a negative coefficient, thereby reducing log price.

Location

Longitude has the largest absolute coefficient of all the variables. It has a negative coefficient and therefore negative effect showing that listings with a higher longitude (that is, listings that are further East) are more likely to be cheaper than those in the West. This makes sense as the upmarket areas on the Atlantic Seaboard and in the City Bowl which are referenced in the literature are the most far West, as seen in Figure 4.12. Looking at mapped prices in Figure 5.2 these areas tend to have higher listing prices. Similarly, within the Southern Suburbs the more expensive listings are further West and on the West Coast the listings tend to be more expensive compared to the Northern Suburbs listings East of them.

All mapped values from Figures 5.2 onwards are presented in even buckets of 10 with values of light yellow being the midpoint of zero, orange and red being negative and green being positive.



(A) Point

(B) Mean ward

(C) Mean suburb

FIGURE 5.2: Mapped log price

All suburbs except the Eastern Suburbs have a negative coefficient showing that, on average, listings on the Atlantic Seaboard are more expensive than them given that Atlantic Seaboard is the base level in the suburbs categorical variable. The average listing price in the Eastern Suburbs is lower than that of the Atlantic Seaboard but higher than the South Eastern and the Southern Suburbs, which are directly West of it. This shows that East-West price effect only holds true up until the Eastern Suburbs and so the Eastern Suburbs have a positive coefficient to offset some of the overall negative longitude effect. South Peninsula, West Coast and South East Surburbs have the largest negative effect showing the biggest average price difference compared to the Atlantic seaboard. With the exception of the Eastern suburbs, these three regions are the most far South, North and East from the Atlantic seaboard respectively showing how price

decreases in all directions from the Atlantic Seaboard, evident in Figure 5.3.

Distance to the airport and to the closest attraction have very small, near zero effects on price, however even though the coefficient and therefore effect on price is small, exclusion results in reduced performance as seen in Model 3 to 5. Where the exact location is provided, listings are, on average, cheaper than those that where not, potentially showing that Airbnbs that prioritise security by not revealing exact location are on average more expensive. Overall location variables have large absolute t-statistics and small errors showing the value of the spatial information in predicting price.

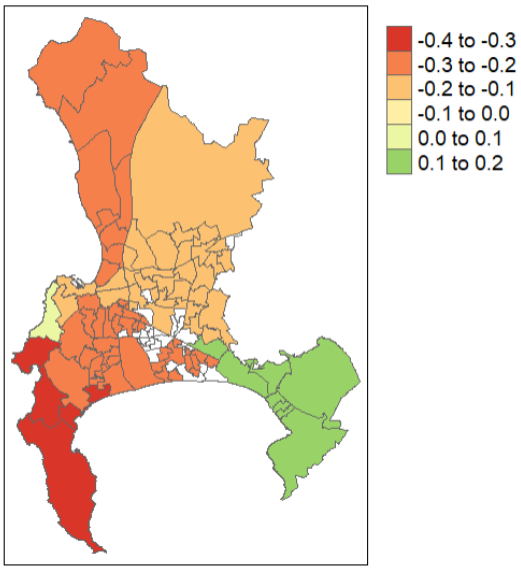


FIGURE 5.3: Model 2 suburb coefficients

Reviews and Availability

Listings with a review score of 93 to 100 and 100 are on average more expensive than those with less than 93 (0-93). With the difference between (0 - 93) and (93 - 100) being less significant than the difference between (0 - 93) and 100, evident in (93 - 100)'s smaller coefficient and larger p-value. Interestingly review scores with null ratings are also on average more expensive, the effect is also much larger than those of (93 - 100) and 100.

Also interestingly, number of reviews has a very small negative coefficient showing that if a listing has more reviews it is on average slightly cheaper. When considering this as well as the positive effects of 30 and 365 day availability, it could be that more expensive listings are available more often and therefore booked and reviewed less often showing that there is higher demand for more affordable listings in the Cape Town Airbnb market.

Listings that are only available for short term rental are more expensive than those that are not and if a listing has a cancellation policy it is, on average, less expensive than those without one. If a listing is able to be booked instantly it is, on average, cheaper, potentially indicating that more expensive and therefore more upmarket listings prefer to review or apply some filter to the guests that they allow to rent their listing.

Costs

Listings requiring a security deposit are on average more expensive, potentially indicating more or higher quality amenities. Intuitively, listings that charge for additional people as well as cleaning have lower base prices compared to all inclusive listings that are on average more expensive.

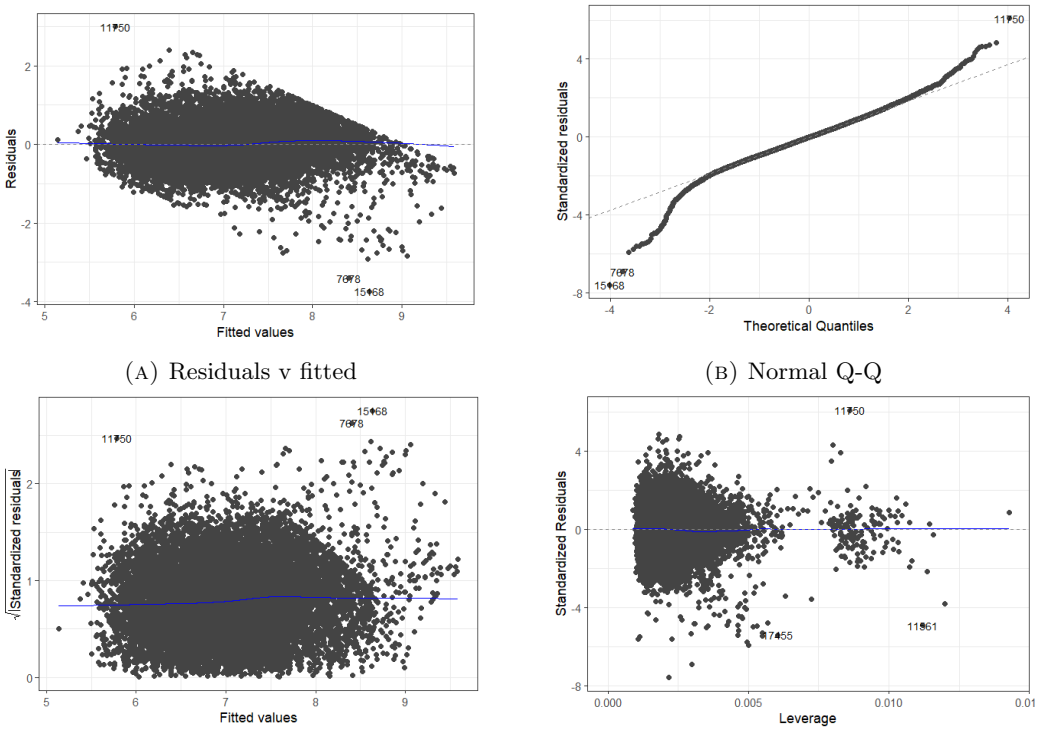
Looking at model diagnostics gives an indication of whether the model's assumptions are upheld.

Diagnostics

Most of the residuals are within the acceptable -2 to 2 range in Figure 5.4 (A) showing that the linearity assumption is largely upheld. While there seems to be a slight negative trend in the residuals for higher fitted values in Figure 5.4 (A), when considered with the residuals in Figure 5.4 (C) a relationship is less obvious indicating homoscedasticity. The errors appear to be normally distributed in Figure 5.4 (B) with the exception of the highlighted outliers. The leverage values in Figure 5.4 (D) are small and only one outlier, listing 11561 stands out.

Investigating this listing as well as other listings with higher leverage values, they appear to be Private / Shared rooms that accommodate few people and are expensive or accommodate many but are cheap. The former may be a data error or a strategy by the host who may only be willing to rent in the case that they were able to charge above-market rates. The latter is a type of listing such a hostel, guesthouse or hotel where the 'accommodates' reflects the number of guests the whole property can accommodate and not how many a room can accommodate. While this could be viewed as a data error, it appears to be quite common and is indicative of a specific type of listing. Both of these instances are features of the data and so the listings are left included in the data and analysis. The RMSLE performance metric helps mitigate these effects.

Looking at the predicted vs observed (PVO) log prices in Figure 5.5, it is evident that the OLS has high variance, shown by the wide scatter and deviation of points away from the diagonal line on the left and by the thickness of the black band on the right.



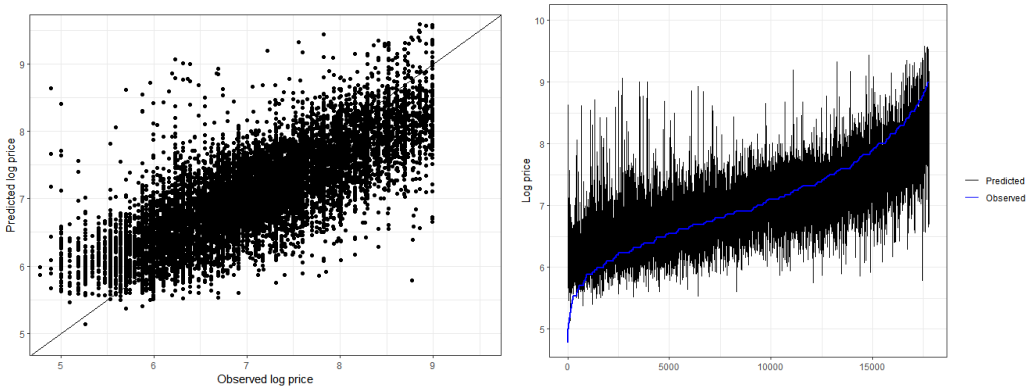
(A) Residuals v fitted

(B) Normal Q-Q

(C) Scale vs location

(D) Residuals vs Leverage

FIGURE 5.4: Model 2 diagnostics

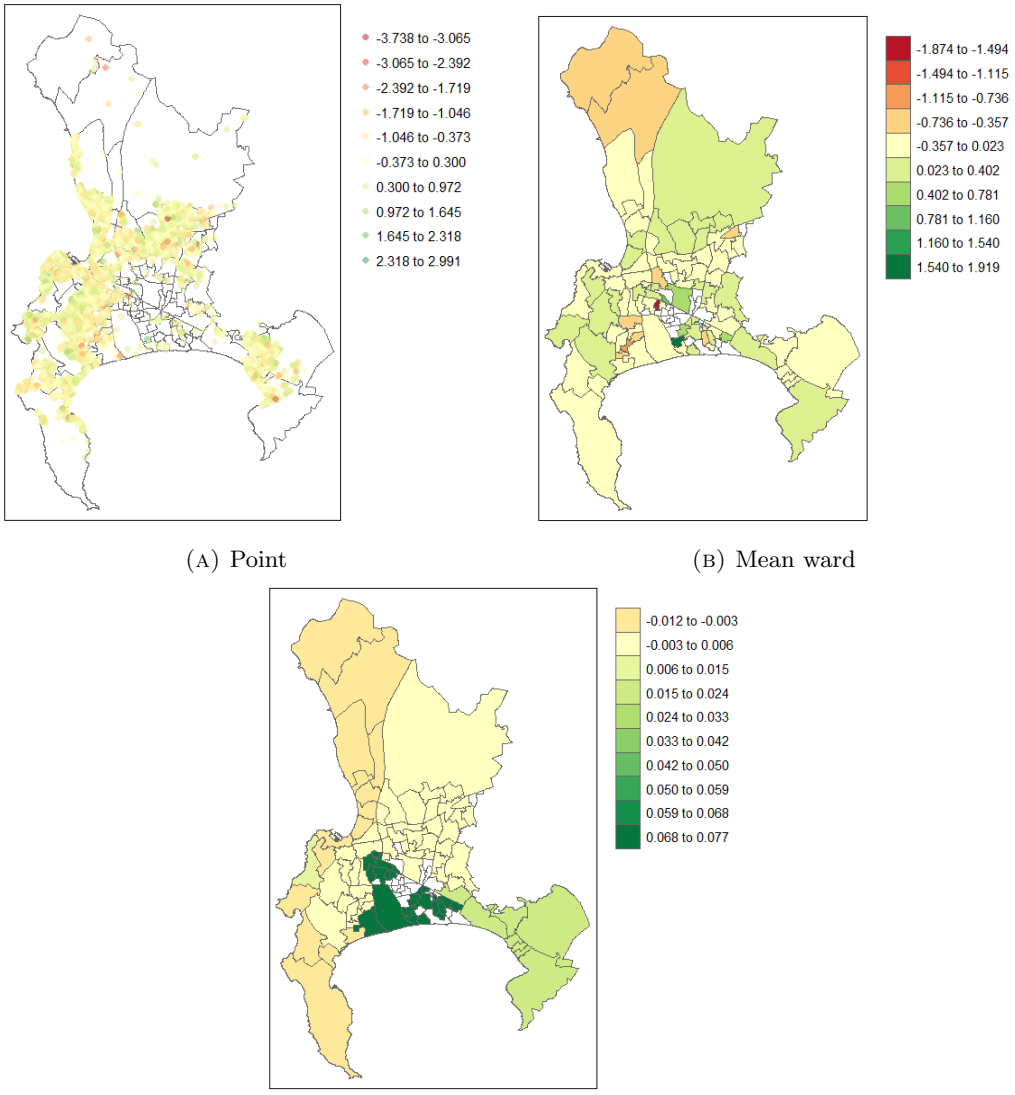


(A) PVO scatter

(B) PVO ordered

FIGURE 5.5: OLS training PVOs

Darker red and darker green dots represent the higher over and under predicted listings in the mapped residuals in Figure 5.6 (A). Averaging the residuals of the listings in a ward, Figure 5.6 (B) shows most wards averaging between -0.357 to 0.402. Wards in the far North of the West Coast and in the South Eastern suburbs, where there are few listings have higher average absolute residuals. Zooming into Figure 5.6 (B), Figure 5.7 highlights a few of the worst predicted wards as Gugulethu, Mitchell's Plain and Cafda Village.



(A) Point (B) Mean ward

(C) Mean suburb

FIGURE 5.6: OLS training residuals

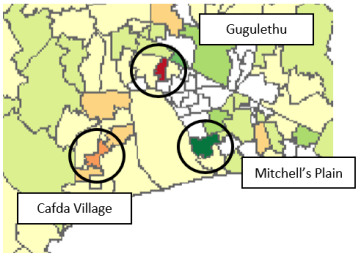


FIGURE 5.7: OLS training poor performance wards

Spatial dependence and Moran's I

Moran's I is a statistical test that can be conducted to assess the presence of spatial autocorrelation. The test has a null hypothesis of there being no autocorrelation and the alternate

hypothesis is that there is positive spatial correlation at some specified neighbourhood.

The test is performed on the training set dependent variable, log price, as well as on the training set OLS residuals. Variations of neighbourhoods using fixed and adaptive Gaussian kernels are tested and the results are presented in Tables 5.4 and 5.5 respectively.

Mapping log price at the listing, average ward and average suburb level as in Figure 5.2, spatial dependence appears to exist and this is confirmed as significant positive autocorrelation was found in log price by the Moran's I test at all the kernel specifications in Table 5.4. For adaptive kernels, a fewer number of neighbours saw a higher correlation and for the fixed kernel, a bandwidth of 0.25km or 250m had the largest correlation with the correlation decreasing for smaller and larger bandwidths thereafter. The largest correlations for adaptive and fixed kernels were therefore 0.264 and 0.250 respectively.

Adaptive kernels are widely chosen over fixed kernels in the literature as they are able to handle varying data with non constant densities. Since the density of Cape Town's Airbnbs varies across the city (as seen in Figure 4.11) and since the adaptive kernel returned the highest correlation value of 0.264, in order to simplify the remainder of the analysis, only an adaptive kernel is used in analysis.

TABLE 5.4: Moran's I on log price

			I	**E(I)**	**var(I)**	**Std. dev**	***p*-value**
		knn					
Gaussian	**Adaptive**	5	0.264	0.000	0.000	57.89	0.000
		10	0.257	0.000	0.000	78.97	0.000
		15	0.250	0.000	0.000	93.72	0.000
		25	0.240	0.000	0.000	116.29	0.000
		50	0.226	0.000	0.000	154.87	0.000
		b (km)					
	Fixed	0.125	0.189	0.000	0.000	12.85	0.000
		0.25	0.250	0.000	0.000	26.85	0.000
		0.5	0.215	0.000	0.000	36.76	0.000
		0.75	0.205	0.000	0.000	45.77	0.000
		1	0.189	0.000	0.000	50.89	0.000
		2	0.146	0.000	0.000	72.25	0.000

The adaptive Gaussian kernel is used to test for spatial autocorrelation in the Model 2 training residuals. The tests return significant positive spatial autocorrelation for all kernel specifications, however the autocorrelations were much smaller than the y variable correlations. This is to be expected given that the regression model includes multiple locational variables which would address some of the spatial dependence. Autocorrelation decreases with more neighbours, as

was with the log price Moran's I tests which is inline with Tobler's law of everything being related to everything else, but nearer things more so than distant things (Tobler, 1970).

TABLE 5.5: Moran's I on regression residuals

		I	E(I)	var(I)	Std. dev	*p*-value
	knn					
Gaussian adaptive	5	0.119	0.000	0.000	26.22	0.000
	10	0.115	0.000	0.000	35.74	0.000
	15	0.109	0.000	0.000	41.49	0.000
	25	0.101	0.000	0.000	49.69	0.000
	50	0.090	0.000	0.000	63.23	0.000

Spatial lag model

While the inclusion of spatial variables in the OLS model did reduce spatial autocorrelation, it did not fully account for it and so spatial models are explored to address the remaining dependence that may exist in the data.

As with OLS, five-fold cross validation is used to determine which adaptive Gaussian kernel spatial lag model performs the best. The first three model specifications from Table 5.1 performed the best in OLS and so are also run for the lag model. The only difference in the spatial lag specification is the exclusion of latitude and longitude as they are incorporated through the use of the spatial weights matrix.

The cross validation RMSE and RMSLE as well as the model results from the training conducted on the whole training set after cross validation are shown in Table 5.6. Validation RMSE and RMSLE does not vary across the full, step and third model and all models' performance increases with an increasing number of neighbours. The lowest RMSE and RMSLE are 1008 and 0.490 respectively. This is lower than the 1019 and 0.493 from the OLS models in Table 5.2 showing that, given that a lower error is associated with better performance, there is potential for the spatial lag model to improve performance over OLS. All models also have a lower AIC than their OLS 'equivalent', showing better fit.

The lag coefficient, ρ, that measures the size of the lagged effect, is significant for all models but its magnitude increases with an increase in number of neighbours. This, as well as validation errors, are inversely related to residual autocorrelation which increases with an increase in neighbours, showing that addressing spatial dependence in this fashion does not necessarily improve prediction. Akin to overfitting, only including a fewer number of neighbours does not allow the model to generalise sufficiently to predict on out of sample data.

TABLE 5.6: Spatial lag models

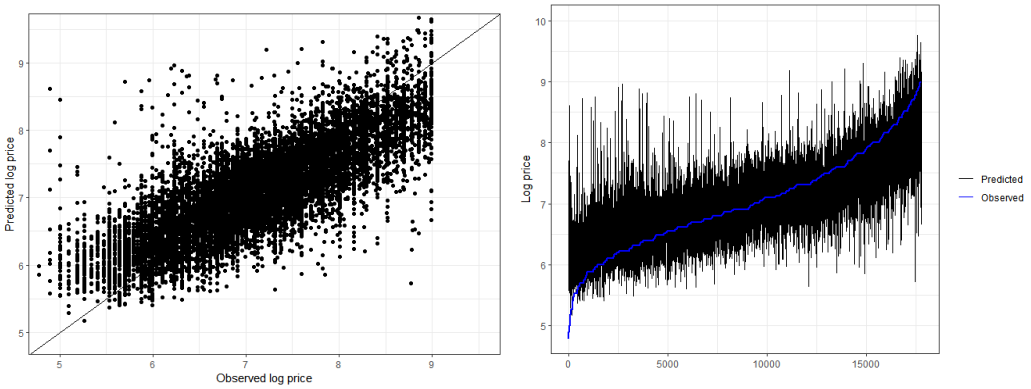
	knn	Validation RMSE	Validation RMSLE	Model AIC	OLS AIC	Likelihood ratio stat	ρ	p-value
Full model	**5**	1021	0.492	24697	25529	834	0.219	0.000
	10	1016	0.491	24377	25529	1154	0.307	0.000
	15	1012	0.490	24262	25529	1269	0.352	0.000
	25	1008	0.490	24137	25529	1394	0.405	0.000
	50	1008	0.491	24027	25529	1504	0.471	0.000
Step model	**5**	1021	0.492	24695	25528	834	0.219	0.000
	10	1016	0.491	24375	25528	1155	0.307	0.000
	15	1012	0.490	24260	25528	1269	0.352	0.000
	25	1007	0.490	24136	25528	1394	0.404	0.000
	50	1008	0.491	24026	25528	1504	0.470	0.000
Model 3	**5**	1022	0.493	24751	25656	907	0.226	0.000
	10	1016	0.492	24418	25656	1241	0.312	0.000
	15	1011	0.491	24301	25656	1357	0.356	0.000
	25	1007	0.491	24175	25656	1483	0.406	0.000
	50	1007	0.491	24113	25656	1545	0.420	0.000

Since performance between models one, two and three does not vary, the step model, model two, is chosen as the final model so as to stay consistent with the OLS specification. Fifty nearest neighbours are used for the final Gaussian adaptive kernel specification as this specification had the lowest validation errors.

After retraining the lag model on the full training set, the predicted vs observed (PVO) training log prices are shown in Figure 5.8. The PVO results look similar to those of OLS with large variation and over and under prediction on either end of the log price spectrum.

The mapped residuals in Figure 5.9 (B) also look similar to that of OLS with prices for listings in Gugulethu and Mitchell's Plain still being poorly predicted. In Figure 5.9 (C) the Atlantic Seaboard is the only suburb with high under prediction.

The final model specification is shown in Table 5.7, however the coefficients only represent the direct effect of the variable. Given that the model also includes the effects of the variables from other observations, there is also an indirect effect which is shown in Table 5.8.

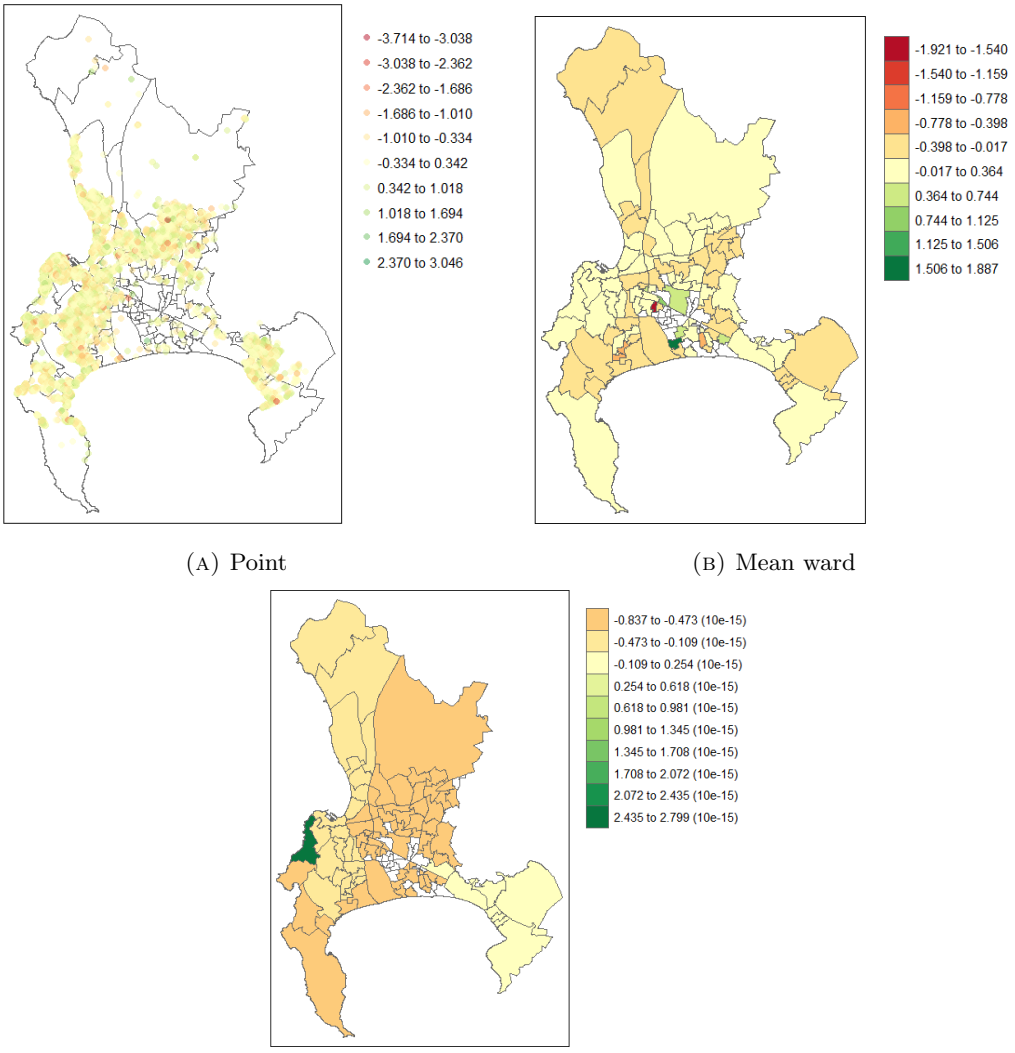


(A) PVO scatter (B) PVO ordered

FIGURE 5.8: Lag training PVOs

TABLE 5.7: OLS and spatial lag and spatial error models

		OLS		**Spatial lag**		**Spatial error**	
		Coeff	t-stat	Coeff	z-stat	Coeff	z-stat
Category	Variable	(Std err)	(p-value)	(Std err)	(p-value)	(Std err)	(p-value)
	(Intercept)	40.57	15.53	3.48	44.15	6.23	141.76
		(2.61)	(0.00)	(0.08)	(0.00)	(0.04)	(0.00)
Basic	Accommodates	0.17	86.75	0.17	88.40	0.17	87.14
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	Room type: Hotel room	−0.06	−2.42	−0.03	−1.13	−0.04	−1.62
		(0.02)	(0.02)	(0.02)	(0.26)	(0.02)	(0.10)
	Room type: Private/ shared room	−0.42	−39.09	−0.39	−37.73	−0.41	−38.61
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
Amenities	Wifi	0.10	7.16	0.07	5.72	0.07	5.62
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Free parking	0.09	8.98	0.08	8.06	0.09	9.41
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Pool	0.18	22.85	0.14	18.07	0.18	21.95
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Hot tub	0.06	3.85	0.05	3.55	0.05	3.60
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Kitchen	−0.12	−8.50	−0.10	−7.15	−0.11	−7.72
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Aircon	0.23	25.53	0.20	22.47	0.21	23.26
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Washer	0.11	11.78	0.11	12.17	0.11	11.90
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Beach / water front	0.07	5.69	0.02	2.05	0.04	2.97
		(0.01)	(0.00)	(0.01)	(0.04)	(0.01)	(0.00)
Host	Verified host identity	0.04	4.62	0.04	4.57	0.04	4.85
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Host listings count	0.00	6.93	0.00	3.75	0.00	5.97
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	House rules	−0.02	−2.08	−0.00	−0.62	−0.01	−0.81
		(0.01)	(0.04)	(0.01)	(0.53)	(0.01)	(0.42)
Location	Longitude	−1.86	−13.15				
		(0.14)	(0.00)				
	Exact location	−0.04	−4.93	−0.03	−4.00	−0.04	−4.31
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Suburbs: City Bowl	−0.16	−11.02	−0.09	−6.52	−0.21	−11.53
		(0.01)	(0.00)	(0.01)	(0.00)	(0.02)	(0.00)
	Suburbs: Eastern	0.15	2.35	−0.12	−2.59	−0.41	−6.27
		(0.06)	(0.02)	(0.05)	(0.01)	(0.06)	(0.00)
	Suburbs: Northern	−0.19	−5.43	−0.20	−6.86	−0.44	−11.08
		(0.03)	(0.00)	(0.03)	(0.00)	(0.04)	(0.00)
	Suburbs: South East	−0.26	−5.25	−0.16	−3.33	−0.41	−6.41
		(0.05)	(0.00)	(0.05)	(0.00)	(0.06)	(0.00)
	Suburbs: South Peninsula	−0.31	−19.56	−0.22	−14.01	−0.36	−17.28
		(0.02)	(0.00)	(0.02)	(0.00)	(0.02)	(0.00)
	Suburbs: Southern	−0.23	−11.70	−0.18	−10.06	−0.30	−11.88
		(0.02)	(0.00)	(0.02)	(0.00)	(0.02)	(0.00)
	Suburbs: West Coast	−0.29	−14.73	−0.27	−17.10	−0.46	−22.22
		(0.02)	(0.00)	(0.02)	(0.00)	(0.02)	(0.00)
	Distance to airport	0.00	4.57	0.00	3.37	0.00	8.70
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	Distance to nearest attraction	0.00	2.42	−0.00	−5.35	−0.00	−3.54
		(0.00)	(0.02)	(0.00)	(0.00)	(0.00)	(0.00)
Reviews	Number of reviews	−0.00	−7.72	−0.00	−8.29	−0.00	−8.16
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	Reviews: 100 (0.01)	0.14	12.38	0.12	10.66	0.12	10.67
		(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	
	Reviews: 93-99	0.02	1.95	0.00	0.40	0.00	0.31
		(0.01)	(0.05)	(0.01)	(0.69)	(0.01)	(0.76)
	Reviews: null	0.25	21.80	0.23	20.95	0.23	20.98
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
Costs	Cleaning fee	−0.06	−7.20	−0.06	−6.65	−0.06	−6.58
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Security deposit	0.08	8.27	0.08	8.57	0.07	8.35
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	All inclusive	0.08	7.51	0.08	8.11	0.08	7.84
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Additional charge	−0.21	−23.15	−0.20	−22.58	−0.20	−22.44
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
Availability	Short term only	0.04	4.56	0.05	5.48	0.05	5.22
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Availability 30 days	0.08	7.02	0.08	7.12	0.08	7.01
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Availability 365 days	0.05	3.93	0.05	4.21	0.05	3.83
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Instantly bookable	−0.05	−6.64	−0.04	−5.73	−0.05	−6.17
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	Cancellation policy	−0.13	−16.20	−0.11	−14.60	−0.12	−14.85
		(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)
	ρ			0.40	0.00		
				(0.01)	(0.00)		
	λ					0.28	0.00
						(0.01)	(0.00)
Log Likelihood				−12027		−12401	
AIC (Linear model)				25527		25527	
AIC (Spatial model)				24135		24882	
LR test: statistic				1394		646	
LR test: p-value				0.00		0.00	



(A) Point (B) Mean ward

(C) Mean suburb

FIGURE 5.9: Mapped spatial lag model training residuals

Spatial error model

Addressing spatial dependence through a weighted error term, the SEM results are shown in Table 5.9. As with the SLM, cross-validation performance did not vary between full, step and model 3 specifications but model AIC is better than the OLS equivalent for all specifications. The λ coefficient is also significant for all model specifications with the coefficient size growing with an increasing number of neighbours.

Cross validation errors worsen with a larger number of neighbours, the lowest RMSE and RMSLE is 1025 and 0.496 respectively. These are higher when compared to both OLS and the lag models indicating that, even while λ is significant and therefore that error terms are autoregessesion, addressing the dependence with this specification may not improve predictive performance.

TABLE 5.8: Direct and indirect lag effects

Category	Variable	Direct	Indirect	Total
Basic	Accommodates	0.17	0.11	0.29
	Room type: Hotel room	-0.03	-0.02	-0.04
	Room type: Private/ shared room	-0.39	-0.26	-0.65
Amenities	Wifi	0.07	0.05	0.12
	Free parking	0.08	0.05	0.13
	Pool	0.14	0.09	0.24
	Hot tub	0.05	0.03	0.08
	Kitchen	-0.10	-0.07	-0.16
	Aircon	0.20	0.13	0.33
	Wash	0.11	0.07	0.19
	Beach / water front	0.02	0.02	0.04
Host	Verified host identity	0.04	0.03	0.06
	Host listings count	0.00	0.00	0.00
	House rules	0.00	0.00	-0.01
Location	Exact location	-0.03	-0.02	-0.05
	Suburb: City Bowl	-0.09	-0.06	-0.15
	Suburb: Eastern	-0.12	-0.08	-0.21
	Suburb: Northern	-0.20	-0.13	-0.34
	Suburb: South East	-0.16	-0.11	-0.27
	Suburb: South Peninsula	-0.22	-0.15	-0.37
	Suburb: Southern	-0.18	-0.12	-0.31
	Suburb: West Coast	-0.27	-0.18	-0.46
	Distance to airport	0.00	0.00	0.00
	Distance to nearest attraction	0.00	0.00	0.00
Reviews	Number of reviews	0.00	0.00	0.00
	Reviews: 100	0.12	0.08	0.20
	Reviews: 93.99	0.00	0.00	0.01
	Reviews: null rating	0.23	0.15	0.39
Costs	Cleaning fee	-0.06	-0.04	-0.10
	Security deposit	0.08	0.05	0.13
	All inclusive	0.08	0.05	0.14
	Additional charge	-0.20	-0.13	-0.33
Availability	Short term only	0.05	0.03	0.08
	Availability 30 days	0.08	0.05	0.14
	Availability 365 days	0.06	0.04	0.09
	Instantly bookable	-0.04	-0.03	-0.07
	Cancellation policy	-0.11	-0.08	-0.19

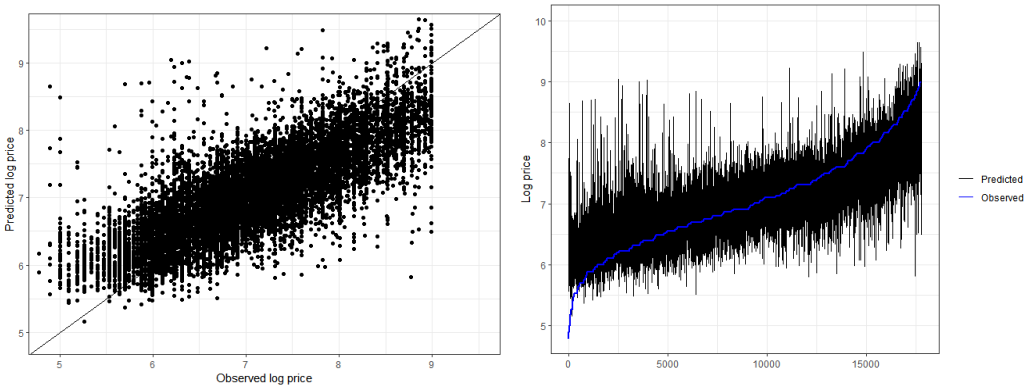
TABLE 5.9: Spatial error models

	knn	Validation RMSE	Validation RMSLE	Model AIC	OLS AIC	Likelihood ratio stat	λ	p-value
Full model	**5**	1026	0.496	24885	25529	646	0.276	0.000
	10	1027	0.497	24559	25529	972	0.412	0.000
	15	1028	0.497	24402	25529	1129	0.493	0.000
	25	1030	0.498	24249	25529	1282	0.588	0.000
	50	1030	0.498	24087	25529	1444	0.711	0.000
Step model	**5**	1025	0.496	24883	25528	647	0.275	0.000
	10	1027	0.497	24558	25528	971	0.412	0.000
	15	1028	0.497	24403	25528	1127	0.492	0.000
	25	1029	0.498	24249	25528	1281	0.587	0.000
	50	1030	0.498	24089	25528	1441	0.710	0.000
Model 3	**5**	1028	0.498	24965	25656	693	0.284	0.000
	10	1029	0.499	24622	25656	1036	0.422	0.000
	15	1030	0.499	24459	25656	1199	0.503	0.000
	25	1031	0.500	24295	25656	1363	0.598	0.000
	50	1031	0.500	24127	25656	1532	0.720	0.000

With this in mind, step model 2 with 5 neighbours is chosen as the final error model. The summary for this model is available in Table 5.7. The coefficients do not differ greatly from OLS

model 2 showing model stability.

From Table 5.10 it is evident that with a p-value of 0.946, residuals of the error model are not significantly correlated, showing that the model was able to address the spatial autocorrelation that was unaddressed by OLS.



(A) PVO scatter (B) PVO ordered

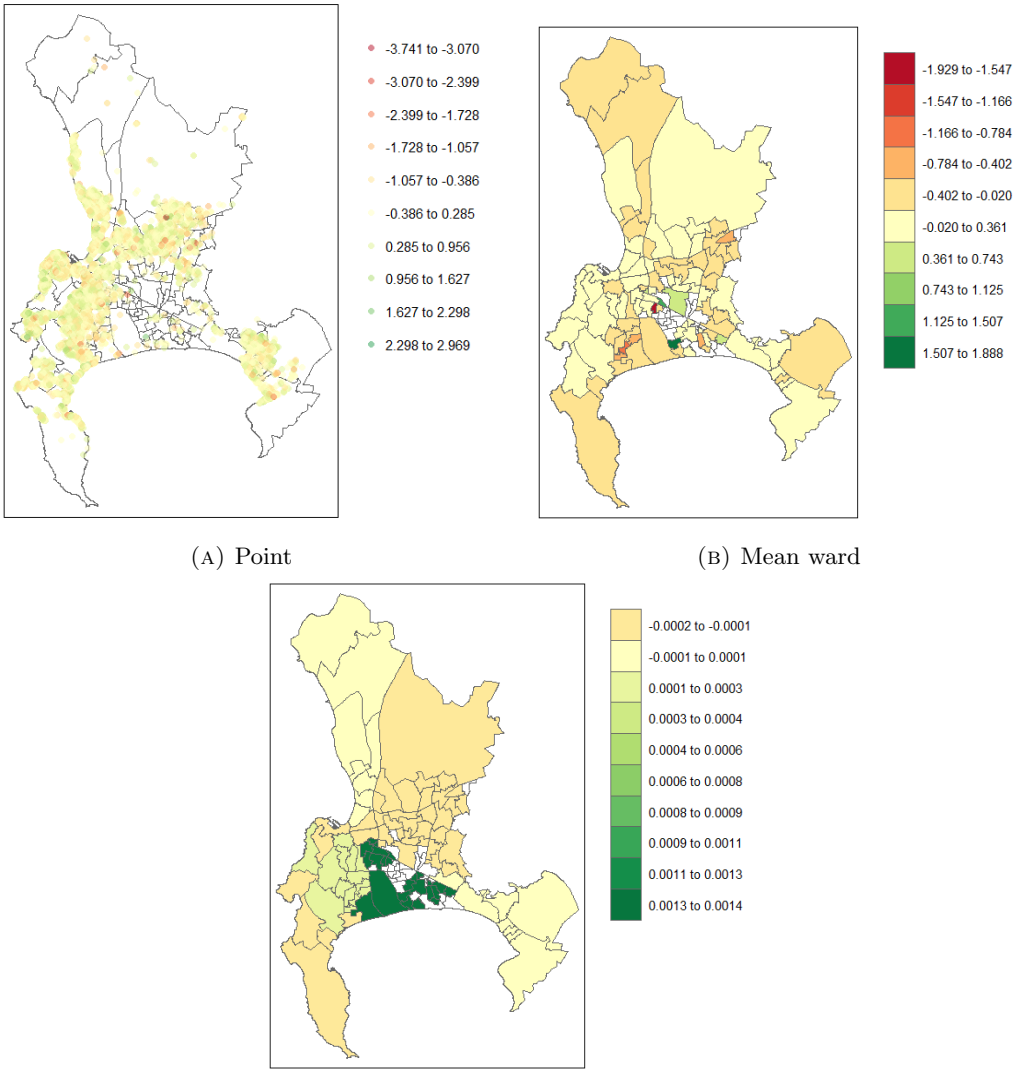
FIGURE 5.10: Error training PVOs

Even after addressing some of the spatial autocorrelation, the predicted vs observed log prices in Figure 5.10 look similar to that of OLS and the lag model. The mapped residuals in Figure 5.11 show the Atlantic seaboard still being under predicted in addition to listings in the Eastern suburbs, West coast and Southern suburbs.

TABLE 5.10: Moran's I on spatial regression residuals

Model	**I**	**E(I)**	**var(I)**	**Std. dev**	***p*-value**
Spatial lag	0.010	0.000	0.000	4.94	0.000
Spatial error	-0.007	0.000	0.000	-1.61	0.946
GWR	0.007	0.000	0.000	13.06	0.000

Given that both the lag and error model parameters are found to be statistically significant, robust Lagrangian multiplier tests are conducted to test the significance of each effect in the presence of the other. The results from Table 5.11 show that for different kernel specifications, a near zero p-value shows the lag and error effect to be significant, even in the presence of each other.



(A) Point (B) Mean ward

(C) Mean suburb

FIGURE 5.11: Mapped spatial error model training residuals

TABLE 5.11: Lagrangian multiplier tests

	Robust lag			Robust Error		
k	**LM**	**df**	***p*-value**	**LM**	**df**	***p*-value**
5	308	1	0.00	53	1	0.00
10	457	1	0.00	208	1	0.00
15	523	1	0.00	401	1	0.00
25	634	1	0.00	834	1	0.00
50	761	1	0.00	2030	1	0.00
75	768	1	0.00	3106	1	0.00
125	746	1	0.00	4811	1	0.00
200	658	1	0.00	6104	1	0.00
300	536	1	0.00	6567	1	0.00
500	447	1	0.00	6794	1	0.00
750	354	1	0.00	6569	1	0.00
1000	313	1	0.00	6550	1	0.00

Geographically weighted regression (GWR)

Accounting for spatial dependence and heterogeneity, a GWR is explored as the last linear spatial model. Results from five-fold cross validation using the same variables from the lag and

error models are shown in Table 5.12. Validation errors decrease with an increase in neighbours, minimising between 400 to 600 before increasing again. The lowest RMSE is 1024 and the lowest RMSLE is 0.4776. This is a higher validation RMSE but a lower RMSLE than previous models suggesting that the GWR might not greatly improve overall performance but does perform better when reducing the effect of outliers. Prioritising RMSLE, followed by RMSE, the choice is between 400 and 500 neighbours. Given that 500 is closer to 600, which achieved the lowest RMSE, the 500 neighbour model is chosen as the final model.

TABLE 5.12: GWR validation errors

knn	Validation RMSE	Validation RMSLE
5	1643	0.711
10	1539	0.576
25	1432	0.523
50	1337	0.509
75	1224	0.497
150	1087	0.483
200	1038	0.476
400	1026	0.476
500	1026	0.476
600	1024	0.477
700	1027	0.476
800	1035	0.478

Running this model, the parameters are shown in Table 5.13. There is large deviation from the global model for numerous parameters indicating the varying effect of certain variables in certain areas, in other words, spatial heterogeneity. The mapped coefficient variation for four variables is provided in Figure 5.13. In Figure 5.13 (A) the incremental increase in log prices with every increase in number of guests accommodated is less in the East than in the West. In Figure 5.13 (D), the presence of a pool in the City Bowl increases the listing prices less than it would elsewhere.

In Table 5.10 the GWR residuals have a very small, negligible, correlation of 0.007 showing, similarly to the error model, that spatial models are able to reduce spatial autocorrelation.

The predicted vs observed prices in Figure 5.13 appear to show slightly less variation than those seen previously, with a thickness of the black band decreasing slightly.

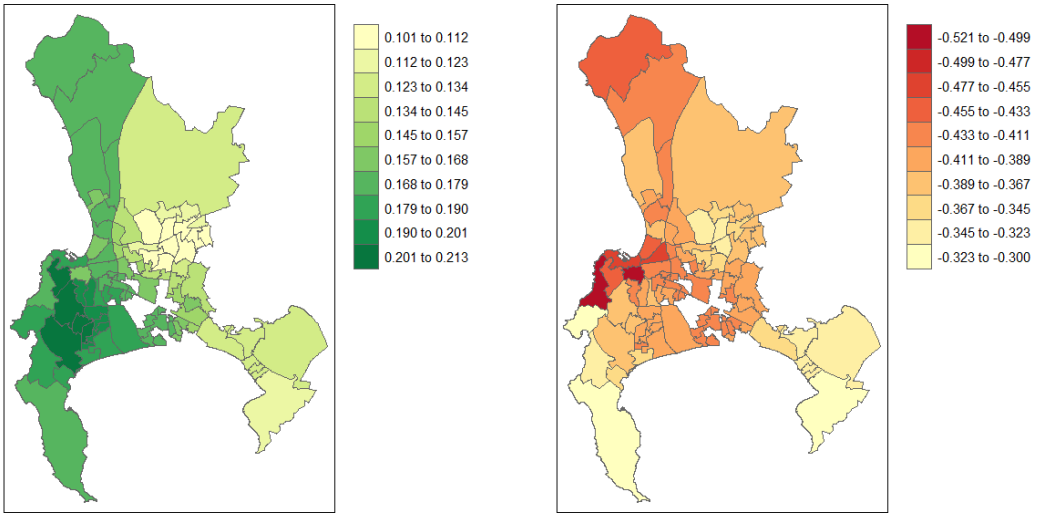
Mapped training residuals in Figure (B) 5.14 shows that on average, the residuals in the South Eastern suburbs have reduced but that the few previously mentioned wards are still problematic.

TABLE 5.13: GWR coefficient variation

Category	Variable	Min	Median	Max	Global
	(Intercept)	-1.94E+14	5.69	3.83E+13	6.21
Basic	Accommodates	0.10	0.18	0.23	0.17
	Room type: Hotel room	-0.35	-0.06	0.37	-0.06
	Room type: Private/ shared room	-0.57	-0.39	-0.25	-0.42
Amenities	Wifi	-0.08	0.08	0.21	0.10
	Free parking	-0.25	0.07	0.17	0.10
	Pool	0.05	0.17	0.30	0.19
	Hot tub	-0.05	0.03	0.16	0.06
	Kitchen	-0.32	-0.11	0.08	-0.12
	Aircon	0.06	0.19	0.30	0.23
	Wash	-0.01	0.11	0.17	0.11
	Beach / water front	-0.45	0.05	0.29	0.07
Host	Verified host identity	-0.07	0.03	0.12	0.04
	Host listings count	0.00	0.00	0.03	0.00
	House rules	-0.09	-0.01	0.09	-0.02
Location	Exact location	-0.13	-0.03	0.04	-0.04
	Suburb: City Bowl	-3.83E+13	-0.12	1.94E+14	-0.21
	Suburb: Eastern	-3.83E+13	-0.05	1.94E+14	-0.40
	Suburb: Northern	-3.83E+13	-0.13	1.94E+14	-0.43
	Suburb: South East	-3.83E+13	-0.17	1.94E+14	-0.40
	Suburb: South Peninsula	-3.83E+13	-0.33	1.94E+14	-0.35
	Suburb: Southern	-3.83E+13	-0.16	1.94E+14	-0.29
	Suburb: West Coast	-3.83E+13	-0.33	1.94E+14	-0.46
	Distance to airport	0.00	0.00	0.00	0.00
	Distance to nearest attraction	0.00	0.00	0.00	0.00
Reviews	Number of reviews	0.00	0.00	0.00	0.00
	Reviews: 100	-0.04	0.12	0.22	0.15
	Reviews: 93.99	-0.06	0.00	0.12	0.03
	Reviews: null rating	0.12	0.24	0.36	0.25
Costs	Cleaning fee	-0.20	-0.05	0.06	-0.06
	Security deposit	0.01	0.08	0.20	0.08
	All inclusive	-0.01	0.07	0.16	0.08
	Additional charge	-0.38	-0.21	-0.05	-0.21
Availability	Short term only	-0.12	0.04	0.15	0.04
	Availability 30 days	-0.11	0.08	0.19	0.09
	Availability 365 days	-0.13	0.04	0.21	0.05
	Instantly bookable	-0.15	-0.03	0.09	-0.05
	Cancellation policy	-0.19	-0.14	0.02	-0.13

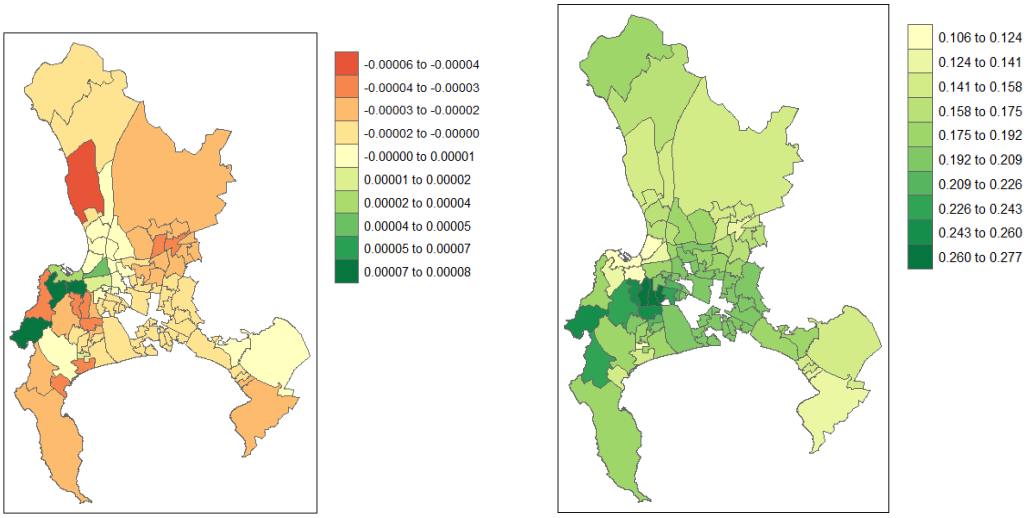
Random forest

The linear regression diagnostics alluded to potential non-linearity. In order to make non-linear predictions, random forests (RFs) are used. Given tree-based methods' ability to deduce variable importance, all variables are used in the model, the same as Model 1 in Table 5.1. Five-fold cross validation is used to determine how many variables, $Mtry$, should be considered at each split. The RMSE and RMSLE are reported in Table 5.14 along with $Min\ n$, the minimum node size. A $Mtry$ value of 19 variables returned the lowest validation RMSE and RMSLE of 882 and 0.439. This is slightly more variables than the traditional suggested p/3, that is, one third of all variables.



(A) Accommodates

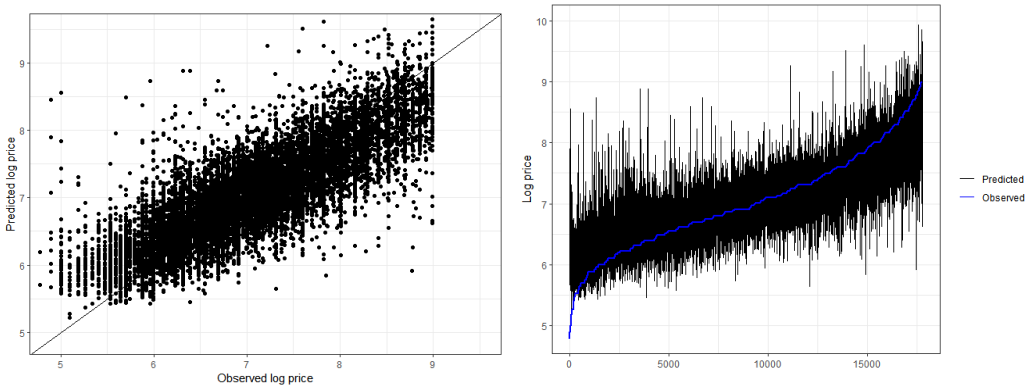
(B) Private / shared room



(C) Distance to nearest attraction

(D) Pool

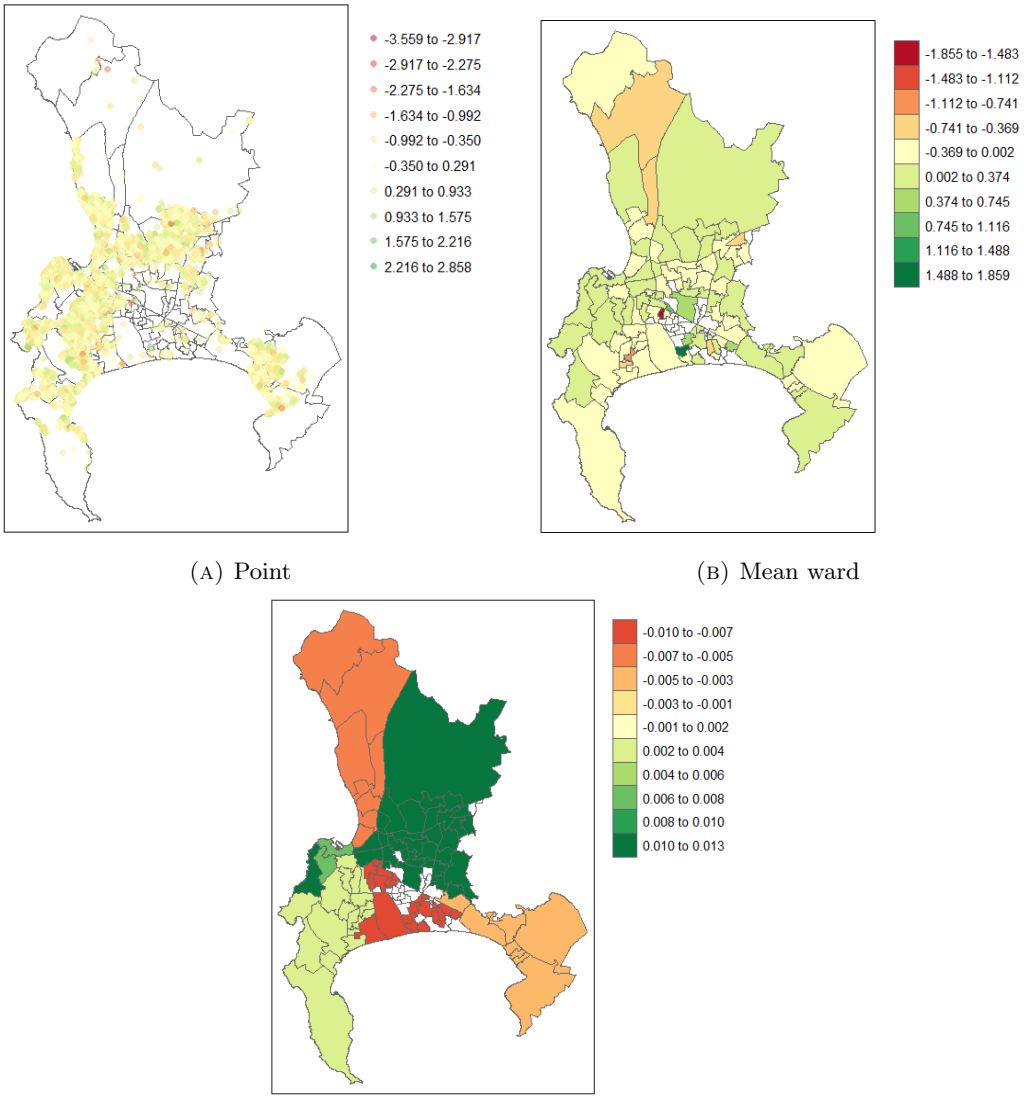
FIGURE 5.12: Mapped GWR coefficient variation



(A) PVO scatter

(B) PVO ordered

FIGURE 5.13: GWR training PVOs



(A) Point (B) Mean ward

(C) Mean suburb

FIGURE 5.14: Mapped GWR training residuals

TABLE 5.14: Random forest validation errors

Mtry	Min n	Validation RMSE	Validation RMSLE
5	**21**	930	0.446
8	**39**	911	0.446
19	**13**	882	0.439
20	**26**	887	0.442
28	**7**	884	0.442

Training the RF model on all training data and 19 variables at each split, the resultant explained variance was 70.15 and the RMSE and RMSLE were 872 and 0.435 respectively. This improvement from linear regression shows the importance of accommodating the non-linear nature of relationships.

Variable importance based on impurity is shown in Figure 5.15. As with previous models,

accommodates is the most important variable by far. Four out of the top eight variables are locational variables, showing the importance of both absolute position (latitude and longitude) as well as relative position (distance to nearest attraction and airport) in prediction. Similar to the OLS models, airconditioning and a pool are top variables, again showing insight into the warmer climate, summer-related aspect of the Cape Town Airbnb market. Variable importance is minimal after the first 17 variables, showing similarly to previous models that the likes of whether the host is a superhost and whether exact location is provided is less important.

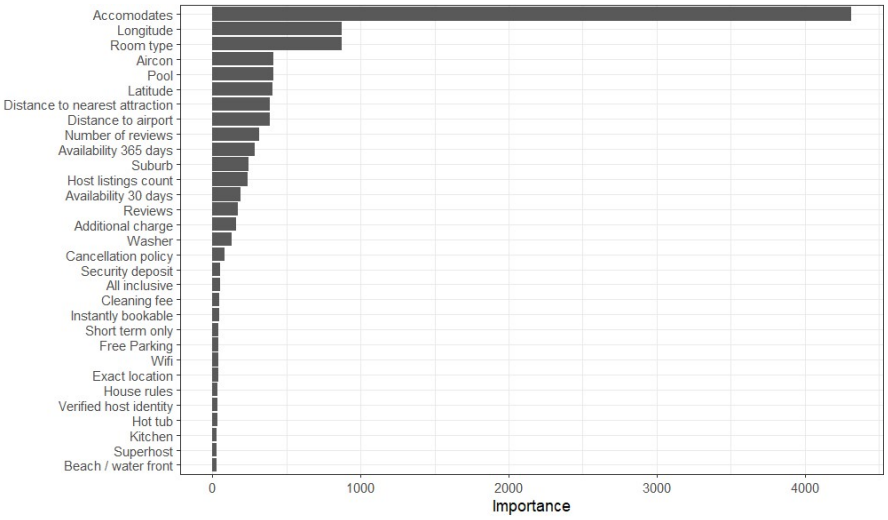


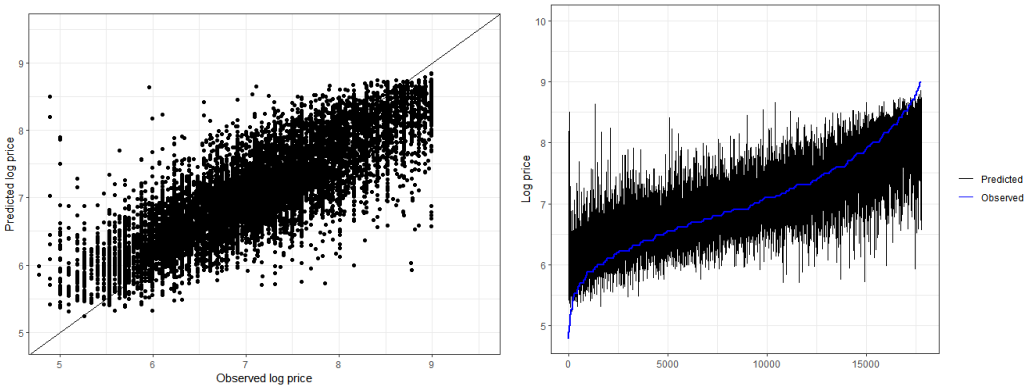
FIGURE 5.15: Random forest variable importance

The predicted versus observed log price plots in Figure 5.16 again show a reduction in variability but also slight improvement in the predictions in the tails, in other words, there is slight less over prediction for low prices and less under prediction for high prices.

The mapped residuals in Figure 5.17 shows that on average, more under prediction than over prediction occurs at the ward level and that at the suburb level, prices in the Atlantic Seaboard, City bowl, Southern Suburbs and Eastern suburbs appearing to be over predicted.

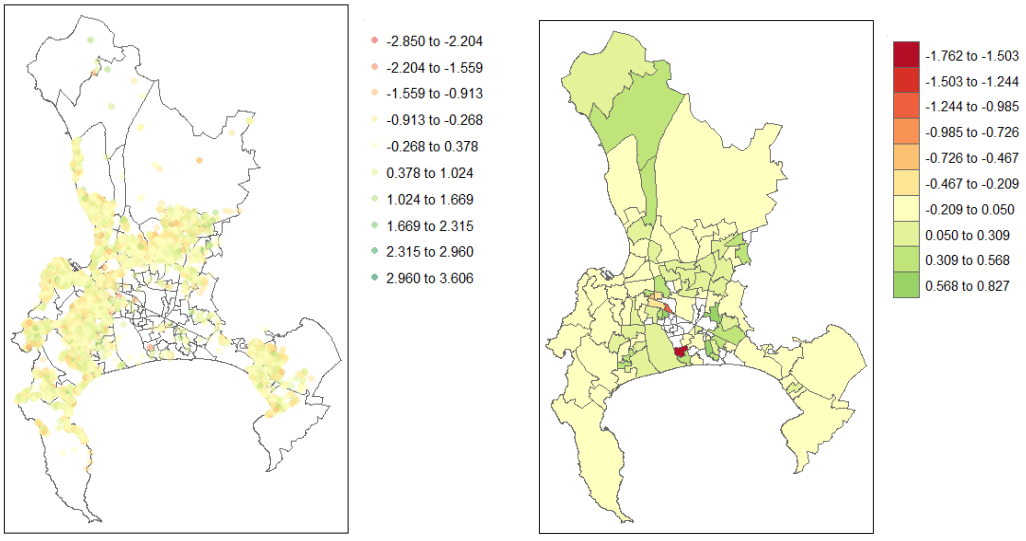
Gradient boosting machine

As with the other models, five-fold cross validation is performed to decide on final model parameters. Three values of learning rate, sample size and max tree depth were investigated in conjunction with the other model value defaults. The resultant validation errors are given in Table 5.15. Low learning rates reduce both RMSE and RMSLE for all tree depths and sample sizes. The model with both the lowest RMSE and RMSLE has a sample size of 0.4, tree depth of 15 and learning rate of 0.01, these are consequently chosen to be final model parameters.

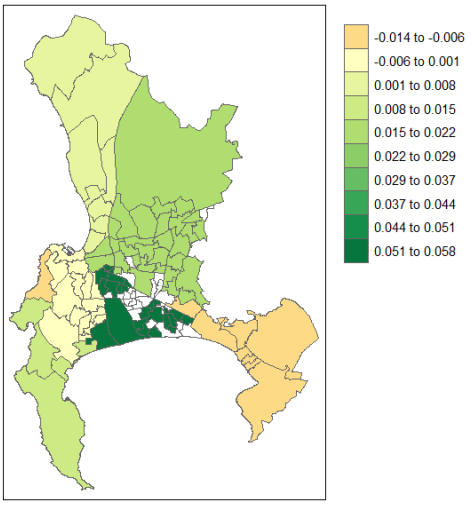


(A) PVO scatter (B) PVO ordered

FIGURE 5.16: Random forest training PVOs



(A) Point (B) Mean ward



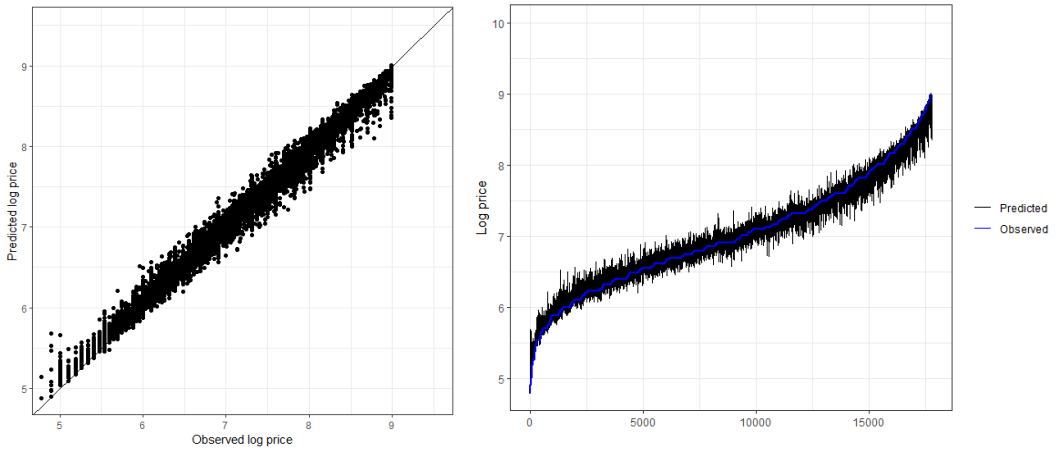
(C) Mean suburb

FIGURE 5.17: Mapped random forest training residuals

TABLE 5.15: GBM validation errors

		RMSE			RMSLE		
		Learning rate			Learning rate		
Sample size	**Max tree depth**	**0.1**	**0.05**	**0.01**	**0.1**	**0.05**	**0.01**
	15	873	855	836	0.443	0.430	0.418
0.4	**25**	875	857	840	0.440	0.429	0.419
	40	874	854	841	0.445	0.430	0.419
	15	868	853	837	0.438	0.428	0.418
0.5	**25**	867	856	842	0.437	0.428	0.419
	40	869	854	842	0.438	0.428	0.419
	15	861	849	843	0.435	0.425	0.420
0.6	**25**	869	856	844	0.437	0.427	0.420
	40	867	857	844	0.437	0.429	0.421

In the GBM predicted vs observed plots in Figure 5.18 it is evident how the sequential training on errors allows the the GBM to learn very well as there is very little variance and minimal bias, even in the tails.

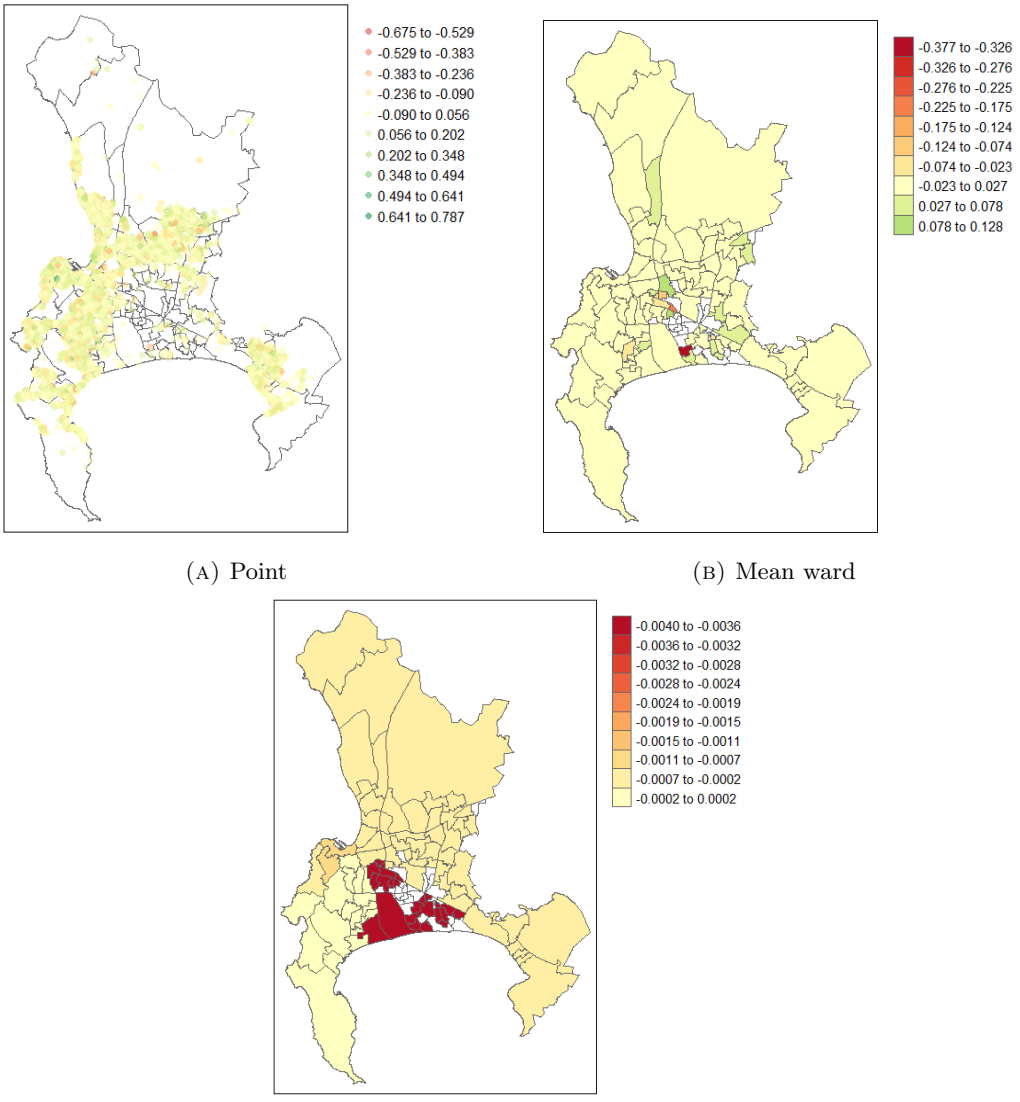


(A) PVO scatter (B) PVO ordered

FIGURE 5.18: GBM training PVOs

However, even with the improved global validation errors, from Figure 5.19 (B) the model still over predicts in Mitchell's Plain.

The GBM variable importance is given in Figure 5.20. As with all other models, accommodates is the most important by far. The next four most important variables are all locational variables.



(A) Point

(B) Mean ward

(C) Mean suburb

FIGURE 5.19: Mapped GBM training residuals

5.3 Test set performance

After training all the various models on the entire training set, they are then fit on the 20% test set data to judge their final performance on unseen data. The RMSE and RMSLE values are provided in Table 5.16.

The machine learning models, random forest and gradient boosting machine, ultimately achieve the best predictive performance having not only the lowest RMSLE but also RMSE showing better performance even with or without down weighting outliers. The GBM slightly outperforms the random forest achieving the lowest RMSE of 831 and RMSLE of 0.415.

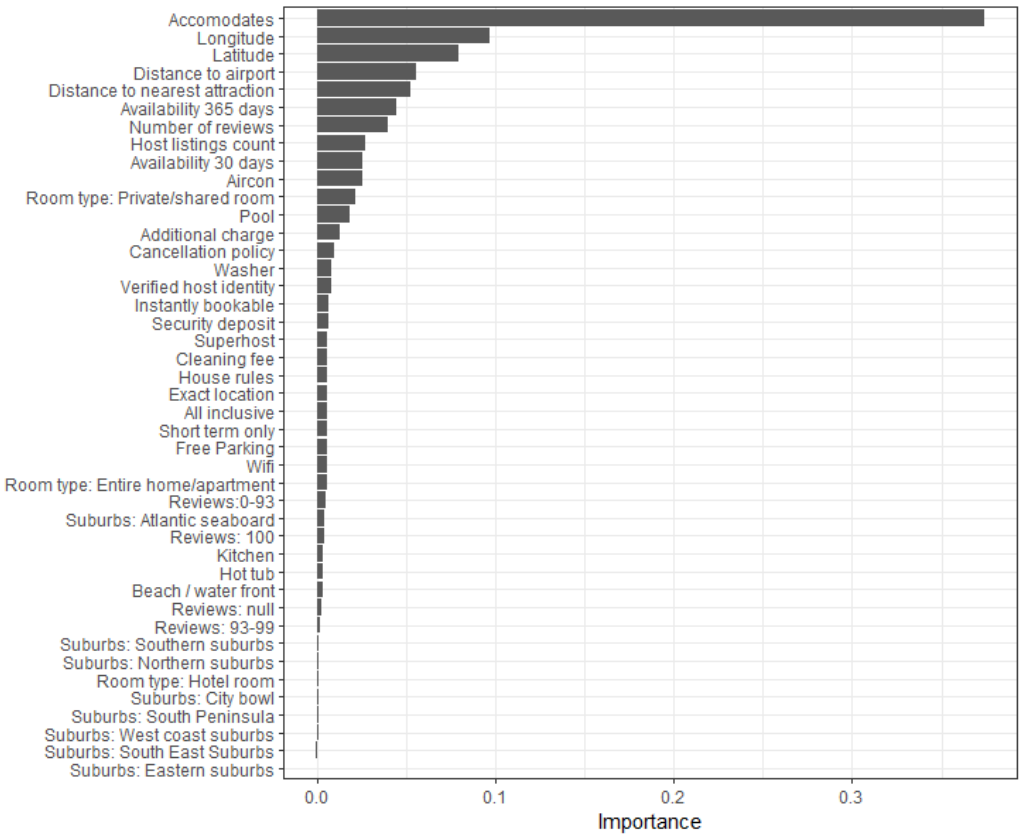


FIGURE 5.20: GBM variable importance

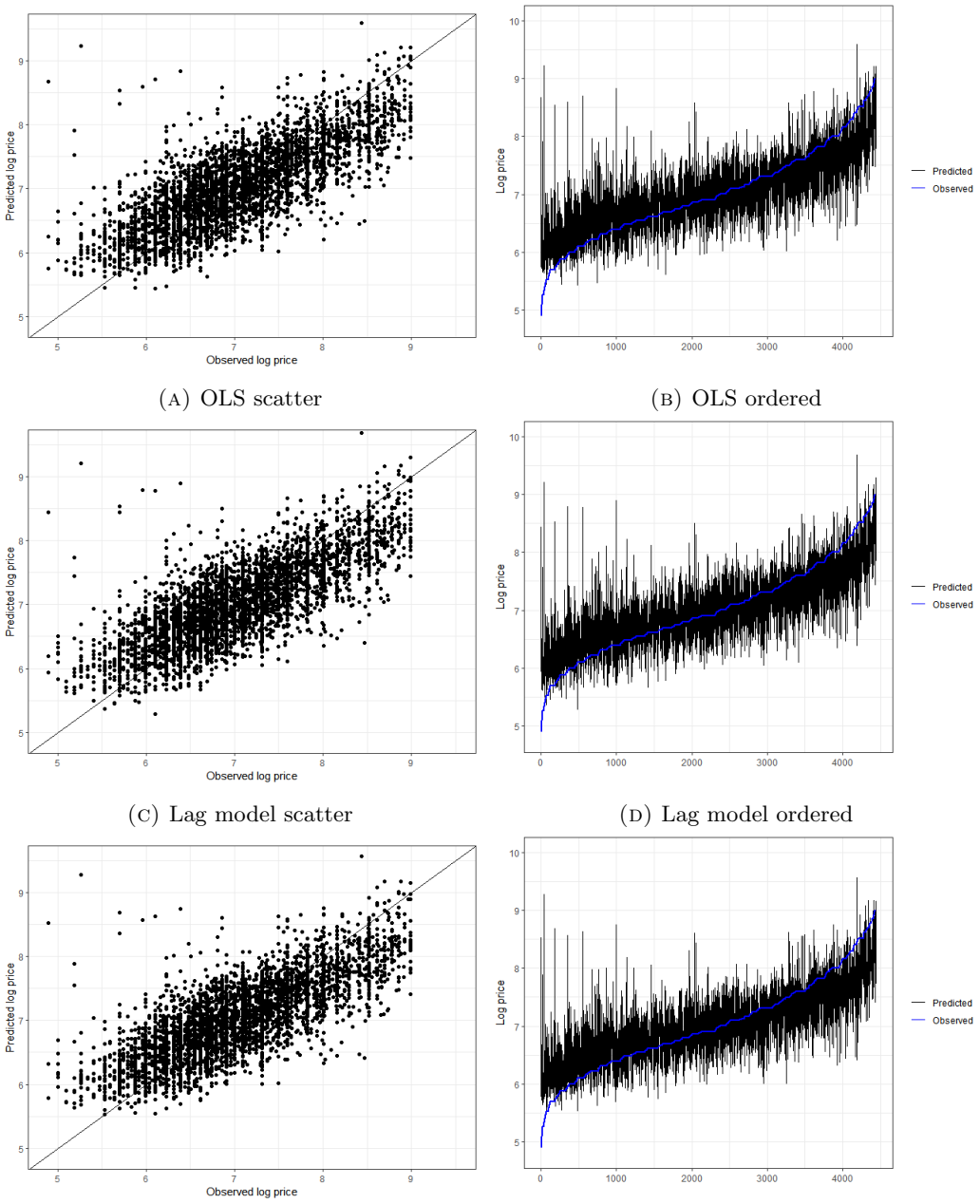
Despite the spatial lag and spatial error models reducing residual correlation in training, they did not improve performance over the OLS models, as was the case for the cross validation set. GWR did, however, outperform OLS as well as the other spatial models while also accounting for spatial dependence.

TABLE 5.16: Test set model performance

	RMSE	**RMSLE**
Regression	995	0.499
Spatial lag	1001	0.499
Spatial error	1007	0.501
Geographically weighted regression	981	0.470
Random forest	876	0.440
Gradient boosting machine	831	0.415

Figure 5.21 and Figure 5.22 plot the predicted log price versus observed log price for the various models. All models tend to over predict for smaller values of log price and under predict for high values while performance is better for mid range log price values. The OLS, lag and error models in Figure 5.21 tend to be more scattered while the GWR, random forest and GBM in Figure 5.22 are less scattered showing less variance with the most noticeable improvement being from GWR to random forest. Overall, the ordered PVOs on the right are quite similar across

all models showing that there is not a bias of over or under prediction across the whole model, only in the tails, while the scatter plots show that from OLS to GBM, there is a reduction in variance, with the points moving closer to the diagonal. The reduction in 'large spikes' in the ordered plots also indicate less listings being severely over or under predicted.



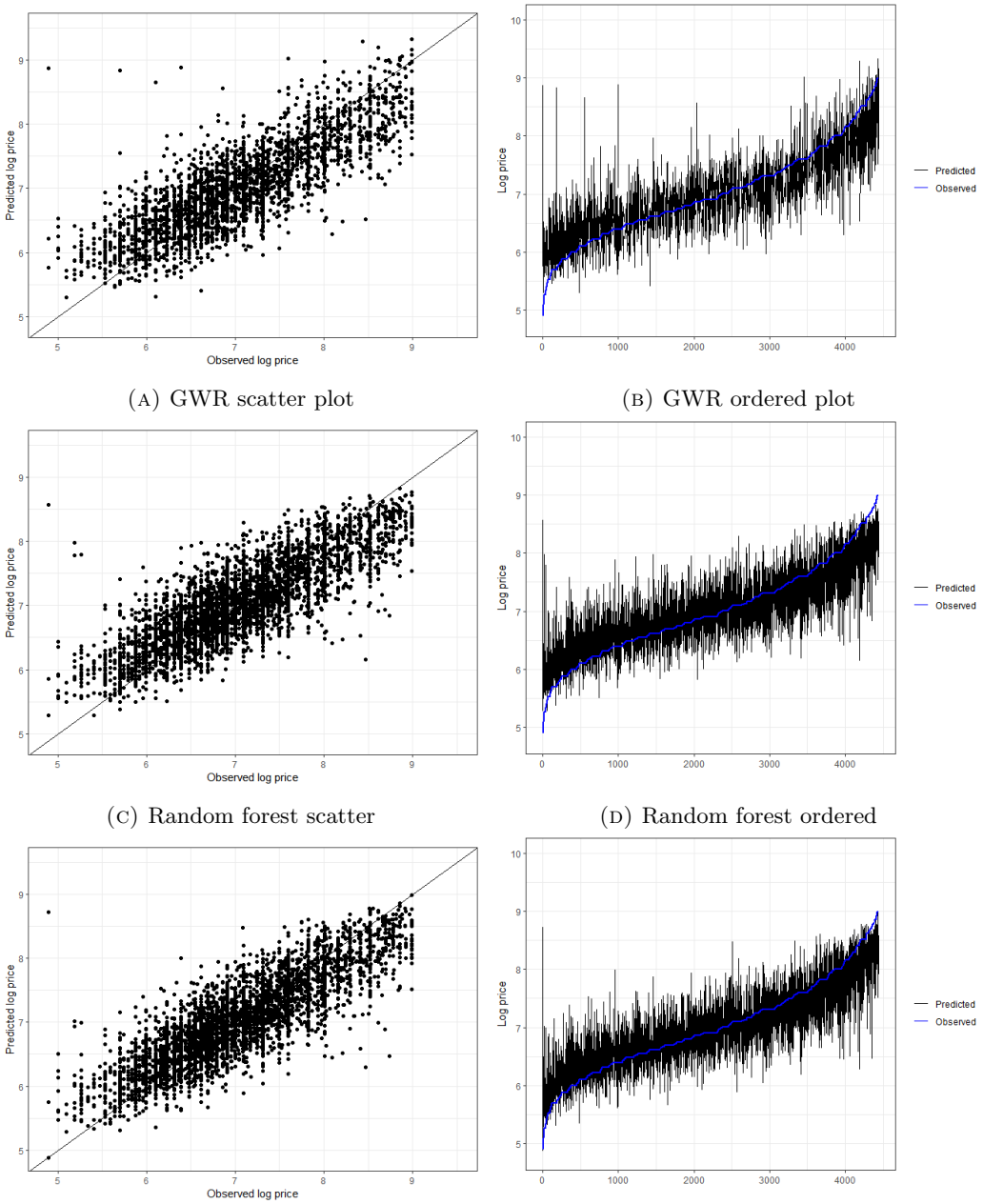
(A) OLS scatter (B) OLS ordered

(C) Lag model scatter (D) Lag model ordered

(E) Error model scatter (F) Error model ordered

FIGURE 5.21: OLS, lag and error model test set scattered and ordered predicted vs observed

Figures 5.23 and 5.24 show the residual values mapped for the various models. Given the number of listings, it is difficult to highlight specific instances or areas of over or under prediction.



(A) GWR scatter plot

(B) GWR ordered plot

(C) Random forest scatter

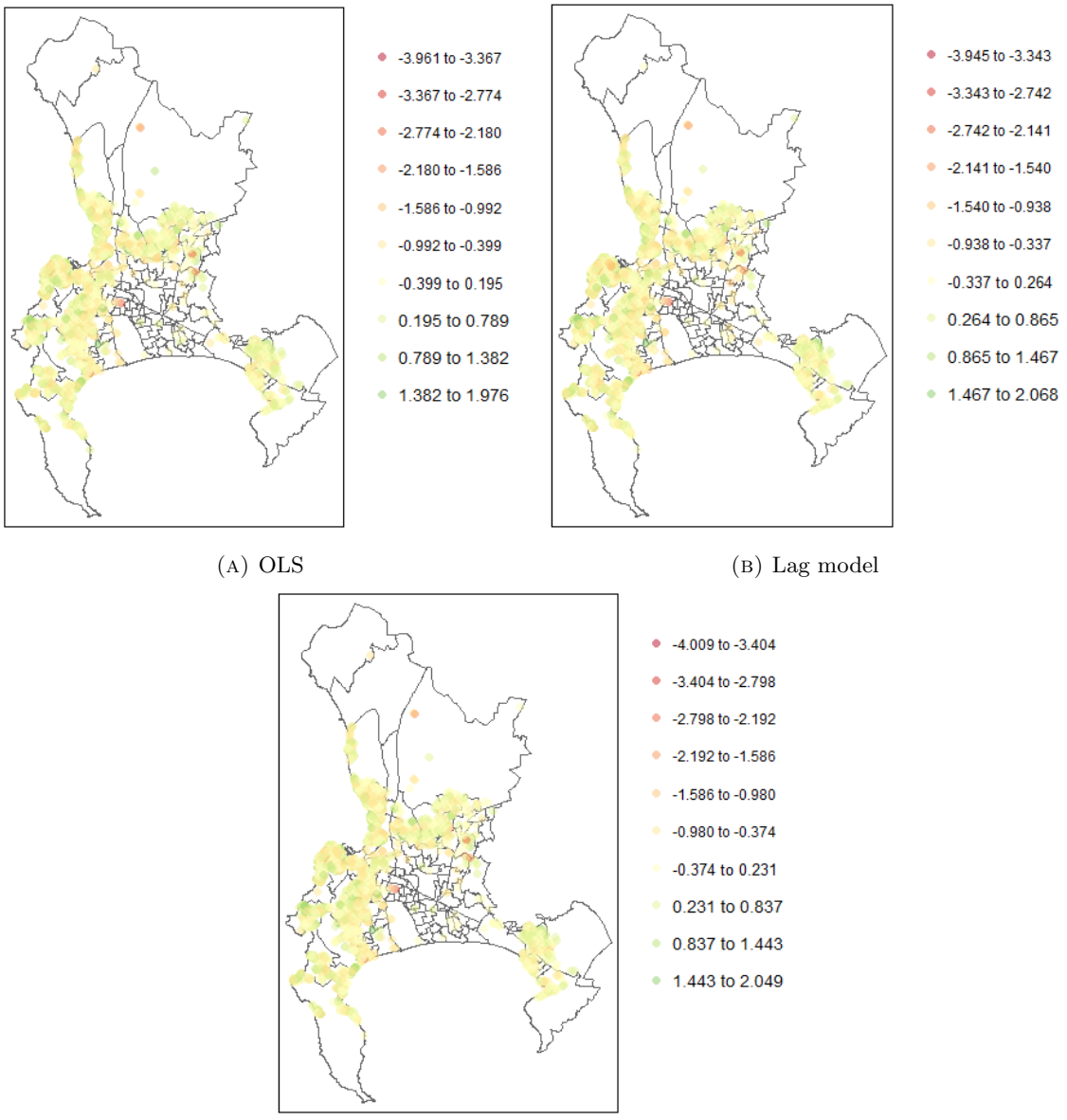
(D) Random forest ordered

(E) GBM scatter

(F) GBM ordered

FIGURE 5.22: GWR, random forest and GBM test set scattered and ordered predicted vs observed

Taking the mean residual value of all listings in a specific ward, Figures 5.25 and 5.26 provide clearer spatial views. For all models, most wards are the lightest shade of yellow corresponding to the smallest absolute error. From the OLS to the GBM models, the contrast in colours decreases showing a reduction in variation and overall smaller absolute residuals. In general, there are more orange and red wards than there are green showing that on average, over prediction is more common than under prediction. This is unsurprising given that model choices were based on RMSLE which penalises underestimation more than overestimation. For OLS, spatial lag and



(A) OLS

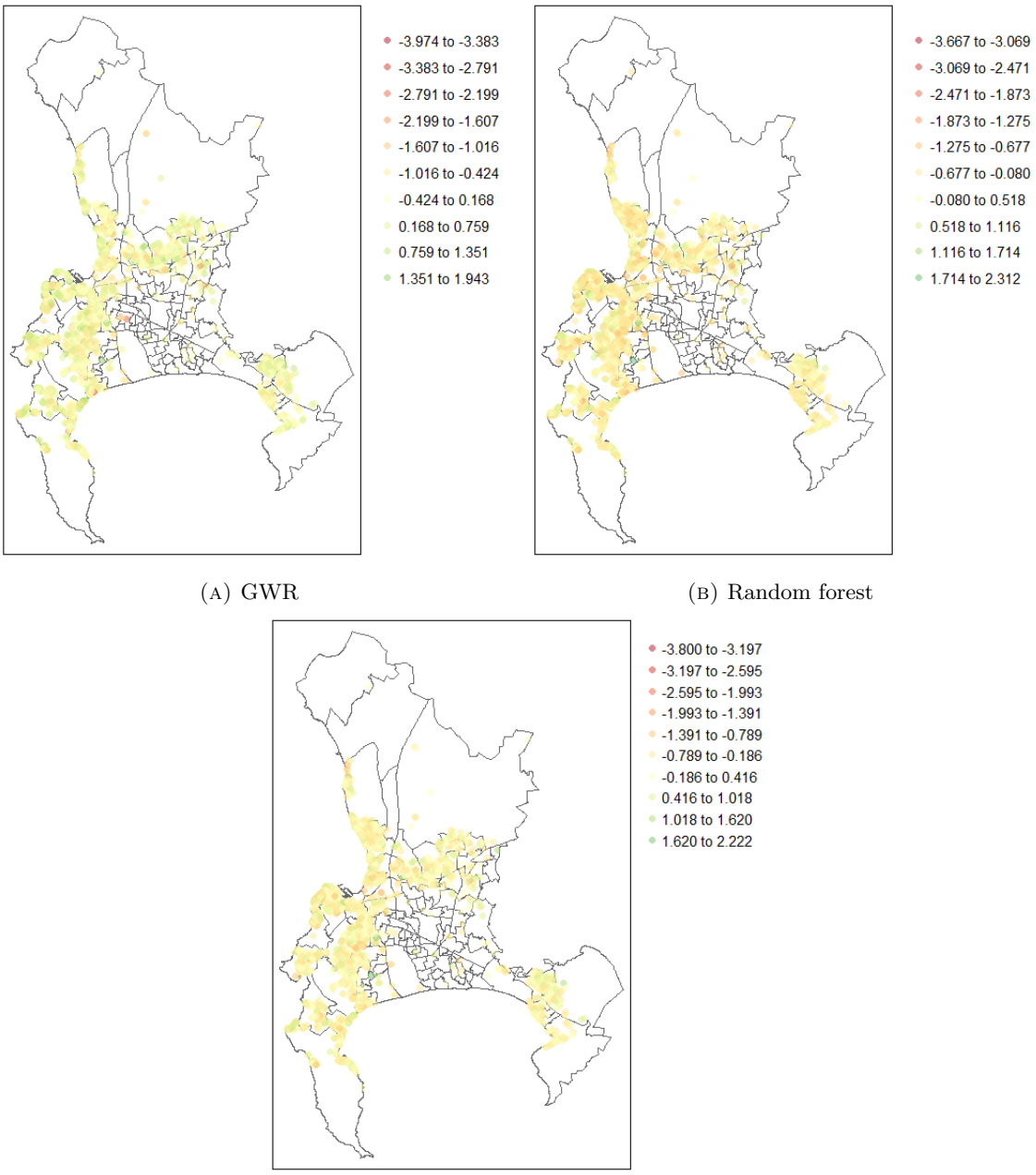
(B) Lag model

(C) Error model

FIGURE 5.23: Mapped OLS, lag and error model testing residuals

error models and GWR, the largest over prediction (as indicated by areas in dark red) are for the area of Manenberg, which is along side Gugulethu and Nyanga. The RF and GBM also over predict in this area, but not to the extent of the the other models. The area most consistently under predicted (represented by dark green) is Cafda Village which was also highlighted as problematic in the training predictions.

Averaging the residuals to the suburb level in Figures 5.27 and 5.28, only the GWR and GBM have green suburbs of under prediction. The South Eastern suburb containing the aforementioned Manenberg, Gugulethu and Mitchell's Plain is by far the suburb with the worst prediction.



(A) GWR

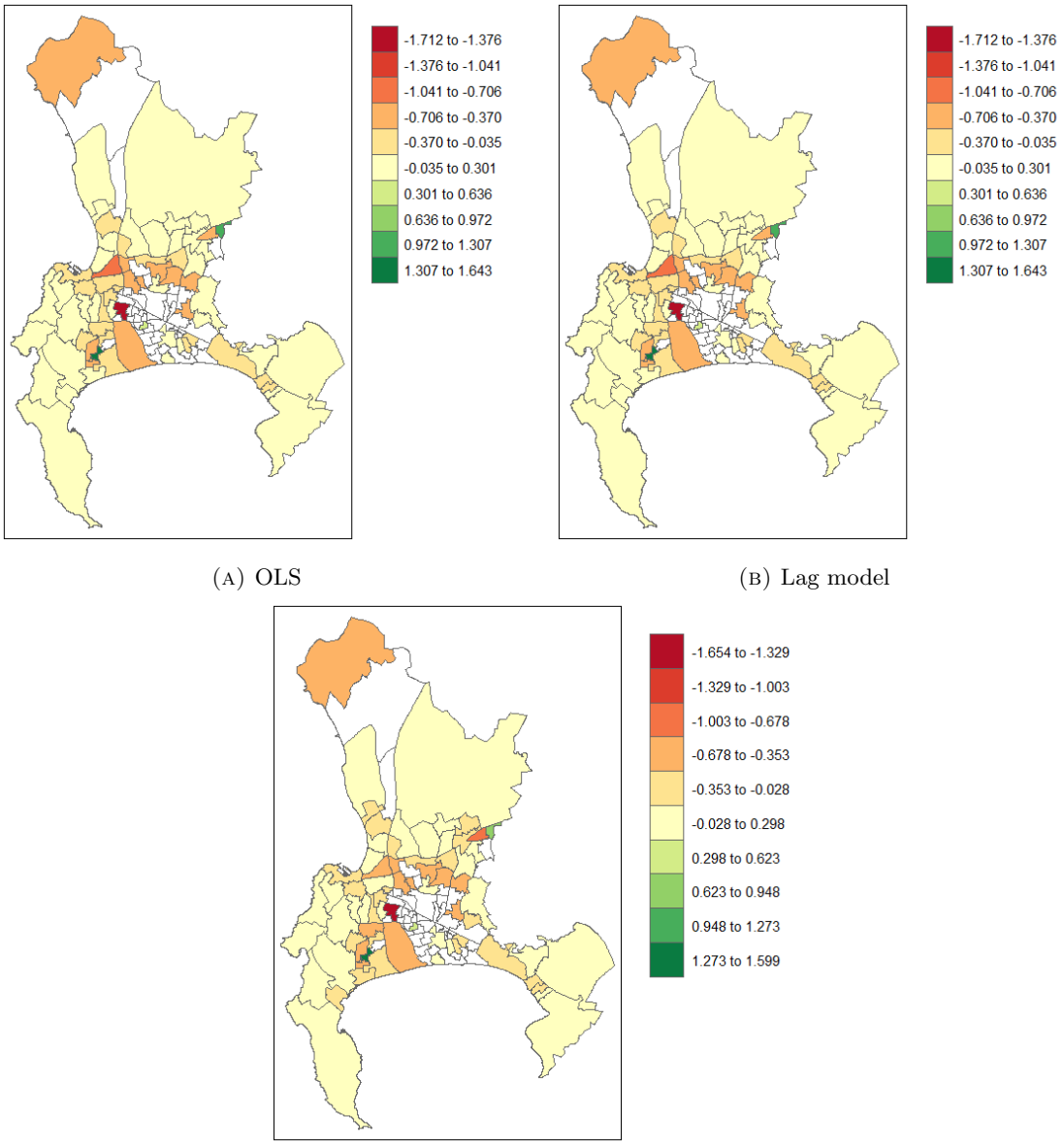
(B) Random forest

(C) GBM

FIGURE 5.24: Mapped GWR, random forest and GBM testing residuals

In this chapter various linear, spatial and machine learning models were applied to the Cape Town Airbnb data set.

Through the use of a Gaussian adaptive kernel, the spatial models were able to address the effect of spatial dependence and slightly improve prediction while doing so, in the case of GWR. The GWR also showed how the importance of some variables such as a pool amenity varies by location, with it adding less to listing price in the City Bowl compared to other suburbs.



(A) OLS

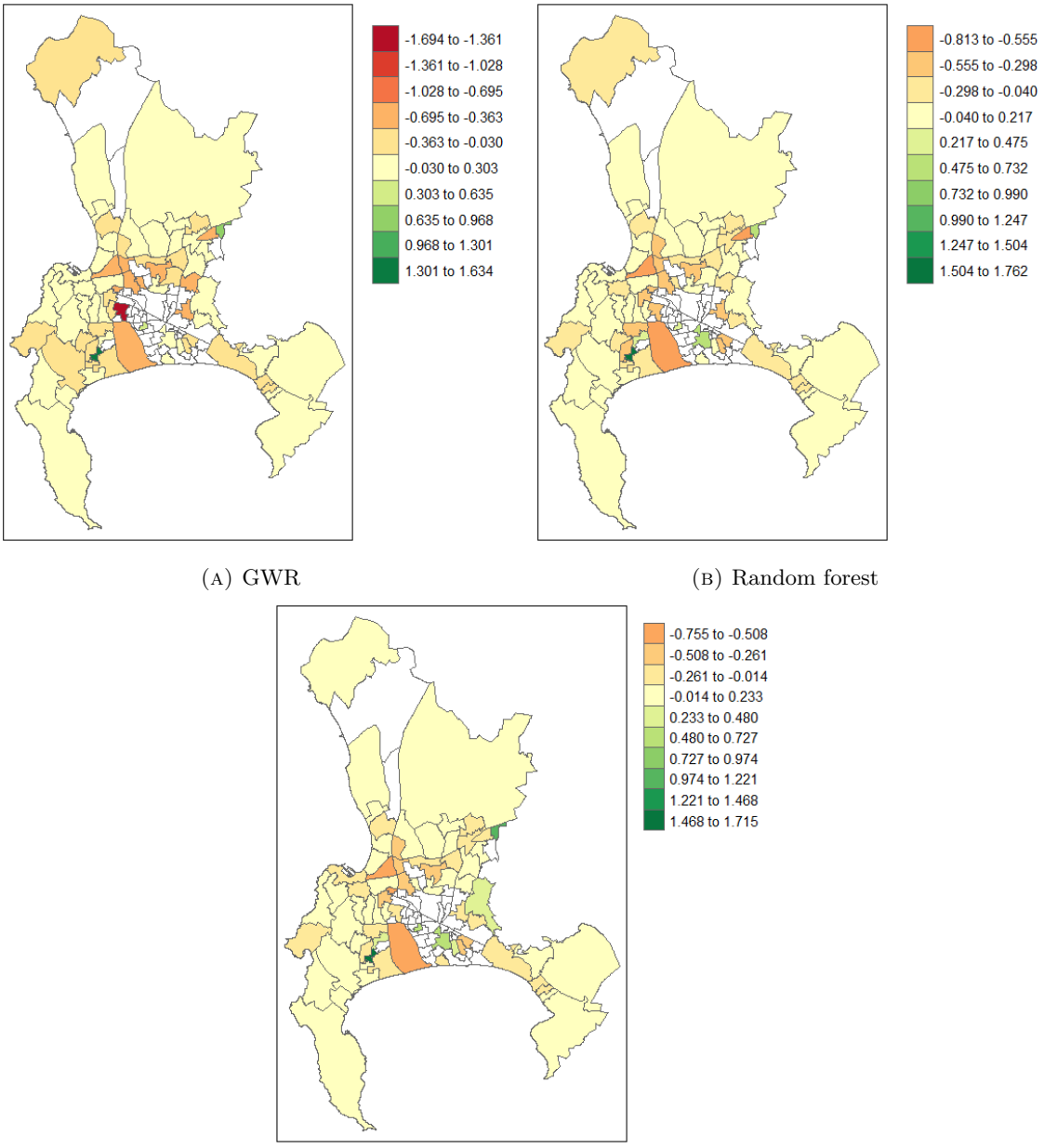
(B) Lag model

(C) Error model

FIGURE 5.25: Mapped OLS, lag and error model mean ward testing residuals

Non linear machine learning models performed the best in terms of prediction. Across all models, basic variables such as accommodates and room type, location variables such as latitude and longitude and amenity variables such as a pool and air conditioning are found to be important price determinants, while Airbnb-specific variables such as superhost or verified identity are less important. For all models, predictive performance was not consistent across price or location. All models under predicted on the most expensive listings and under predicted the cheapest listings and also consistently performed poorly in the South East Suburbs.

The next chapter provides conclusions and suggests possible next steps for further research.

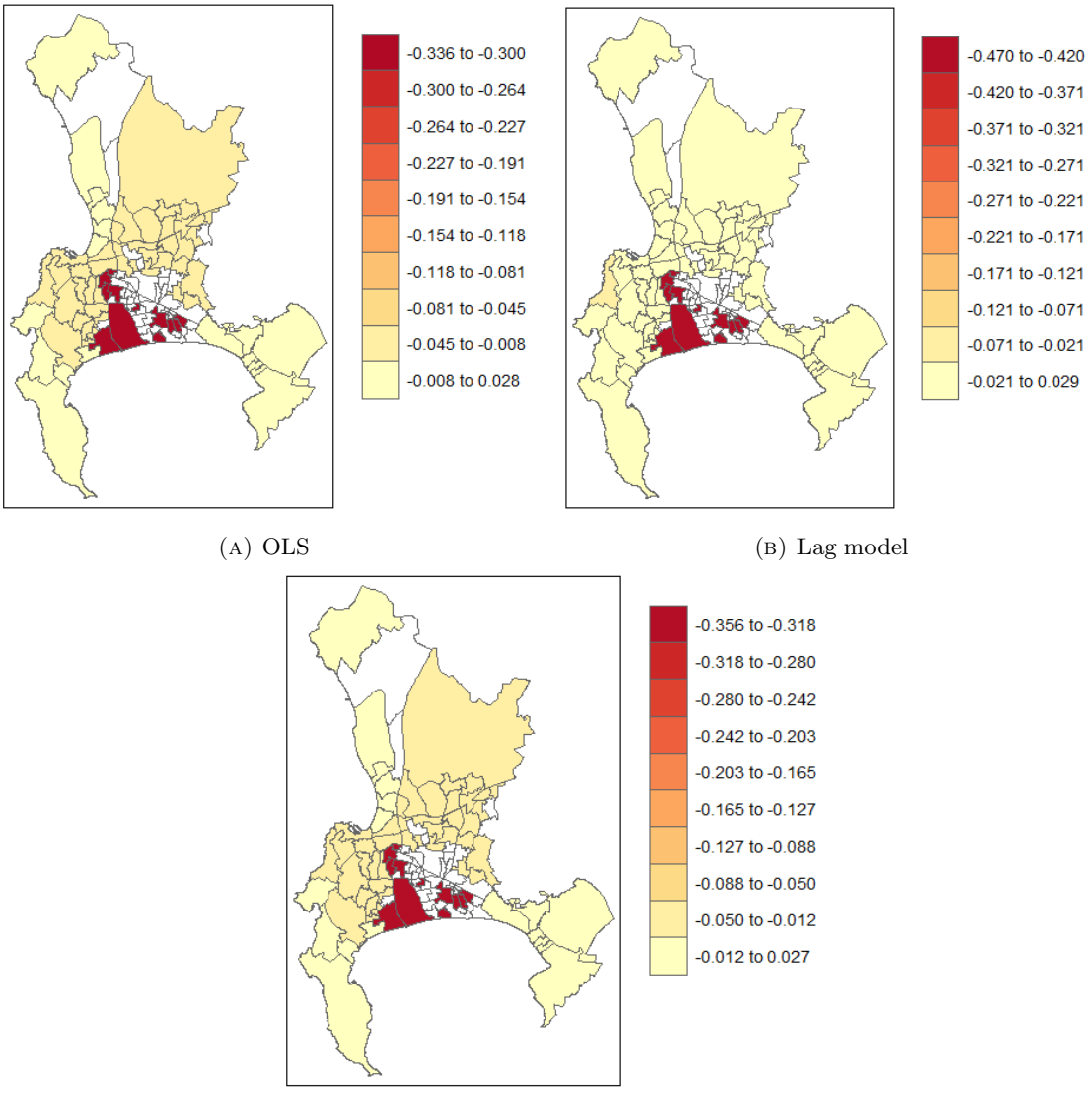


(A) GWR

(B) Random forest

(C) GBM

FIGURE 5.26: Mapped GWR, random forest and GBM mean ward testing residuals

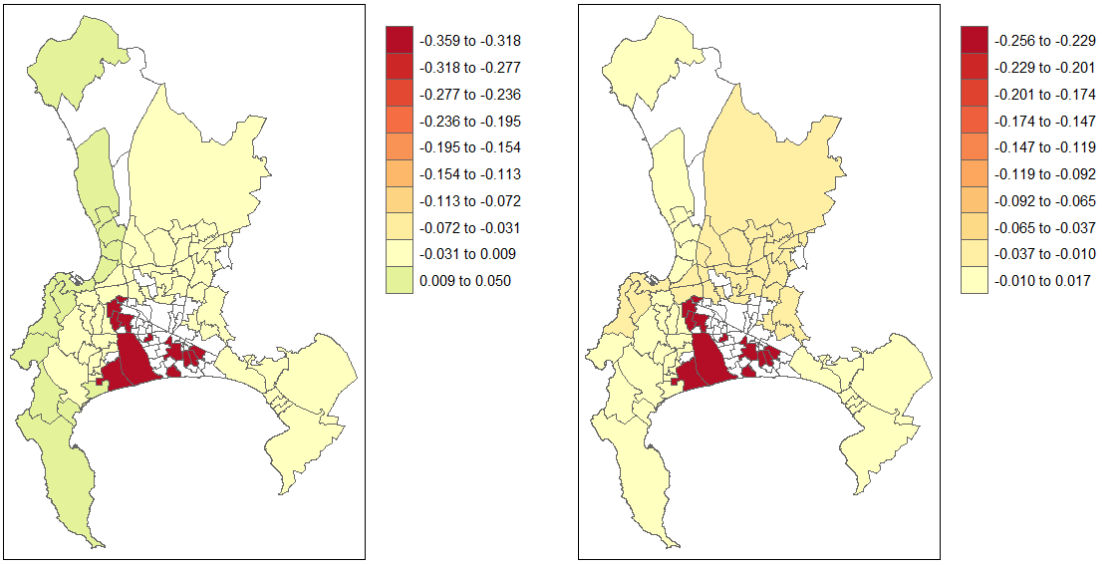


(A) OLS

(B) Lag model

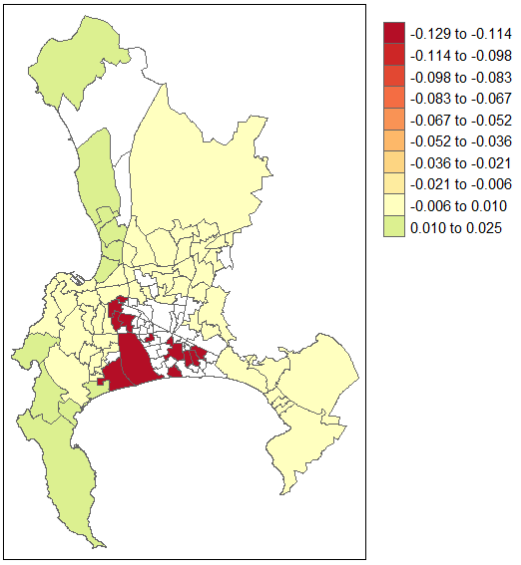
(C) Error model

FIGURE 5.27: Mapped OLS, lag and error model mean suburb testing residuals



(A) GWR

(B) Random forest



(C) GBM

FIGURE 5.28: Mapped GWR, random forest and GBM mean suburb testing residuals

Chapter 6

Conclusion

This chapter concludes the thesis by summarising its objectives and findings as well as suggesting potential improvements and next steps.

Literature and results review

Starting by investigating the literature, the hedonic pricing framework which models property, and vacation rental prices as functions of their parts was described. Factors relating to the physical aspects of the property were found to be the most important price determinants but factors relating to location and amenities were also important. Extending the framework to Airbnb models, various linear, non linear, spatial and non spatial methods have been used to predict price and explore price determinants. The same physical, location and amenity factors were found to be important in the Airbnb market as in the property and vacation rental market. Non linear and spatial models are frequently found to outperform their linear counterparts with GWR being a particular popular spatial method. Adaptive kernel functions were also frequently chosen over fixed kernels.

When the Cape Town Airbnb data set was subjected to linear and spatial models, the adaptive kernel captured more significant spatial correlation compared to a fixed kernel as in the literature. While the spatial lag and spatial error models did not perform better than OLS regression, they were able to address spatial dependence without a reduction in performance. The GWR model, however, was able to improve predictive performance and address spatial dependence simultaneously.

Various machine learning models were explored in the literature with the best performing frequently being neural networks and tree based methods. Only tree based methods were applied to the Cape Town data set and both random forests and GBMs performed the best out of all the implemented models. The GBM model slightly outperformed the random forest, making it the most powerful model that was implemented. Across all models, the physical property, location and amenity variables found to be important in the literature were also important in Cape Town.

Some insights into the spatial variability of the market were also discovered. Airbnbs are largely clustered in the City Bowl and Atlantic Seaboard and these areas tend to be more expensive and are highly contrasted by those listings in areas such as the South Eastern suburbs which are less densely clustered and tend to be less expensive. This contrast was highlighted qualitatively in the literature and can be seen quantitatively in the analysis in this thesis.

Potential further research

Method improvements

Spatial weights matrices have infinite possible specifications and are core to spatial models. Further investigation into an appropriate specification could potentially improve model performance and so should be explored. Alternatively, given that there is spatial variability in the Cape Town data, implementing different models in different areas, as proposed by Yacim and Boshoff (2018) on Cape Town housing data, could also be a possible way to improve local and aggregated global model performance.

While this thesis implemented spatial models and machine learning models independently, the growing field of spatial machine learning which combines the two could be explored. A specific method that could be implemented is a geographic random forest, recently introduced by Georganos et al. (2021).

Data improvements

A data-related improvement that could be made relates to the way in which distance is measured. This thesis used Haversine distance, however, road based measurements or alternate measurements could also be incorporated.

As highlighted in the literature, work has been done in the areas of natural language processing and computer vision to extract host generated and user generated content information that could be fed into predictive models. This data is available for the Cape Town data set and so could be included in future research.

Lastly, another potential analysis improvement could be to conduct a longitudinal analysis. This thesis used cross sectional data, only considering a single point in time. Given the effects of seasonality as well as changing market dynamics acknowledged in the literature, longitudinal analysis could aid in predictive performance and understanding.

Conclusion

This thesis set out to explore the problem of Airbnb price prediction, specifically in Cape Town, South Africa, and in doing so aimed to discover important price determinants as well as spatial effects. It also aimed to bridge the gap in the literature as this Cape Town Airbnb price prediction research, to the best of our knowledge, had not been conducted in South Africa to date. Significant spatial effects were found to exist in the market. Machine learning models were found to best predict price and traditional physical, location and amenity factors were found to be important price determinants.

Appendix A

Appendix A

A.1 Variables initially excluded

TABLE A.1: Initially excluded variables

	Variable Name	Description	Reason for exclusion
1	listing_url	URL to listing	Non informative
2	listing_id	Numeric listing identifier	Non informative
3	scrape_id	Inside Airbnb "Scrape" this was part of	Non informative, all listings have the same scrape ID
4	last_scraped	UTC. The date and time this listing was "scraped".	Non informative, all listings have the same last scraped date
5	thumbnail_url	URL to the Airbnb hosted thumbnail image for the listing	All null
6	medium_url	URL to the Airbnb hosted medium sized image for the listing	All null
7	picture_url	URL to the Airbnb hosted regular sized image for the listing	Non informative
8	xl_picture_url	URL to the Airbnb hosted XL sized image for the listing	All null
9	host_url	The Airbnb page for the host	Non informative
10	host_name	Name of the host. Usually just the first name(s).	Non informative
11	host_since	The date the host/user was created. For hosts that are Airbnb guests this could be the date they registered as a guest.	Non informative
12	host_response_time	The time the host takes to respond to requests	33% are nulls, information included in superhost variable
13	host_response_rate	The rate at which the host responds to requests	33% are nulls, information included in superhost variable
14	host_acceptance_rate	That rate at which a host accepts booking requests.	99% are nulls
15	host_thumbnail_url	URL to host's thumbnail	Non informative
16	host_picture_url	URL to host's picture	Non informative
17	host_neighbourhood	Neighbourhood where the host is situated	Non informative
18	host_verifications	List of channels host is verified through	Non informative
19	host_has_profile_pic	Boolean indicating whether host is verified	99.7% are False
20	host_id	Numeric host identifier	Non informative
21	neighbourhood	The neighbourhood as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.	All nulls
22	neighbourhood_group _cleansed	The neighbourhood group as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.	All nulls
23	city	User provided text field of approximate location	95% are "Cape Town" / "Kaapstad"
24	state	User provided text field of approximate location	97% are "Western Cape" / "WC"
25	market	Text field of approximate Airbnb Market	99% are "Cape Town"
26	smart_location	Text field of approximate location	94% are "Cape Town" / "Kaapstad"
27	country_code	Country symbol	99.9% are "ZA"
28	country	Country	99.9% are "South Africa"
29	bed_type	Type of bed	99.8% are "Real bed"
30	square_feet	Size of listing in square feet	99% are nulls
31	has_availability	Boolean of whether listing has availability	Non informative, all listings are True
32	calendar_last_scraped	UTC. The date and time calendar was last "scraped".	Non informative, all listings have same date
33	requires_license	Whether the listing/jurisdiction requires a license	Non informative, all listings are False
34	license	The licence/permit/registration number	99% are null
35	jurisdiction_names	Legal jurisdiction	All nulls
36	is_business_travel_ready	Boolean indicating whether the listing is business travel ready	Non informative, all listings are False
37	require_guest_profile _picture	Boolean indicating whether listing requires guest to have profile picture	99.8% of listings are False
38	require_guest_phone _verification	Boolean indicating whether listing requires guest to be verified by phone	99.6% of listings are False
39	calculated_host_listings _count_entire_homes	The number of Entire home/apt listings the host has in the current scrape, in the city/region geography	Non informative
40	calculated_hst_listings _count_private_rooms	The number of Private room listings the host has in the current scrape, in the city/region geography	Non informative
41	calculated_host_listings _count_shared_rooms	The number of Shared room listings the host has in the current scrape, in the city/region geography	Non informative
42	calendar_updated	Date listings calendar was last updated	Non informative
43	host_listings_count	The number of listings the host has internationally (per Airbnb calculations)	Non informative
44	host_total_listings_count	The number of listings the host has internationally (per Airbnb calculations)	Non informative and duplicated
45	minimum_minimum_nights	The minimum minimum number of nights the listing was available for in the last year	Non informative
46	maximum_minimum_nights	The maximum minimum number of nights the listing was available for in the last year	Non informative
47	minimum_maximum_nights	The minimum maximum number of nights the listing was available for in the last year	Non informative
48	maximum_maximum_nights	The maximum maximum number of nights the listing was available for in the last year	Non informative
49	name	Text field of host-provided listing name	Requires extensive cleaning
50	summary	Text field of host-provided listing summary description	Requires extensive cleaning
51	space	Text field of host-provided listing space description	Requires extensive cleaning
52	description	Text field of host-provided listing description	Requires extensive cleaning
53	$\text{neighborhood}_{o}verview$	Text field of host-provided neighbourhood description	Requires extensive cleaning
54	notes	Text field of host-provided additional notes	Requires extensive cleaning
55	transit	Text field of host-provided description of transport access	Requires extensive cleaning
56	access	Text field of host-provided description physical access to the listing	Requires extensive cleaning
57	interaction	Text field of host-provided description required interaction with the host / other guests	Requires extensive cleaning
58	host_about	Text field of host-provided description of themselves	Requires extensive cleaning
59	street	Text field of host-provided of approximate location	93% are a variation of "Cape Town, South Africa"
60	experiences_offered	Description of the experiences offered	Non informative, all listings have "none"
61	weekly_price	Weekly price in Rands	94% are null and the variable is derived from the dependent variable
62	monthly_price	Monthly price in Rands	94% are null and the variable is derived from the dependent variable

Bibliography

Adamiak, Czesław. 2022. "Current state and development of Airbnb accommodation offer in 167 countries." *Current Issues in Tourism* 25(19):3131–3149.

Airbnb. 2018. *Airbnb in South Africa: The Positive Impact of Healthy Tourism*. https://press.airbnb.com/wp-content/uploads/sites/4/2018/09/Airbnb-in-South-Africa-Positive-Impact-of-Healthy-Tourism.pdf.

Airbnb. 2020. *The amenities guests want*. https://www.airbnb.co.za/resources/hosting-homes/a/the-amenities-guests-want-25.

Airbnb. 2022. *Q3 2022 Shareholder letter*. https://s26.q4cdn.com/656283129/files/doc_financials/2022/q3/Airbnb_Q3-2022-Shareholder-Letter_Final.pdf.

AirDNA. 2017. *Airdna Data Service*. https://www.airdna.co/services/datafeed.

Anselin, Luc. 2005. "Exploring spatial data with GeoDaTM: a workbook." *Center for spatially integrated social science* .

Anselin, Luc et al. 2001. "Spatial econometrics." *A companion to theoretical econometrics* 310330.

Anwar, Syed Tariq. 2018. "Growing global in the sharing economy: Lessons from Uber and Airbnb." *Global Business and Organizational Excellence* 37(6):59–68.

Bailey, Martin J, Richard F Muth and Hugh O Nourse. 1963. "A regression method for real estate price index construction." *Journal of the American Statistical Association* 58(304):933–942.

Basuroy, Suman, Yongseok Kim and Davide Proserpio. 2020. "Estimating the impact of Airbnb on the local economy: Evidence from the restaurant industry." *Available at SSRN 3516983* .

Bidanset, Paul E and John R Lombard. 2014. "Evaluating spatial model accuracy in mass real estate appraisal: A comparison of geographically weighted regression and the spatial lag model." *Cityscape* 16(3):169–182.

Bitter, Christopher, Gordon F Mulligan and Sandy Dall'erba. 2007. "Incorporating spatial variation in housing attribute prices: a comparison of geographically weighted regression and the spatial expansion method." *Journal of Geographical Systems* 9(1):7–27.

Bivand, Roger S, Edzer J Pebesma, Virgilio Gomez-Rubio and Edzer Jan Pebesma. 2008. *Applied spatial data analysis with R*. Vol. 747248717 Springer.

Brunsdon, Chris, Stewart Fotheringham and Martin Charlton. 1998. "Geographically weighted regression." *Journal of the Royal Statistical Society: Series D (The Statistician)* 47(3):431–443.

Cai, Tiancheng, Kevin Han and Han Wu. 2019. *Melbourne airbnb price prediction*.

Chen, Ching-Fu and Rochelle Rothschild. 2010. "An application of hedonic pricing analysis to the case of hotel rooms in Taipei." *Tourism Economics* 16(3):685–694.

Crisci, Massimiliano, Federico Benassi, Hamidreza Rabiei-Dastjerdi and Gavin McArdle. 2022. "Spatio-temporal variations and contextual factors of the supply of Airbnb in Rome. An initial investigation." *Letters in Spatial and Resource Sciences* 15(2):237–253.

Deboosere, Robbin, Danielle Jane Kerrigan, David Wachsmuth and Ahmed El-Geneidy. 2019. "Location, location and professionalization: a multilevel hedonic analysis of Airbnb listing prices and revenue." *Regional Studies, Regional Science* 6(1):143–156.

Dogru, Tarik and Osman Pekin. 2017. *What do guests value most in Airbnb accommodations? An application of the hedonic pricing approach*.

Dolnicar, Sara and Samira Zare. 2020. "COVID19 and Airbnb–Disrupting the disruptor." *Annals of tourism research* 83:102961.

Espinet, Josep M, Marc Saez, Germa Coenders and M Fluvià. 2003. "Effect on prices of the attributes of holiday hotels: a hedonic prices approach." *Tourism Economics* 9(2):165–177.

ESRI. 2021. *How Neighborhood Summary Statistics works*. https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-neighborhood-summary-statistics-works.htm.

Fischer, Manfred M and Jinfeng Wang. 2011. *Spatial data analysis: models, methods and techniques*. Springer Science & Business Media.

Friedman, Jerome H. 2001. "Greedy function approximation: a gradient boosting machine." *Annals of statistics* pp. 1189–1232.

Georganos, Stefanos, Tais Grippa, Assane Niang Gadiaga, Catherine Linard, Moritz Lennert, Sabine Vanhuysse, Nicholus Mboga, Eléonore Wolff and Stamatis Kalogirou. 2021. "Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling." *Geocarto International* 36(2):121–136.

Gibbs, Chris, Daniel Guttentag, Ulrike Gretzel, Jym Morton and Alasdair Goodwill. 2018. "Pricing in the sharing economy: a hedonic pricing model applied to Airbnb listings." *Journal of Travel & Tourism Marketing* 35(1):46–56.
URL: *https://doi.org/10.1080/10548408.2017.1308292*

Guttentag, Daniel. 2015. "Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector." *Current issues in Tourism* 18(12):1192–1217.

Harrison Jr, David and Daniel L Rubinfeld. 1978. "Hedonic housing prices and the demand for clean air." *Journal of environmental economics and management* 5(1):81–102.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2017. *The elements of statistical learning: Data mining, inference, and prediction*. springer open.

Helbich, Marco, Wolfgang Brunauer, Eric Vaz and Peter Nijkamp. 2014. "Spatial heterogeneity in hedonic house price models: The case of Austria." *Urban Studies* 51(2):390–411.

Henama, Unathi Sonwabile. 2018. "Disruptive entrepreneurship using Airbnb: the South African experience." *African Journal of Hospitality, Tourism and Leisure* 7(1):1–16.

Henama, Unathi Sonwabile and Lebogang Matholwane Mathole. 2022. Teaching South Africans How to Become Successful Hosts on Airbnb: The Case of the Airbnb Africa Academy. In *Entrepreneurship Education in Tourism and Hospitality Management*. IGI Global pp. 129–148.

Hofäcker, Jana and Matthias Gebauer. 2021. "Airbnb in Townships of South Africa: A New Experience of Township Tourism?" *Urban tourism in the global South: South African perspectives* pp. 129–147.

InsideAirbnb. 2023. *InsideAirbnb Data*. http://http://insideairbnb.com/.

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2013. *An introduction to statistical learning*. Vol. 112 Springer.

Kourtit, Karima, Peter Nijkamp, John Östh and Umut Turk. 2022. "Airbnb and COVID-19: SPACE-TIME vulnerability effects in six world-cities." *Tourism Management* 93:104569.

Li, Mingche M and H James Brown. 1980. "Micro-neighborhood externalities and hedonic housing prices." *Land economics* 56(2):125–141.

Li, Yang, Quan Pan, Tao Yang and Lantian Guo. 2016. Reasonable price recommendation on Airbnb using Multi-Scale clustering. In *2016 35th Chinese Control Conference (CCC)*. IEEE pp. 7038–7041.

Ndaguba, EA. 2021. "Economic impediment of COVID-19 lockdown on Airbnb performance in Cape Town neighbourhood." *Academy of Strategic Management Journal* 20(1):1–16.

Oskam, Jeroen and Albert Boswijk. 2016. "Airbnb: the future of networked hospitality businesses." *Journal of tourism futures* .

Pirie, Gordon. 2017. Urban tourism in Cape Town. In *Urban tourism in the developing world*. Routledge pp. 223–244.

Rosen, Sherwin. 1974. "Hedonic prices and implicit markets: product differentiation in pure competition." *Journal of political economy* 82(1):34–55.

Short, Ryan, Ceri Scott, Kirra Evans, Kim Adonis and Mark Schoeman. 2021. *The foundations of inclusive tourism: The contribution of Airbnb to inclusive growth in South Africa*. https://genesis.imgix.net/uploads/files/Genesis-Analytics-Airbnb-The-foundations-of-inclusive-tourism-13-Sept-2021-Final-report.pdf.

Soler, Ismael P, German Gemar, Marisol B Correia and Francisco Serra. 2019. "Algarve hotel price determinants: A hedonic pricing model." *Tourism Management* 70:311–321.

Subroyen, Juanita, Marita Turpin and Alta de Waal. 2021. "Empowering Peer-to-Peer Platform Role-players by Means of Topic Modelling: A Case Study of Airbnb in Cape Town, South Africa.".

Sun, Shijie, Xingjian Wang and Mingxing Hu. 2022. "Spatial distribution of Airbnb and its influencing factors: A case study of Suzhou, China." *Applied Geography* 139:102641.

Tang, Emily and Kunal Sangani. 2015. "Neighborhood and price prediction for San Francisco Airbnb listings." *Departments of Computer science, Psychology, economics–Stanford University* .

Tobler, Waldo R. 1970. "A computer movie simulating urban growth in the Detroit region." *Economic geography* 46(sup1):234–240.

Tripadvisor. 2021. *Top Attractions in Cape Town.* https://www.tripadvisor.co.za/Attractions-g312659-Activities-oa0-Cape_Town_Central_Western_Cape.html.

Wegmann, Jake and Junfeng Jiao. 2017. "Taming Airbnb: Toward guiding principles for local regulation of urban vacation rentals based on empirical results from five US cities." *Land use policy* 69:494–501.

Xu, Feifei, Mingxing Hu, Liqing La, Jialing Wang and Chao Huang. 2020. "The influence of neighbourhood environment on Airbnb: a geographically weighed regression analysis." *Tourism Geographies* 22(1):192–209.

Yacim, Joseph Awoamim and Douw Gert Brand Boshoff. 2018. "Impact of artificial neural networks training algorithms on accurate prediction of property values." *Journal of Real Estate Research* 40(3):375–418.

Yacim, Joseph Awoamim and Douw Gert Brand Boshoff. 2019. "A Comparison of bandwidth and kernel function selection in geographically weighted regression for house valuation." *Architecture* 10(1).

Yacim, Joseph Awoamim and Douw Gert PhD Boshoff. 2016. Comparison Of Mass Appraisal Models For Effective Prediction Of Property Values. Technical report African Real Estate Society (AfRES).

Yang, Siqi. 2021. Learning-based Airbnb Price Prediction Model. In *2021 2nd International Conference on E-Commerce and Internet Technology (ECIT)*. pp. 283–288.

Zervas, Georgios, Davide Proserpio and John W Byers. 2017. "The rise of the sharing economy: Estimating the impact of Airbnb on the hotel industry." *Journal of marketing research* 54(5):687–705.

Zhang, Zhihua, Rachel JC Chen, Lee D Han and Lu Yang. 2017. "Key factors affecting the price of Airbnb listings: A geographically weighted approach." *Sustainability* 9(9):1635.