

# **Appraising South African Residential Property and Measuring Price Developments**



**UNIVERSITY OF  
KWAZULU-NATAL**

---

**INYUVESI  
YAKWAZULU-NATALI**

Dane Bax

December 2022

# **Appraising South African Residential Property and Measuring Price Developments**

by

Dane Bax

A thesis submitted to the

University of KwaZulu-Natal

in fulfilment of the requirements for the degree

of

Doctor of Philosophy

in

Statistics

Thesis supervisors: Professor Temesgen Zewotir

Professor Delia North



UNIVERSITY OF KWAZULU-NATAL

School of Mathematics, Statistics and Computer Science

WESTVILLE CAMPUS, DURBAN, SOUTH AFRICA

## Declaration – Plagiarism

I, Dane Bax, declare that


1. The research reported in this thesis, except where otherwise indicated, is my original research.
2. This thesis has not been submitted for any degree or examination at any other university.
3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.
4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then
  - a) their words have been rewritten but the general information attributed to them has been referenced, or
  - b) where their exact words have been used, then their writing has been placed in italics and referenced.
5. This thesis does not contain text, graphics or tables copied and pasted from the internet, unless specifically acknowledged, and the source being detailed in the thesis and in the reference sections.

### **Student:**

Name: Mr Dane Bax

Date: 2022-12-05

Signature:

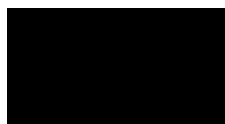


### **Supervisor:**

Name: Professor Temesgen Zewotir

Date: 2022-12-15

Signature:

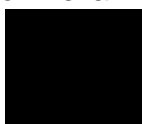


### **Co-supervisor:**

Name: Professor Delia North

Date: 2022-12-15

Signature:



## **Disclaimer**

This document describes work undertaken as a PhD programme of study at the University of KwaZulu-Natal (UKZN). All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institution.

## **Abstract**

Housing wealth is well established as one of the most important sources of wealth for households and investors. However, owning a home is a fundamental human need, making monitoring residential property prices a social endeavour as well as an economic one, especially under times of economic uncertainty. Residential property prices also have a direct effect on the macroeconomy because of how they influence wealth effects where increased consumption by households is experienced through gains in households balance sheets due to increased equity. Collecting correct and adequate data is vitally important in analysing property market movements and developments, particularly given globalization, and the interlinked nature of financial markets. Although measuring residential property price developments is an important economic and social activity, matching properties over time is extremely difficult because the sale of homes is typically infrequent, characteristics vary, and homes are uniquely located in space. This thesis focuses on appraising several residential property types located throughout South Africa from January 2013 to August 2017, investigating different modelling approaches with the aim of developing a residential property price index. Various methods exist to create residential property price indices, however, hedonic models have proven useful as a quality adjusted approach where pure price changes are measured and not simply changes in the composition of samples over time. Before fitting any models to appraise homes, an autoencoder was built to detect anomalous data, due to human error at the data entry stage. The autoencoder identified improbable data resulting in a final data set of 415 200 records, once duplicate records were identified and removed. This study first investigated generalised linear models as a candidate approach to appraise homes in South Africa which showed possible alternatives to the ubiquitous log linear model. Relaxing functional form assumptions and considering the nested locational structure of homes, hierarchical generalised linear models were considered as the next candidate method. Partitioning around the medoids was applied to find additional spatial groupings which were treated as random effects along with the suburb. The findings showed that the marginal utility of structural attributes was non-linear and smooth functions of covariates were an appropriate treatment. Furthermore, the use of random effects helped account for the spatial heterogeneity of homes through partial pooling. Finally, machine learning algorithms were investigated because of minimal assumptions about

the data generating process and the possibility of complex non-linear and interaction effects. Random forests, gradient boosted machines and neural networks were adopted to fit these appraisal functions. The gradient boosted machines had the best goodness of fit, showing non-linear relationships between the structural characteristics of homes and listing prices. Partial dependence plots were able to quantify the marginal utility over the distributions of different structural characteristics. The results show that larger sized homes do not necessarily yield a premium and a diminished return is evident, similar to the results of the hierarchical generalised additive models. The variable importance plots showed that location was the most important predictor followed by the number of bathrooms and the size of a home. The gradient boosted machines achieved the lowest out of sample error and were used to develop the residential property price index. A chained, dual imputation Fisher index was applied to the gradient boosted machines showing nominal and real price developments at a country and provincial level. The chained, dual imputation Fisher index provided less noisy estimates than a simple median mix adjusted index. Although listing prices were used and not transacted prices, the trend was similar to the ABSA Global Property Guide. In order to make this research useful to property market participants, a web application was developed to show how the proposed methodology can be democratised by property portals and real estate agencies. The Listing Price Index Calculator was created to easily communicate the results through a front-end interface, showing how property portals and real estate agencies can leverage their data to aid sellers in determining listing prices to go to market with, help buyers obtain an average estimate of the home they wish to purchase and guide property market participants on price developments.

**KEYWORDS:** Generalised Linear Models; Hedonic Regression; Hierarchical Generalised Additive Models; Machine Learning; Real Estate Economics; Spatial Modelling

## **Acknowledgements**

The culmination of this research has been challenging and I would like to thank my wife, Pamela Bax, and my family especially my mother, Megan Bax, for their unwavering support. I would also like to thank my supervisors, Professor Zewotir and Professor North, for their endearing guidance and support.

## List of Publications

Bax, D., North, D., & Zewotir, T. (2019). A gamma generalised linear model as an alternative to log linear real estate price functions. *Journal of Economic and Financial Sciences*, 12(1). <https://doi.org/10.4102/jef.v12i1.476>

Bax, D., North, D., & Zewotir, T. (2020). A machine learning web application to estimate listing prices of South African homes. *Journal of African Real Estate Research*, 4(2), 1–23. <https://doi.org/10.15641/jarer.v4i2.802>

Bax, D., North, D., & Zewotir, T. (2021). Appraising residential property using hierarchical generalised additive models. *Journal of Property Research*, 38(1), 1–15. <https://doi.org/10.1080/09599916.2021.1888774>



## Contents

|  |      |
|--|------|
| Declaration – Plagiarism.....                                  | i    |
| Disclaimer.....  | ii   |
| Abstract.....  | iii  |
| Acknowledgements.....  | v    |
| List of Publications.....                                      | vi   |
| Contents.....  | vii  |
| List of Figures.....   | viii |
| List of Tables.....  | ix   |
| Abbreviations.....   | x    |
| Chapter One: Introduction.....                                 | 1    |
| 1.1    Review of Residential Property Pricing Methodology..... | 2    |
| 1.1.1    Mean or Median Indices.....                           | 3    |
| 1.1.2    Repeat Sales.....                                     | 4    |
| 1.1.3    Hedonic Pricing.....                                  | 5    |
| 1.1.4    South African Methodologies.....                      | 9    |
| 1.2    Research Gap and Objectives.....                        | 10   |
| 1.3    The Significance and Contribution of the Study.....     | 11   |
| 1.4    Thesis Outline.....                                     | 12   |
| Chapter Two: The Data and Exploratory Data Analysis.....       | 13   |
| 2.1    Description of the Data.....                            | 13   |
| 2.2    Summary Statistics.....                                 | 17   |
| 2.3    Spatial Nature of the Data.....                         | 20   |
| 2.4    Summary.....  | 22   |
| Chapter Three: Generalised Linear Models.....                  | 24   |
| 3.1    Model Description and Motivation.....                   | 24   |
| 3.2    Results and Discussion.....                             | 27   |
| 3.3    Summary.....  | 37   |

|   |    |
|---|----|
| Chapter Four: Hierarchical Generalised Additive Models..... | 38 |
| 4.1    Model Description and Motivation.....                | 38 |
| 4.2    Partitioning Around the Mediods.....                 | 42 |
| 4.3    Results and Discussion.....                          | 43 |
| 4.4    Summary.....   | 51 |
| Chapter Five: Machine Learning Methods.....                 | 52 |
| 5.1    Model Description and Motivation.....                | 52 |
| 5.1.1    Gradient Boosted Machines.....                     | 53 |
| 5.1.2    Random Forests.....                                | 54 |
| 5.1.3    Artificial Neural Networks.....                    | 55 |
| 5.2    Model Training and Validation.....                   | 57 |
| 5.3    Results and Discussion.....                          | 60 |
| 5.4    Summary.....   | 67 |
| Chapter Six: Listing Price Index and Web Application.....   | 68 |
| 6.1    Hedonic Model Selection.....                         | 68 |
| 6.2    Index Number Theory.....                             | 69 |
| 6.3    Project Process.....                                 | 71 |
| 6.4    Results and Discussion.....                          | 74 |
| 6.5    Listing Price Index Web Application.....             | 78 |
| 6.6    Summary.....   | 81 |
| Chapter Seven: Discussion and Conclusion.....               | 82 |
| 7.1    Conclusion.....                                      | 82 |
| 7.2    Limitations and Future Work.....                     | 85 |
| References.....   | 86 |
| Appendix.....   | 95 |

## List of Figures

|   |    |
|---|----|
| Figure 2.1. Autoencoder diagram.....  | 15 |
| Figure 2.2. Autoencoder reconstruction error.....                                 | 16 |
| Figure 2.3: Property type listing price histograms.....                           | 18 |
| Figure 2.4: Parallel plot.....  | 19 |
| Figure 2.5: South African municipalities.....                                     | 21 |
| Figure 3.1: GLM fitted versus residual plots..                                    | 29 |
| Figure 3.2: GLM Q-Q plots.....  | 30 |
| Figure 3.3: GLM Coefficients..  | 33 |
| Figure 3.4: Gamma model residuals against transformed size covariate.....         | 34 |
| Figure 3.5: Gamma model variogram plots. ....                                     | 36 |
| Figure 4.1: PAM Spatial clusters.....   | 44 |
| Figure 4.2: PAM silhouette plots.....   | 45 |
| Figure 4.3: HGAM variogram plots.....   | 47 |
| Figure 4.4: HGAM diagnostic residual plots.....                                   | 49 |
| Figure 4.5 Smooth covariate plots.....  | 50 |
| Figure 5.1 Example of feed forward artificial neural network architecture.....    | 56 |
| Figure 5.2 Example of feed forward artificial neural network process diagram..... | 57 |
| Figure 5.3: GBM Variogram plots.....  | 63 |
| Figure 5.4: Partial dependence calibration plots.....                             | 65 |
| Figure 5.5 Variable importance plots.....   | 66 |
| Figure 6.1 Deployment architecture.....   | 72 |
| Figure 6.2: CRISP-DM framework.....   | 73 |
| Figure 6.3 Provincial listing price growth 2013-2017.....                         | 78 |
| Figure 6.4 LPIC example one.....  | 79 |
| Figure 6.5 LPIC example two.....  | 80 |

## List of Tables

|   |    |
|---|----|
| Table 2.1. Description of the data.....                               | 13 |
| Table 2.2. Clean data summary statistics..                            | 14 |
| Table 2.3. Final data summary statistics.....                         | 16 |
| Table 2.4: Correlation matrix.....                                    | 17 |
| Table 2.5: Yearly property type frequency table.....                  | 20 |
| Table 3.1: Comparison of GLM summaries.....                           | 28 |
| Table 3.2: Gamma model results summary.....                           | 32 |
| Table 3.3: Analysis of deviance.....                                  | 35 |
| Table 3.4: GLM permutation test for Moran's I.....                    | 37 |
| Table 4.1: Results and comparison of HGAM's.....                      | 46 |
| Table 4.2: HGAM permutation test for Moran's I..                      | 48 |
| Table 5.1: 5-Fold cross validation structure.....                     | 59 |
| Table 5.2: Cross validation model summaries.....                      | 61 |
| Table 5.3: Combined holdout goodness of fit summaries.....            | 62 |
| Table 5.4: GBM permutation test for Moran's I.....                    | 62 |
| Table 5.5: GBM model summaries.....                                   | 64 |
| Table 6.1: Hedonic model out of sample errors.....                    | 69 |
| Table 6.2 Median mix adjusted index results.....                      | 74 |
| Table 6.3: Nominal South African LPI results.....                     | 74 |
| Table 6.4: Annual inflation rate comparisons.....                     | 75 |
| Table 6.5: Inflation adjusted South African LPI results.....          | 75 |
| Table 6.6: ABSA global property guide annual house price changes..... | 76 |
| Table 6.7: Provincial LPI results.....                                | 76 |

## Abbreviations

|      |   |
|------|---|
| AIC  | Akaike information criterion            |
| GCV  | Generalised Cross Validation            |
| GBM  | Gradient Boosted Machine                |
| GLM  | Generalised Linear Model                |
| HGAM | Hierarchical Generalised Additive Model |
| LPIC | Listing Price Index Calculator          |
| OLS  | Ordinary Least Squares                  |
| PDP  | Partial Dependence Plot                 |
| REML | Restricted Maximum Likelihood           |
| RMSE | Root Mean Squared Error                 |
| RPPI | Residential Property Price Index        |
| VIP  | Variable Importance Plot                |

# Chapter One

## Introduction

Residential property is an important determinant in the cost of living and is perceived as a fundamental source of wealth. The collective value of residential property is closely tracked by households, investors, banks, governments and other economic property establishments (Hill, 2013). Policy makers use property price trends as a metric to gauge financial stability in property markets, in conjunction to assessing conditions of credit markets (de Haan and Diewert, 2011). Periods of economic expansion often correspond with burgeoning house prices. Goodhart and Hofmann (2006) conducted a study where a strong correlation between house prices and increased economic activity was found for 16 industrialised economies. Empirical evidence shows a causal relationship between developments in housing prices and the real economy where residential property prices have the propensity to drive macroeconomic developments as a leading indicator (Brunauer *et al*, 2012).

Residential property plays a major role in global economies, Girouard and Blöndal (2001) identified housing wealth as a factor in global business cycles. Their research showed that consumption spending and household borrowing increased based on wealth effects off the back of increased house prices where households were influenced to spend based on an increase in their net asset value. This wealth effect ameliorates household liquidity constraints which can influence consumption spending and aggregate demand. Ando and Modigliani (1963) state that consumers distribute increases in expected wealth over time and the marginal propensity to consume from that wealth, has a similar magnitude. Although Case *et al* (2005) found that wealth effects from the housing market does not have the same degree of influence as wealth effects from the stock market. Housing wealth most often serves as collateral and increases in home prices can ameliorate loan repayments resulting in positive spill overs and hike consumption spending. Additional spill over effects includes booms in housing investment, often resulting in increased employment (Hill, 2013). Higher house prices tend to stimulate construction activity and bolster employment and income for workers in the housing market (de Haan and Diewert, 2011). Increased sales in existing housing stock are experienced under burgeoning house prices, which leads to augmented tax revenues from property transfer taxes (Hill, 2013). Regulation

plays an important role in home price developments, manifesting in strengthening or dampening the effect of the mortgage market through the availability of loans to households and investors.

Residential property prices are an input into measuring the aggregate wealth of a country, South Africa capitalises the value of residential property on the household balance sheet in the set of national accounts. Macroeconomists along with central banks measure residential property price inflation to identify bubbles and understand how house price inflation relates to recessions (Silver, 2016). The basis of economics is to maximize the value of an objective through effective decision making (Greene, 2003). The purchase or sale of a home can be characterized as a significant financial transaction for a household and changes in house prices are likely to influence timing and affordability decisions (Els and Von Fintel, 2010). Residential property price indexes (RPPI's) make inter-area comparisons possible which augment household and investor purchasing decisions (de Haan and Diewert, 2011). Efficient housing markets should reflect full information of the market where this information can be analysed and used as a macroeconomic indicator, for use by credit lenders and policy makers (Diewert, 2007).

Literature supports the view that measuring residential property price developments is of significant economic importance. However, several factors complicate the construction of RPPI's. Constructing price indices for homogenous goods is relatively straight forward, however, homes are uniquely located in space with varying sets of characteristics, making them heterogenous and much more difficult to measure temporal price changes. Moreover, properties maybe subject to depreciation and renovations.

### **1.1. Review of Residential Property Pricing Methodology**

Residential property markets constitute many differentiated or heterogenous properties comprising of a myriad of different prices. Market forces are responsible for setting the different prices of residential properties which is contingent on each individual property's set of characteristics. Generally, the market will settle on a set of prices for the various assortment of residential properties that will clear the market through the reconciliation of supply and demand (Day, 2003). Due to the heterogenous

nature of residential property, property valuations can be a complex process. The valuation of a property is a function of how potential buyers perceive its worth in comparison to similar properties in the market. In South Africa, estate agencies use a technique known as Comparative Market Analysis (CMA) to derive property valuations to mandate properties to market (Private Property, 2019). Estate agencies compile CMA's by comparing similar properties that have sold recently in the same area of interest, accounting for various sources of information such as home characteristics, time on market, important neighbourhood factors and the initial listing price. CMA's assist in determining how much a prospective buyer will pay for a property, taking cognisance of local market conditions and supply and demand forces.

Measuring pure price changes over time requires residential property prices to be adjusted for quality changes. Measuring like-with-like properties in successive periods is critical to price index measurements, known as quality adjusted indices (Silver, 2016). Different quality levels such as physical attributes and location must be considered and be comparable over time with care taken to avoid misspecification, particularly omitted variables. Various methods to measure residential property price developments and control for changes in quality are outlined in literature.

#### 1.1.1 Mean or Median Indices

The simplest measure of residential property price changes is based some measure of central tendency, typically the median as home prices are positively skewed (de Haan and Diewert, 2011). These are simple to construct as no data on housing characteristics is required. However, these indices typically produce noisy estimates of price change as changes in the mix of properties from period to period is common. Furthermore, simple mean or median price indices will be subject to bias when the composition of homes changes over time. The number of homes transacted does not necessarily represent the total housing stock resulting in bias which is also systemic to other property price index methods. Developing a property price index using listing price data certainly provides more homes in the sample design which is the focus of this study. A general technique for reducing sample selection bias is post stratification of a sample, also known a mix adjustment. Mix adjustment is the simplest way to control for changes in the composition of homes and facilitates the creation of price indices for different housing segments. Mix adjustment involves separating the sample



of homes into homogeneous strata. Firstly, a measure of the change in the mean or median for each stratum is calculated, the aggregate mix adjusted residential property price index is then constructed as a weighted average of indices for each stratum. For  $M$  strata, the aggregate mix adjusted index is given by:

$$P^{0t} = \sum_{i=1}^M w_m^0 P_m^{0t}, \quad (1.1)$$

where  $P_m^{0t}$  is the index for stratum  $m$ , comparing the mean or median price in the comparison period  $t$  with the mean or median price in an earlier base period 0.  $w_m^0$  relates to the share of properties in each stratum, denoting the weights of stratum  $m$ . Using the stock value shares of strata is suitable to track the price change of housing stock (Silver, 2016). The choice of granularity for creating the strata is important because the effectiveness of the stratification will depend upon the stratification variables. Very detailed strata based on both physical and locational attributes increases homogeneity and reduces the quality mix problem, however, a trade off exists where increasing the number of strata reduces the average number of observations per stratum. The detail of the stratification scheme should be constructed based on the availability of strata defining characteristics for all sample data.

### 1.1.2 Repeat Sales

A well-known set of residential property price indices are the Standard and Poor's Case-Shiller home price indices for the United States which use the repeat sales methodology. The repeat sales methodology compares properties that are transacted more than once over the sample period. Standard repeat sales indices require data on the price, sales date, and address of sold properties, pooling the data over all periods in the observation period. By only examining properties that have transacted repeatedly over a sample period, quality control is theoretically achieved, however, this approach lends itself towards smaller sample sizes (Bourassa, Hoesli and Sun, 2007). Another drawback is the inability to deal with depreciation and renovations which is true for simple mean or median mix adjustment indices too (de Haan and Diewert, 2011). The standard repeat sales methodology equation is given by:

$$\ln(P_n^t/P_n^s) = \sum_{i=0}^T \gamma^i D_n^i + \mu_n^t,$$

where the left-hand side of the equation denotes the change in price, determined by the difference in transaction times denoted by  $D_n^t$ , a dummy variable indicating the period the resale occurs with  $\mu_n^t$  denoting the error term. The repeat sales index from period 0 to period  $t$  is derived by exponentiating the coefficients  $\hat{\gamma}^t$ . Case and Schiller (1987, 1989) propose Weighted Least Squares to correct for heteroskedasticity which manifests when two transaction dates are further apart (Shimizu, Nishimura and Watanabe, 2016). Two major restrictions to using this approach in this study include incomplete address data and an inability to know if properties are repeat listings over periods. Furthermore, this approach is inefficient in the sense that property characteristics are not included in the modelling. For these reasons, an alternative technique is investigated.

### 1.1.3 Hedonic Pricing

Hedonic pricing mathematically models residential property prices as a function of structural and location characteristics using regression models (Lyons, 2015). In an extensive study of real estate literature, Hill (2013) found hedonic models have been favoured as a quality adjusted approach over other methods. Hedonic pricing is pervasive in the construction of residential property price indices where regression models are developed in the price estimation procedure (de Haan and Diewert, 2011: Jiang *et al*, 2015). Hedonic regression has been found useful as a quality adjusted methodology where pure price changes are measured and not simply changes in the composition of samples in different periods (Shimizu *et al*, 2010). Critical to the success of developing RRPI's is accounting for changes in the quality-mix of homes over the sample period which translates into measuring price pure changes. A rise in average home prices over time may be attributable to a change in the quality-mix and not pure price change (Hill, 2013). For example, if more 4-bedroom, larger sized homes in a more expensive suburb are transacted in the current period compared to the previous period, bias would ensue where the change in average prices would tend upwards, unless some degree of quality-mix control is implemented (Silver 2016). Hedonic models are an effective way to tackle the quality-mix problem because hedonic models unbundle home prices into implicit prices providing estimated marginal values to home characteristics. Rosen (1974) states that the price of a

product can be measured in terms of its utility bearing characteristics, mathematically given by:

$$P_j = P(\mathbf{Z}_j) = P(Z_{j1}, Z_{j2}, \dots, Z_{jk}),$$

where in the case of this study,  $P_j$  is the price of property  $j$  which is a function of its set of  $\mathbf{Z}_j$  characteristics. This function relates prices of goods to their respective characteristics, specifically heterogeneous goods (Day, 2003). Goodman (1978) suggests a general form that does not impose uniformity of coefficients over space and time given by:

$$P_{nt} = f_{nt}(Z_{1nt}, \dots, Z_{1nt}), \quad (1.2)$$

which refers to the  $i$ th characteristic in the  $n$ th submarket at time  $t$ . This form produces hedonic functions for separate markets at different periods. de Haan and Erwin (2011) outline Ordinary Least Squares (OLS) as a prominent hedonic pricing technique, which estimates the marginal contribution of each property's set of attributes. The implicit price for characteristic  $Z_i$  of property  $j$  is calculated by taking the partial derivative. OLS can take the form of the full linear model (1.3) or the logarithmic linear model (1.4) given by:

$$P_n^t = \beta_0^t + \sum_{k=0}^n \beta_k^t Z_{nk}^t + \epsilon_n^t, \quad (1.3)$$

$$\ln P_n^t = \beta_0^t + \sum_{k=0}^n \beta_k^t Z_{nk}^t + \epsilon_n^t. \quad (1.4)$$

The assumption that the price  $P_n^t$  of property  $n$  in period  $t$  is a function of a fixed number of parameters plus  $\epsilon_n^t$  random error.  $\beta_0^t$  and  $\beta_k^t$  are the intercept and characteristic coefficients to be estimated. Two main approaches exist using this technique. Firstly, the time dummy approach where a single OLS is run on the pooled cross-sectional data. In this case the characteristic coefficients are fixed over time with a time coefficient that varies between periods (de Haan and Erwin, 2011). A disadvantage of this approach is the problem of temporal fixity which means adding new periods to the data will result in changes to the coefficient estimates, resulting in revisions (Hill 2013).

The second main approach is the characteristics approach where separate OLS regressions are run for the respective periods allowing the characteristic coefficients to vary period to period which is far more reasonable than the fixed time dummy approach (de Haan and Erwin, 2011). The characteristics approach is similar to the form proposed by Goodman (1978) in equation (1.2) as the coefficients of the characteristics are allowed to vary over time. Goodman (1978) states that although no theoretical link exists between the functional notation and specified functional form, log-linear models are often relied upon in hedonic studies. This approach is supported in literature as it often results in residuals with constant variance. The characteristics method deals with temporal fixity and is more popular for computing residential price indexes used by statistical agencies and government bureaus (Hill, 2013).

Day (2003) developed a hedonic house price function for Glasgow, Scotland where the natural logarithm of selling price was regressed on physical and locational property attributes. The research showed that along with the physical attributes of the properties, spatial effects were statistically significant. Bourassa et al (2007) also applied a log linear hedonic model to the Auckland, New Zealand housing market where similarly, spatial and physical attributes were statistically significant. A key finding was that a dummy locational variable was able to account for spatial autocorrelation adequately. Els and Von Fintel (2010) developed pooled log linear and quantile regression models to estimate house price growth in the Western Cape, South Africa. The researchers found that the parametric assumptions of the log linear model were violated and that the explicit functional form was incorrectly specified. This led the researchers to develop a quantile regression model where they found the model coefficients varied across quantiles, indicating that hedonic prices were sensitive across the marginal distribution of characteristics. Du Preez, *et al* (2013) developed a hedonic price function for houses in Walmer, Port Elizabeth South Africa using the local constant estimator where the direct estimate of  $E(y|x)$  is derived with a kernel function that produces a smooth estimate of the densities. In their case  $E(y|x)$  is the expected price of a home  $y$  conditional on a set of  $x$  home characteristics. The researchers found that this non-parametric technique outperformed OLS. A potential problem with transforming property prices to the log scale is that exponentiation of the fitted values produces geometric mean (Olivier *et al*, 2008) or median estimates

depending on whether the distribution of  $\log(x)$  is symmetric (Musset, 2006). In the case of this study,  $\log(x)$  would be the natural logarithm listing prices.

Another potential concern is the assumption that property prices are log normal when a different distribution family may represent the data generating process better. Extending linear modelling to the exponential family of distributions, where fitted values are kept on the original scale, can be accomplished by using generalised linear models (Mc Cullagh and Nelder, 1989). Bax and Chasomeris (2019) developed a hedonic pricing function for apartments in KwaZulu-Natal coastal submarkets using a gamma generalised linear model, keeping estimates on the original scale. All the parametric assumptions were satisfied, and bootstrapping was used to validate model generalisation. However, Rosen (1974) suggests that as the marginal cost of the characteristics increase for the seller the hedonic price function is unlikely to be linear. This is an important tenet because it informs the modelling approach. The modelling approach should not impose explicit functional form such as linearity but rather determine the shape of the relationship between listing price and property characteristics. Generalised additive models determine the shape of the relationship between the response and covariates through non-parametric smoothers which can be useful to describe complex relationships (Hastie and Tibshirani, 1986). Pace (1998) showed that generalised additive models outperformed parametric log linear and polynomial log linear models, accounting for non-linearities effectively, in a study estimating residential property prices in Memphis, USA. Hill and Scholz (2018), conducted a hedonic pricing study for Sydney Australia also using non-parametric smoothers or splines and found that the addition of the geospatial data performed only marginally better compared to the inclusion of postcode dummy covariates. Exploiting the hierarchical structure of homes to model the spatial heterogeneity thereof can be accomplished using hierarchical models (Brunauer *et al*, 2013). Tan *et al* (2019) found that nested models which exploit a hierarchical structure, outperform non-hierarchical linear models when dealing with properties nested at multiple geographic locations in a study of Chinese house prices. Homes are nested within a neighbourhood or blocks which hierarchical models account for, dealing with the lack of independence and avoiding Type 1 errors. Uyar and Brown (2007) support this, showing how hierarchical linear models were effective in capturing spatial cross-classification in schools.

Statistical learning is a recent development in the field statistics that leverages statistics, machine learning and computer science to understand complex data and solve contemporary business and scientific questions (James et al, 2013). Supervised statistical learning develops models used in predictive tasks where an output is estimated as a function of one or more inputs (Kuhn and Johnson, 2018). Supervised statistical learning involves developing predictive models on training data that generalise to unseen holdout data (Hastie, Tibshirani and Friedman, 2005). Statistical learning has also shown to be effective in developing flexible hedonic price models. In a study of Onondaga County, New York, USA Yoo *et al* (2012) compared two tree-based machine learning algorithms to stepwise OLS. Their findings showed that both algorithms resulted in lower prediction errors and that random forests were a useful variable selection method. Gradient boosting machines is another example of statistical learning where decision trees are grown sequentially using information from previous trees. Wezel *et al* (2005) applied gradient boosting, a nonparametric machine learning algorithm, and stepwise OLS to develop hedonic price functions for three different data sets. The findings showed that the gradient boosting algorithm achieved a reduction in the out-of-sample errors in comparison to the stepwise OLS. In a hedonic study of single-family homes in Switzerland between 2005 and 2017, Mayer *et al* (2019) found that gradient boosted machines provided the best accuracy compared to other methods such as OLS, robust linear models, mixed effects models, random forests and neural networks. However, they found that mixed effects or hierarchical models were the next best candidate methodology while neural networks showed “erratic” results.

Different hedonic functions are available to map property characteristics to property prices with varying degrees of effectiveness and transparency. The estimation of hedonic price function is the starting point in developing a hedonic price index where index number theory is then applied to the counterfactual predicted values to produce the property price index.

#### 1.1.4 South African Methodologies

Several economic intuitions in South Africa produce property price indices with varying methodologies. The Absa Bank house price index is available from as far back as 1966 and calculates the average house price of residential properties between 80m<sup>2</sup>

– 400m<sup>2</sup> in size, which are categorized into three segments valued at less than ZAR4.4 million (du Toit, 2016). These indices are then weighted by the volume of approved loans to compile the final index. First National Bank (FNB) compiles a house price index which uses a fixed weighted average approach that is applied to various segments of the market based on the size of properties and number of bedrooms (Loos, 2016). Standard Bank prepares a median price index, where the median is chosen as the central point of tendency, due to it being a more robust measure (Nhleko and Tlatsana, 2009). Absa Bank, Standard Bank and FNB use their own approved mortgages to compile their respective RPPI's. Lightstone Property applies the repeat sales methodology in contrast to deriving average or median indices, which they consider to be less influential by the mix of transaction properties and adheres to international index methodology standards (Lightstone Property, 2017). Lightstone create indices using the South African Deeds Office data. A drawback of the current implementation of these indices, is they discount certain property types. The United Kingdom makes use of a variety of house price index methodologies including median mix adjusted methods, repeat sales and hedonic pricing whilst the United States uses the well-known Standard and Poor's Case Shiller index, a repeat sales methodology index (Silver, 2016). Hedonic pricing does not appear to feature as a candidate methodology used by South African institutions although it appears extensively in literature.

## **1.2 Research Gap and Objectives**

Estimating residential property prices and price inflation is important and well established in real estate economic literature. Several South African institutions provide residential property price indices. However, hedonic regression does not appear to feature as a candidate methodology although it is ubiquitous in literature. Furthermore, the price indices produced by South African residential institutions do not extend to different property types. The South African studies that have been identified, focus on geographic segments of the South African property market and not the market in its entirety. No previous South African real estate price studies have been identified that investigate the use of hierarchical generalised additive models or machine learning models to develop hedonic listing price functions for various property types throughout South Africa although hedonic methods have shown to be effective in international studies. This study bridges these gaps by developing hedonic pricing

functions to estimate property listing prices and price developments for different types of homes throughout the South African residential property market, making use of both statistical and machine learning approaches. To achieve this objective, generalised linear models, hierarchical generalised additive models and several machine learning algorithms were investigated. Several generalised linear models were created including log linear models acting as baseline comparison to existing literature. Hierarchical generalised additive models and machine learning algorithms which relax the assumption of explicit function form with hierarchical models accounting for the nested structure of homes were constructed and fitted. Amongst all the models considered, the model with the lowest out of sample error is then selected and used to create a RPPI. Finally, a web application is developed to easily communicate the results thereof. The web application allows users to ascertain average home prices and price developments, very simply, through a front-end web interface without having to trawl through a plethora of property listings.

### **1.3 The Significance and Contribution of the Study**

The beneficiaries of the research include homeowners wishing to sell their homes, potential buyers wanting to obtain an average estimate along with property portals and real estate agencies. People wishing to sell their homes are faced with the challenging question of what price to list their homes for on the market. They have several resources available to help determine this themselves, print material such as real estate listing publications or online sources, including real estate agency websites and property portals. Online property portals aggregate property listings from real estate agencies and disseminate these pooled listings through online user interfaces such as smartphone applications and websites. South African examples of property portals include Private Property and Property24, and some international examples include Zillow, Zoopla and Rightmove. Regardless of the source of information that sellers may use, they are faced with the time-consuming task of trawling through a plethora of listings in order to gauge what price their homes could fetch on the market. Alternatively, interested sellers may seek the help of professional real estate agents to value their homes using comparative market analyses.

Similar to estate agents, a hedonic model performs a CMA, however, it does so with a mathematical model. Potential buyers may want to compare a property of interest to



the average price of properties with the same set of characteristics in the same location. Hedonic models make this task possible, enabling potential buyers to obtain an understanding of price developments over time through a RPPI. This study contributes to the industry by developing hedonic models to appraise residential property and measure price developments over time, comparing several statistical learning methods and making the results consumable through a practical web application or data product. Property portals and real estate agencies are well positioned to leverage their extensive market data, developing such models to guide sellers in their price setting endeavours, resulting in increased traffic through innovative data products such as the one presented in this research. Furthermore, this study contributes to the field of applied statistics by showing the importance and efficacy of iteratively applying different methods, ranging from parametric, semi-parametric to non-parametric, including novel machine learning approaches to solve contemporary scientific and business problems.

#### **1.4 Thesis Outline**

The rest of this thesis proceeds with the following chapters. Chapter two introduces the data of the study and presents the pre-modelling processing steps such as cleaning the and augmenting the data. Summary statistics are presented in both tabular and graphical form through exploratory data analysis. Chapter three begins the modelling of home prices by investigating generalised linear models as candidate functions to estimate homes prices, fitting various distributions and link functions. Chapter four explores hierarchical generalised additive models to create hedonic price functions. Spatial clustering is used to created homogenous groups based on distances between suburbs. Chapter five develops and compares two tree-based machine learning algorithms and neural networks where 5-fold cross validation is used to tune model hyperparameters. Partial dependence and variable importance plots are used to understand the effects of the covariates on the response. Chapter six discusses index number theory and develops the RPPI using the best candidate model. Furthermore, the web application is created and illustrated, showing the practical economic value of the study. Finally, chapter seven provides an overall conclusion for this thesis, summarizing the main findings and presents suggestions for further research.

## Chapter Two

### The Data and Exploratory Data Analysis

Effective data analysis and modelling requires data integrity. Data cleaning and preparation is a fundamental component of any statistical analysis and study (de Jonge and van der Loo, 2013). The chapter that ensues addresses the data cleaning and preparation involved in this research. Furthermore, this chapter explores the data and presents important summary statistics.

#### 2.1. Description of the Data

The data used in this study was obtained from an online property portal, Private Property (<https://www.privateproperty.co.za/>), which aggregates real estate agencies listings data throughout South Africa. The period of the data spanned January 2013 to August 2017. The different property types include apartments, houses, townhouses, clusters, duplexes and simplexes. The data was enriched by obtaining the spatial coordinates of the suburb of each property using a geocoding application programming interface. This information is used to inspect spatial dependency. Table 2.1 provides a description of the data.

**Table 2.1.** Description of the data

| Variable      | Description   |
|---------------|---|
| Listing Price | The advertised price of the property in ZAR                         |
| Size          | The size of the physical structure of the property in square meters |
| Bedrooms      | The number of bedrooms in the property                              |
| Bathrooms     | The number of bathrooms in the property                             |
| Property Type | The type of property  |
| Suburb        | The suburb the property is located                                  |
| Province      | The province the property is located                                |
| Listing Date  | The advertisement date of the property on the portal                |
| Latitude      | The latitude coordinates of the suburb the property is located      |
| Longitude     | The longitude coordinates of the suburb the property is located     |

Adjusting for quality change in residential properties over time is important when measuring residential property prices. Changes in property prices should reflect pure price changes, not simply changes in the composition of samples at different points in time which can be mostly accounted for by measuring key property attributes (Hill, 2011). Key property attributes cited by de Haan and Erwin (2011) include the size of

the property, the location of the property, the type of the property, the age and material used in the construction of the property and other physical attributes such as the number of bedrooms, bathrooms etc.

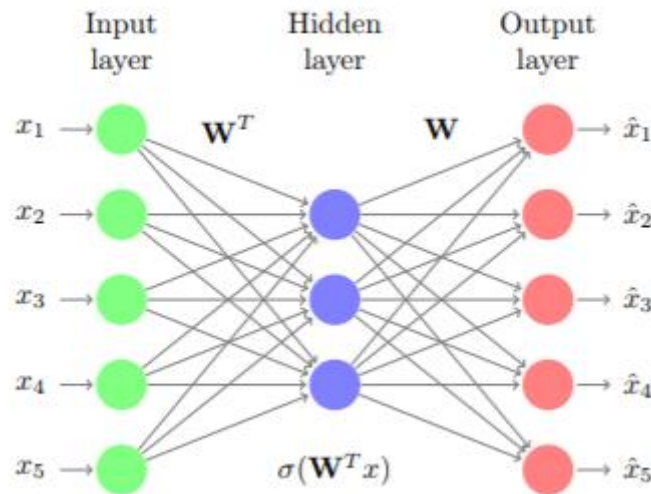
The data required considerable cleaning including capturing categorical variables correctly and ensuring numeric variables were constrained to the correct data type. Missing fields were removed resulting in the removal of observations. Duplicate property listings were identified in each period using row wise string matching and subsequently removed. The data showed a large spread in the numeric variables of interest, shown in Table 2.2.

**Table 2.2.** Clean data summary statistics

|              | Listing Price | Size   | Lot    | Bedrooms | Bathrooms |
|--------------|---------------|--------|--------|----------|-----------|
| Minimum      | 1 000         | 2      | 2      | 0        | 0         |
| 1st Quartile | 950 000       | 98     | 132    | 2        | 2         |
| Median       | 700 000       | 200    | 566    | 3        | 2         |
| Mean         | 2 461 210     | 259.8  | 1 163  | 3.135    | 2.252     |
| 3rd Quartile | 2 950 000     | 330    | 1 014  | 4        | 3         |
| Maximum      | 200 000 000   | 85 102 | 99 999 | 78       | 78        |

The way the data was obtained by the property portal could be subject to incorrect data capturing as real estate agents manually populate the information before it is disseminated via automatic feeds to the property portal. A risk of incorrect data capturing exists and is a fair assumption through the examination of Table 2.2 where the maximum and minimum values seem improbable. An autoencoder, which is a deep learning neural network, was developed to identify anomalous data points using the h2o open-source machine learning stack from Ledell *et al* (2019). Autoencoders generalise the concept of non-linear principal component analysis where the feature space is reduced via a bottleneck at the hidden middle layers, learning the non-linear representation of the inputs, with the output layer aimed at reproducing the input layer given this restricted representation (Hastie *et al*, 2015). The network is able to learn the identity of the data via a non-linear reduced representation of the original data, where a high reconstruction error for data points indicates non-matching of the learned pattern (Candel *et al*, 2018). Reasonable lower limits were set on certain variables

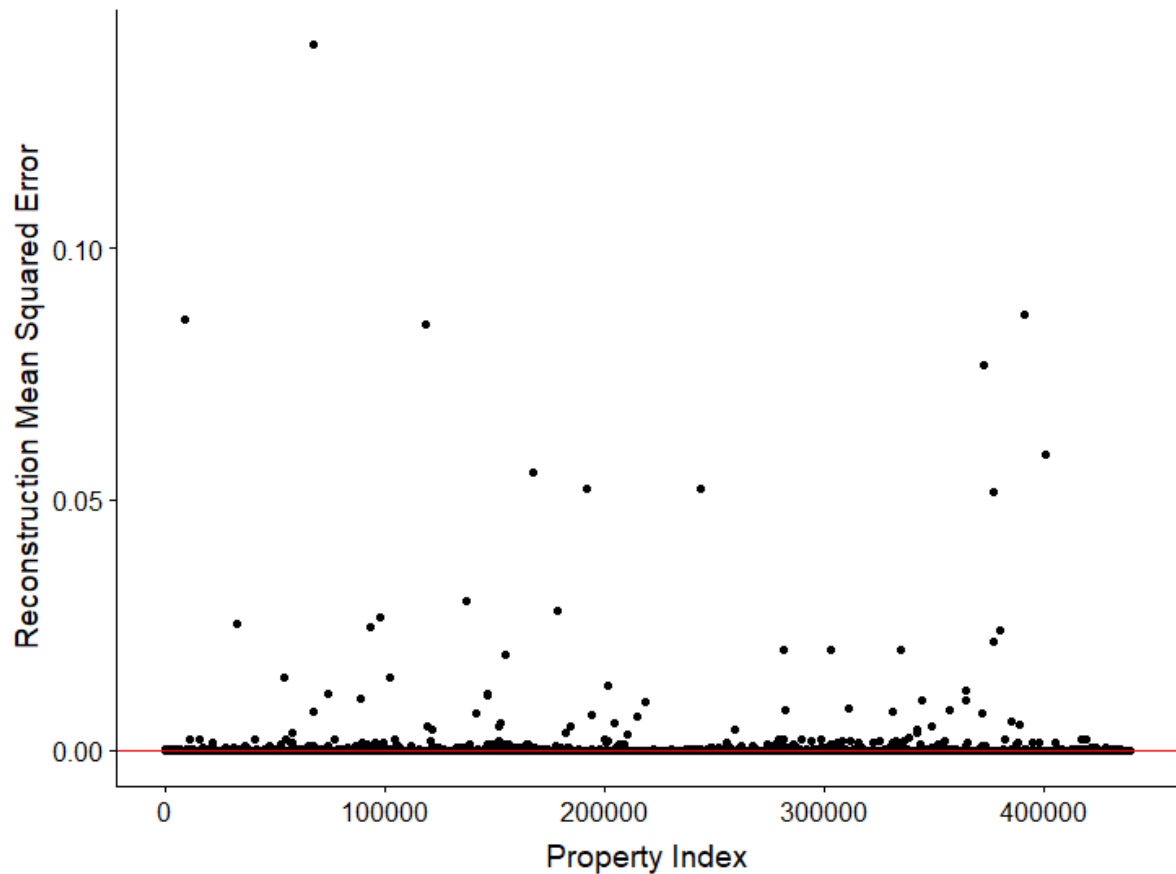
using the ABSA bank property price index, the oldest price index in South Africa, as a guideline (du Toit, 2016). Listing price was set to  $\geq$  ZAR 200 000 and size was set to  $\geq$  35 square meters. An example of the architecture of an autoencoder is presented in Figure 2.1 where through hidden layers and neurons the identity of the data is mapped via non-linear reduction and reconstruction.



**Figure 2.1.** Autoencoder diagram

Source: Hastie *et al* (2015)

Linear combinations of the input vector  $x$  are created from a  $\rho \times m$  matrix of weights  $W$  where  $m < p$  and each of the linear combinations are passed through a non-linear function  $\sigma$  by means of the vector function  $h(x) = \sigma(W^T x)$ . Modelling the output layer is then given by  $Wh(x) = W\sigma(W^T x)$ . For  $i = 1, \dots, N$ , of the input vectors  $x_i$ ,  $W$  is estimated via nonconvex optimization (Hastie *et al*, 2015). This presents a basic example of an autoencoder without bias terms which are omitted for simplicity. The results of the autoencoder were promising at identifying anomalous data where properties with a reconstruction mean squared error  $\geq 9.39e-07$  were deemed anomalous and subsequently discounted as illustrated in Figure 2.2.



**Figure 2.2.** Autoencoder reconstruction error

After accounting for data enriching, cleansing and anomaly detection, the final dataset consisted of 415 200 properties. A summary of the data is presented in Table 2.3 where the spread of the variables is noticeably reduced and more plausible.

**Table 2.3.** Final data summary statistics

|              | Listing Price | Size  | Bedrooms | Bathrooms |
|--------------|---------------|-------|----------|-----------|
| Minimum      | 200 000       | 35    | 1        | 1         |
| 1st Quartile | 958 000       | 100   | 2        | 2         |
| Median       | 1 690 000     | 200   | 3        | 2         |
| Mean         | 2 159 173     | 230.7 | 3.091    | 2.157     |
| 3rd Quartile | 2 799 000     | 315   | 4        | 3         |
| Maximum      | 19 700 000    | 2 080 | 13       | 12        |

## 2.2. Summary Statistics

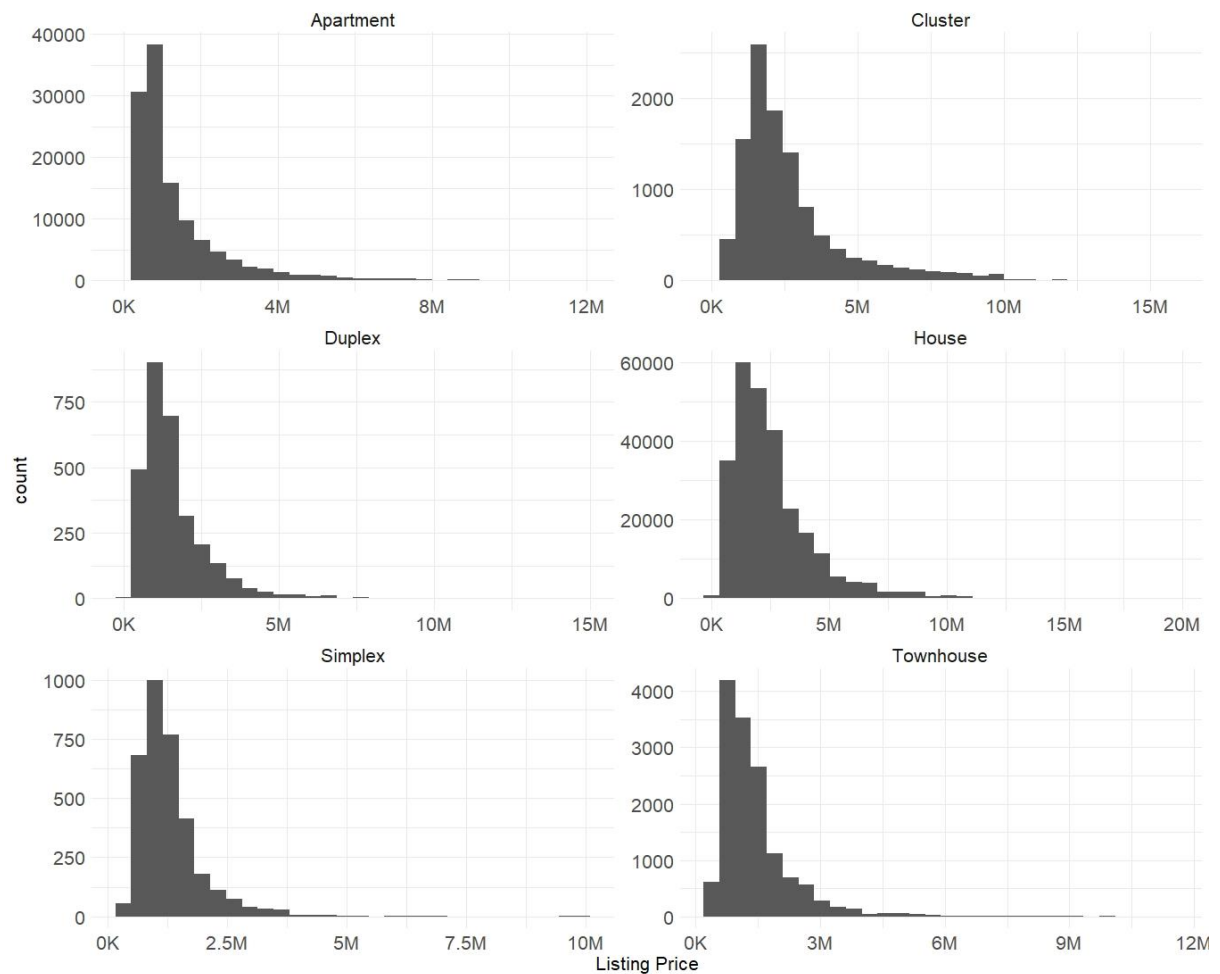
Correlation measures the strength and existence of a relationship between two or more variables (Keller, 2012). Table 2.4 details the pairwise Spearman correlation coefficients between all numeric variables in the data along with the level of significance. The correlation coefficient is the highest between the number of bedrooms and bathrooms. Listing price is most highly correlated with size and number of bathrooms.

**Table 2.4:** Correlation matrix

|               | Listing Price | Size     | Bathrooms |
|---------------|---------------|----------|-----------|
| Listing Price |               |          |           |
| Size          | 0.69****      |          |           |
| Bathrooms     | 0.64****      | 0.69**** |           |
| Bedrooms      | 0.52****      | 0.70**** | 0.76****  |

p < .0001 \*\*\*\*, p < .001 \*\*\*, p < .01 \*\*, p < .05 \*

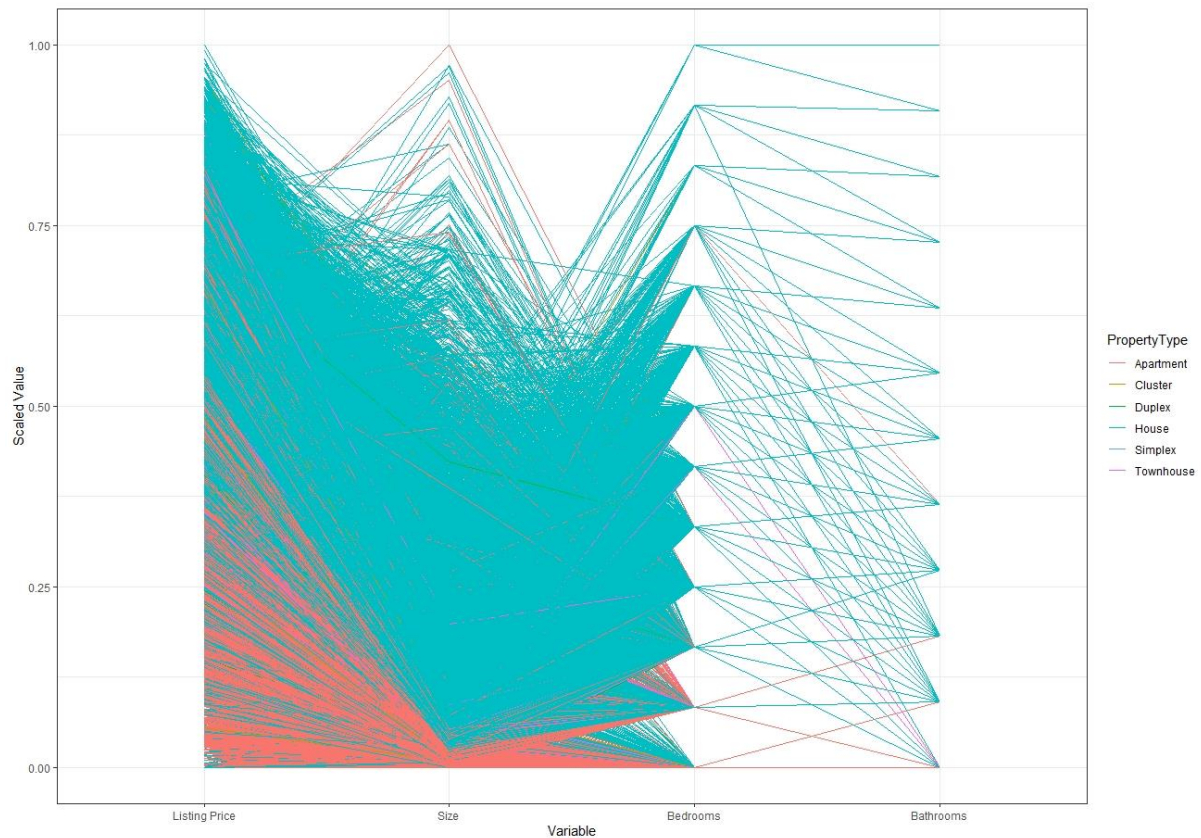
Understanding the distribution of a response variable is important in statistical studies. Choosing the best estimator for a sample distribution is paramount and contingent on the statistical properties (Greene, 2003). Figure 2.3 depicts the histograms of listing price by property type.



**Figure 2.3:** Property type listing price histograms

All property types appear to be positively skewed with long tails at different levels of kurtosis.

Parallel plots are useful data exploratory visualizations for high dimensional, multivariate data (Heinrich and Weiskopf, 2013). Figure 2.4 depicts a parallel plot of the numeric variables grouped by the property type.



**Figure 2.4:** Parallel plot

The parallel plot aims to achieve a graphical summary where different property types have different discriminant profiles. The natural logarithm was applied to listing prices to aid in visualization. All variables are univariately scaled, so the minimum of each variable is zero and the maximum is one. Apartments seem to have more homogeneity with respect to listing price, size, bathrooms and bedrooms with houses appearing to have greater variability.

Paramount to the problem of property prices over time is the need to compare similar properties in successive periods, the concept of like with like comparisons (Silver, 2016). Residential property sales are typically characterised as infrequent which makes measuring price changes challenging (Jiang, Phillips and Yu, 2015). This study makes use of listing prices and not transaction prices which increases the sample size as more properties are available for sale than get sold. Using the listing price is viewed as a valid approach for these reasons (Shimizu *et al*, 2010). A breakdown of the number of property types available for sale in each year of the sample period is provided in Table 2.5.



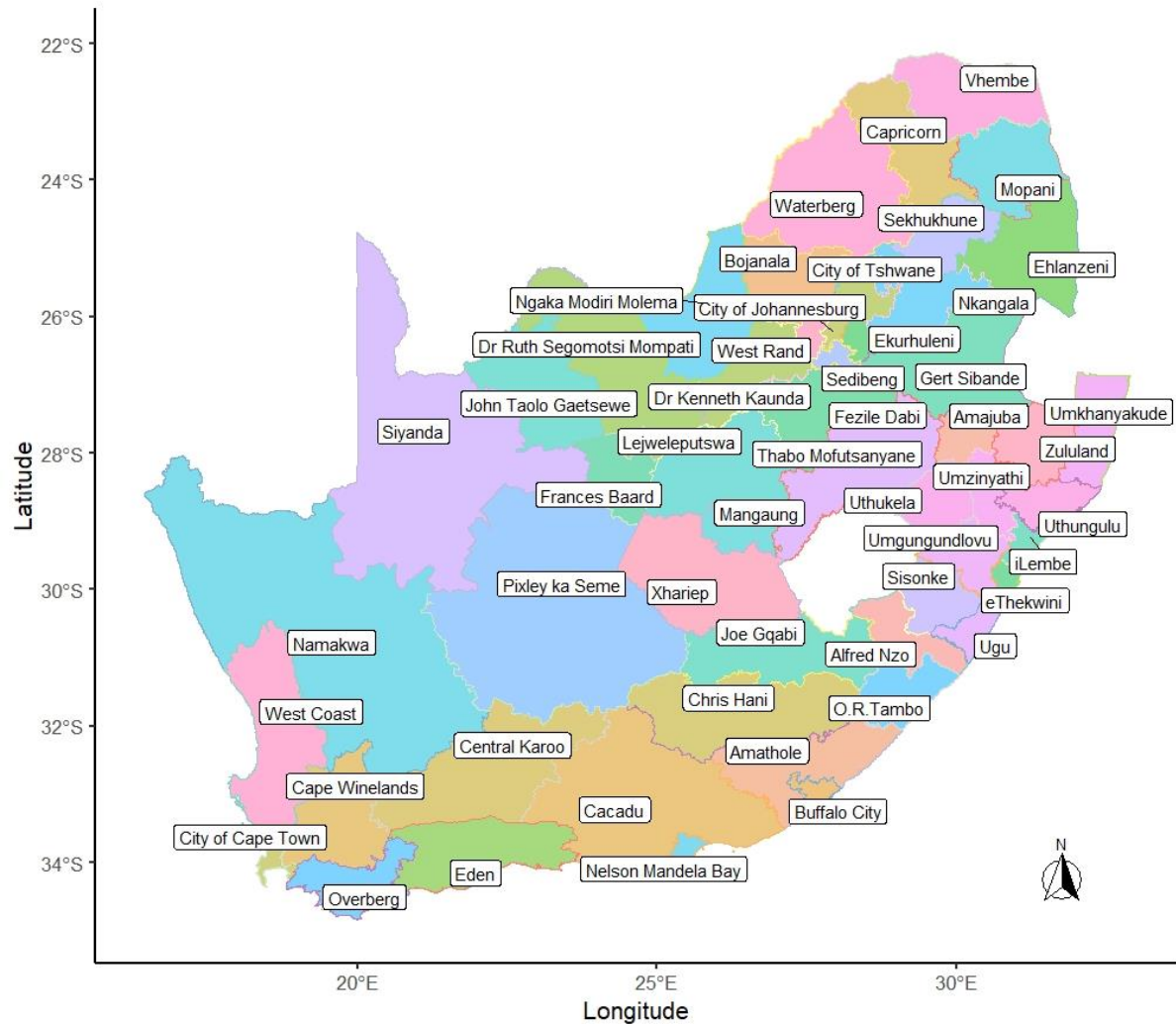
**Table 2.5:** Yearly property type frequency table

| Property Type | 2013   | 2014   | 2015   | 2016   | 2017   |
|---------------|--------|--------|--------|--------|--------|
| Apartment     | 11 732 | 21 025 | 24 598 | 35 799 | 26 181 |
| Cluster       | 713    | 2 573  | 2 313  | 3 147  | 2 053  |
| Duplex        | 342    | 552    | 548    | 843    | 665    |
| House         | 18 812 | 54 851 | 57 010 | 79 927 | 53 810 |
| Simplex       | 610    | 748    | 696    | 789    | 578    |
| Townhouse     | 165    | 2 731  | 2 937  | 4 833  | 3 619  |

There are no years where data did not exist for different property types. Furthermore, houses and apartments account for most of the sample. Properties in 2017 were only captured until August.

### **2.3. Spatial Nature of the Data**

This research analysed property listings throughout South Africa, Figure 2.5 illustrates the main municipalities located in South Africa.



**Figure 2.5:** South African municipalities

Prices of adjacent properties are often related which can lead to correlation in the residuals of regression models, violating the assumption of independence (Bourassa *et al*, 2007). Spatial autocorrelation or dependency is a challenging problem in real estate modelling where correlation manifests in two-dimensional space unlike serial correlation which is one dimensional. Bourassa *et al* (2007) found that the inclusion of a submarket dummy variable accounted for spatial autocorrelation and outperformed geostatistical and lattice approaches. A similar approach was adopted by Bax and Chasomeris (2019) where a fixed location effect was included in the gamma generalised linear, accounting for any spatial autocorrelation in the residuals.

Variograms are useful diagnostic informal checks which assist in understanding the spatial autocorrelation structure. Variograms display the dissimilarity of observations

that vary in space as a function of the distance between them (Ploner, 1999). The sill represents spatially autocorrelated sample locations, and the range is where the distance flattens out and the sample locations are no longer spatially autocorrelated. A variogram will be flat when no correlation or low correlation is present which indicates randomness in the structure (Chiles and Delfiner, 1999). The nugget effect is an important concept in variograms and describes the variability between observations that are closely spaced which could be inherent in the data or due to the sampling component (Clark, 2010). Therefore, in the context of this study, a large nugget effect could be the product of closely clustered properties with similarly signed and order of magnitude residuals that would overestimate the amount of spatial dependency. A prevalent test developed by Moran (1948), is a two-dimensional specification test for spatial autocorrelation, analogous to a test of univariate time series correlation (Anselin, 2006):

$$I = \frac{e'we/s_0}{(e'e)/n},$$

where  $e$  represents the regression model residuals and  $w$  is the spatial weighting matrix and  $s_0$  is a standardization factor that relates to the sum of weights for the non-zero cross-products.

The spatial dependency of residential properties is an important aspect to consider in the development of hedonic regression models, this research makes use of both informal and formal checks to ensure the assumptions of independence is accounted for.

## 2.4. Summary

This chapter presented the essential data cleaning and preparation required to model the data. An autoencoder, a deep learning neural network, was built to identify and remove anomalous data. Exploratory data analysis was performed, presenting important summary statistics and visualisations. Finally, the spatial scope and nature of the data was presented with methods on how the assumption of independence will be tested both informally and formally in the following chapters. Following from the exploratory data analysis, generalised linear models are investigated next the first

candidate methods to develop residential property appraisal functions because of the ability fit exponential family of distributions which could be appropriate given the distribution of listing prices illustrated in the exploratory data analysis.

## Chapter Three

### Generalised Linear Models

The chapter that ensues introduces the methodology of generalised linear models to expand on the typical log linear hedonic model approach which is pervasive in real estate econometric literature. Extending the scope to the exponential family of distributions, various linear models with different distributions and link functions are compared and measured against different goodness of fit criteria.

#### 3.1 Model Description and Motivation

Generalised linear models are a natural extension of classical linear models where properties such as linearity and computing parameter estimates are similar (Mc Cullagh and Nelder, 1989). Generalised linear models are characterised by three components. Firstly, a stochastic or random component representing a response variable  $Y$ , consisting of independent observations  $(y_1, y_2, \dots, y_n)$ , belonging to a class of an exponential family distribution in the form of:

$$f(y; \theta, \phi) = \exp\left\{\frac{\theta y - b(\theta)}{\phi} + c(y, \phi)\right\},$$

where  $\phi$  is a dispersion parameter and  $b(\cdot)$ ,  $c(\cdot)$  are known functions and the range of  $Y$  does not depend on  $\theta$  or  $\phi$ . For a random response variable  $Y$  with distribution of form  $E(Y) = \mu$ . Secondly, a systematic component which consists of a set of covariates  $(x_1, x_2, \dots, x_p)$  which combine linearly with the coefficients to produce the linear predictor  $\eta$ . Therefore  $\eta = \beta X$ . Finally, a link function which connects the stochastic and systematic components where  $\eta = \mu$ .

This generalisation takes the form:

$$\eta_i = g(\mu_i),$$

where  $g(\cdot)$  denotes the link function and  $\eta = \mu$  through the link function. The link function relates the conditional mean to the systematic component, namely the covariates (Jones, 2010). This formulation allows for the exponential family of

distributions including normal, however, the link function may become any monotonic differentiable function, which then allows extensions to distributions such as Poisson, binomial and gamma amongst others (Mc Cullagh and Nelder, 1989). This means that generalised linear models are suitable for modelling continuous data as well as count and binary data.

Generalised linear models obtain maximum likelihood estimates of parameters belonging to an exponential distribution family using the iterative reweighted least squares algorithm where the link function makes the systematic effects linear (Nelder and Wedderburn, 1972). Maximum likelihood estimates are a vector of parameter estimates produced by a model function which makes the observed data probable, given the model function (Lindsey, 1997).

The primary goodness of fit measure for generalised linear models is called the deviance which is the logarithm of a ratio of likelihoods (Mc Cullagh and Nelder, 1989). The analysis of deviance makes model assessment and comparison possible in terms of the choice of covariates. Given a set of data, two extreme models are possible. Firstly, a null model with one parameter which represents a common  $\mu$  for all the  $y$ s. Secondly, a complete model where all the  $y$ s are different, matching the data completely. Fitting a model with more than one parameter represents a saturated model that can be compared to the null model (Dobson and Barnett, 2008). The fitting of  $n$  parameters is performed by maximizing the likelihood of matching the model to the likelihood of the data through the deviance that differs based on the distribution.

For the normal distribution, the deviance is simply the sum of squares just like ordinary least squares which means that fitting a normal or log normal distribution, where the natural logarithm of the response is taken, is equivalent to fitting a linear or log linear ordinary least squares model. For generalised linear models, the saturated model should have a lower deviance than the null model, indicating the inclusion of  $n$  parameters are a better fit. Guisan and Zimmernam (2000) propose that variance reduction in model formulation is generally a desired characteristic of the goodness of fit as with generalised linear models, where deviance reduction can be converted to an equivalent  $R^2$  statistic:

$$D^2 = (Nulldeviance - Residualdeviance) / Nulldeviance,$$

where  $D^2$  is the deviance explained or the amount of deviance accounted for by the model. Naturally, this leads to an understanding of the residuals of generalised linear models where the deviance residuals are reported as a measure of discrepancy. Deviance residuals are calculated as follows:

$$sign(y_i - \hat{\mu})\sqrt{\hat{d}_i^2}.$$

This formulation shows that deviance residuals are calculated by taking the signed square root of the  $i$ th observation to the total model deviance (Jackson, 2008). One can begin to understand the quality of fit that reflects the choice of the link function and linear predictor using deviance residuals (Nelder and Wedderburn, 1972). McCullagh and Nelder (1989) state that through the appropriate link function and linearity of the systematic component, the desired error distribution of the deviance residuals can be achieved which should resemble normal theory residual plots, except for certain plots in the case of binomial errors. Standardized deviance residuals are approximately normal which is preferable to Pearson residuals that tend to reflect any skewness of the underlying distribution. Plotting the standardized deviance residuals against the fitted values can provide an informal check of the goodness of fit depending on the type of generalised linear model, where any curvature could suggest the incorrect choice of link function, omitted independent variables or the omission of quadratic terms in the independent variables (Davidson and Snell, 1991).

The selection of generalised linear models in this study involved choosing the appropriate distribution of  $Y$  and choosing the relationship between  $\eta$  and  $\mu$ . Three candidate combinations of model families and link functions were fit to the data, specifically, the gamma log model, the normal log model and the log normal identity model.

### 3.2 Results and Discussion

The data was split into two sets, training and validation where 70% of the data was used for training and 30% was used for validating the models. This was done to test model generalisability on unseen data for the development of future models. The holdout data for a given model provides a more robust estimate of the generalisation error compared to the training error (Blum, Kalai and Langford, 1999). Partitioning the data into training and holdout sets for each year involved writing a function to ensure that the splits were random, and that the distribution of the response was similar for each split and to the original data. The function ensured that each suburb factor level was present in each split. Model performance and generalisation was tested using the root mean squared error (RMSE) which is a measure of spread that compares the closeness of the model outcomes to the observed data (Gujarati, 2004). A lower RMSE is indicative of less variability between model estimates and the observed data. The Akaike information criterion (AIC) statistics were also computed. When comparing models, the AIC is useful for model selection as it provides an assessment of the quality of different models given a set of data (Greene, 2003). A lower AIC is indicative of better fit. AIC concomitantly considers goodness of fit using the likelihood function whilst penalising model complexity through the number of parameters. Model selection was based on a combination of reported statistics namely, deviance explained, holdout RMSE, AIC and model fit based on diagnostic residual plots. Tables 3.1 details the results of each yearly model fit.



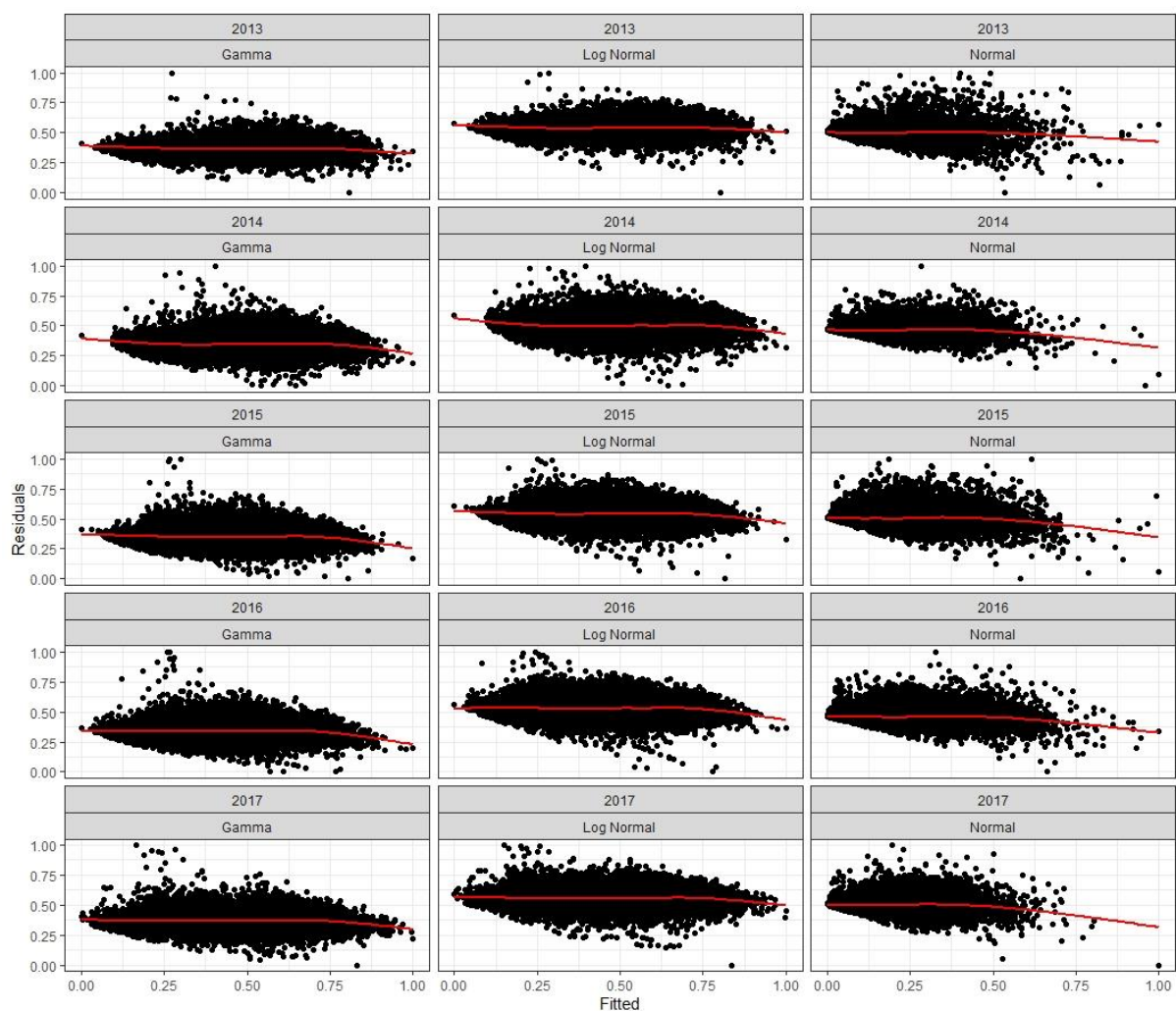
**Table 3.1:** Comparison of GLM summaries

| Year  | Deviance Explained | Training RMSE | Holdout RMSE | AIC     |
|---|--------------------|---------------|--------------|---------|
| <b>Gamma model summary statistics:</b>      |                    |               |              |         |
| 2013  | 0.89               | 708 816       | 719 286      | 665059  |
| 2014  | 0.87               | 762 918       | 764 730      | 1689332 |
| 2015  | 0.87               | 768 004       | 772 905      | 1808202 |
| 2016  | 0.87               | 731 159       | 746 554      | 2557727 |
| 2017  | 0.88               | 723 854       | 724 390      | 1779451 |
| <b>Normal model summary statistics:</b>     |                    |               |              |         |
| 2013  | 0.83               | 666 427       | 704 715      | 692589  |
| 2014  | 0.82               | 724 525       | 743 688      | 1748933 |
| 2015  | 0.83               | 721 649       | 743 687      | 1866417 |
| 2016  | 0.83               | 685 171       | 710 324      | 2637126 |
| 2017  | 0.84               | 682 036       | 693 513      | 1835507 |
| <b>Log normal model summary statistics:</b> |                    |               |              |         |
| 2013  | 0.89               | 709 765       | 716 993      | 664303  |
| 2014  | 0.88               | 766 117       | 762 544      | 1687050 |
| 2015  | 0.87               | 767 251       | 774 243      | 1805578 |
| 2016  | 0.88               | 727 294       | 741 349      | 2553572 |
| 2017  | 0.88               | 724 097       | 726 167      | 1777340 |

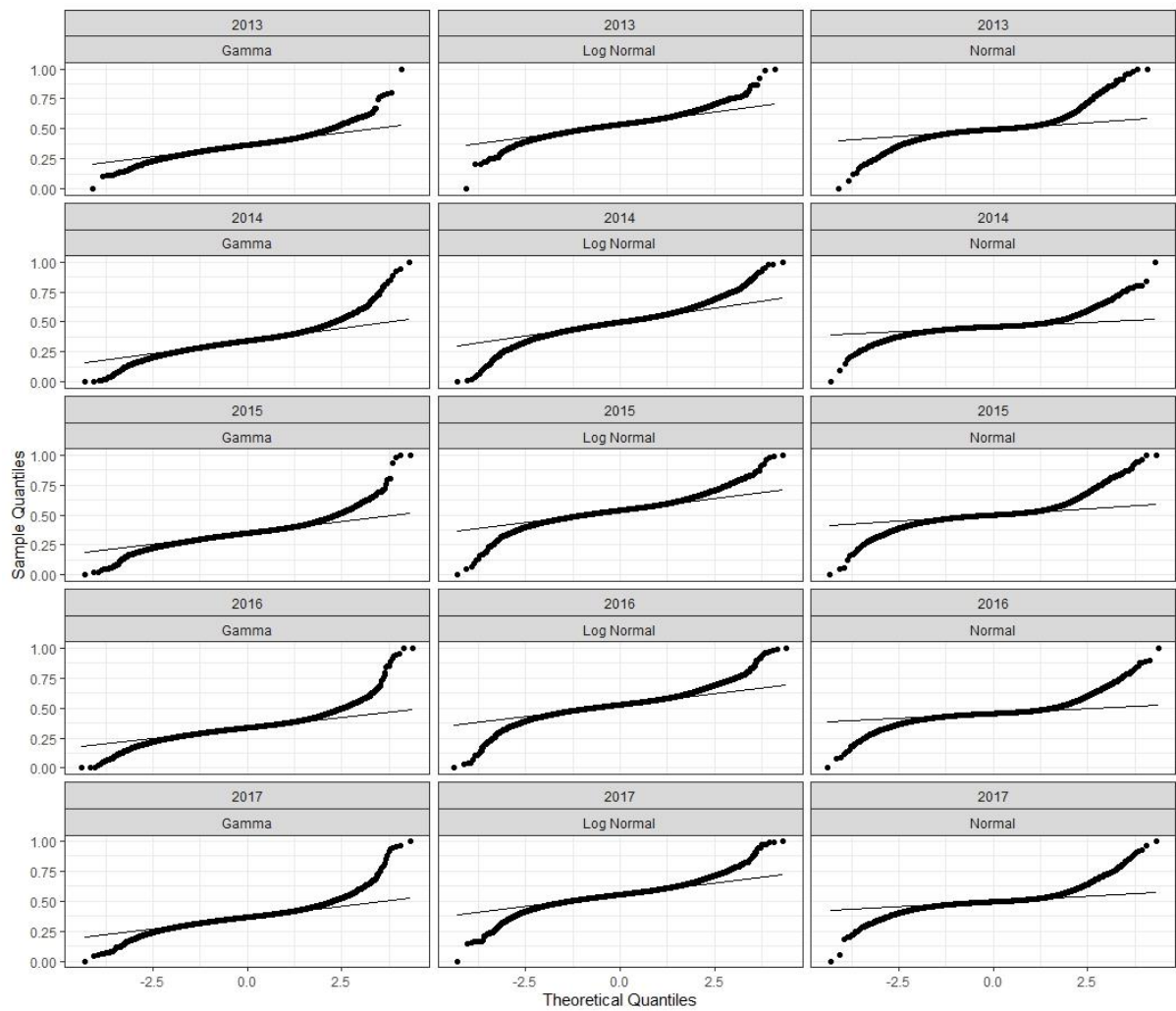
Notes: The deviance explained figures are rounded to two decimal places, all other figures are rounded to the nearest whole number.

Each model produced consistent deviance explained statistics for each year respectively, where the gamma and log normal models shared the highest amount of deviance explained. Moreover, the gamma and log normal models appear very similar in terms of holdout RMSE and AIC statistics. The AICs produced by the log normal models were not directly comparable to the other models as the response variable was on the logarithmic scale. The AICs of the log normal models were made comparable by subtracting the sum of logarithms of the response variable from the likelihood. Based solely on the AICs the log normal models appear to fit the data the best, as they consistently produced the lowest AIC statistics. Considering only the holdout RMSE statistics, the normal model outperformed the two other models with consistently lower RMSE statistics each year. No evidence of over fitting is present as the training and holdout RMSE's are quite similar, indicating the models generalise to unseen data. This suggests model robustness to the introduction of future periods.

Discerning the best model based solely on the goodness of fit measures reported above is difficult and a graphical examination of the residuals is necessary. The goodness of fit residual diagnostic plots for each yearly model are illustrated below in Figure 3.1 and Figure 3.2 below beginning with the gamma model, followed by the log normal and finally the normal. Figure 1 presents the residuals versus fitted values where the y-axis represents the deviance residuals and x-axis represents the linear predictor. Figure 3.2 presents the quantile-quantile (Q-Q) plots for normality. The scale has been normalised to a scale of  $[0, 1]$  for easy comparison of the residuals.



**Figure 3.1:** GLM fitted versus residual plots



**Figure 3.2:** GLM Q-Q plots

The fitted versus residual diagnostic plots for the gamma and log normal models are very similar and do not indicate any discernible pattern in the deviance residuals, one of the required assumptions, however, the normal log link model shows greater signs of heterogeneity at the upper quantiles, violating the assumption of constant variance. Although none of the plots are perfectly normal with deviation at the upper and lower quantiles, Schmidt and Finan (2018) provide empirical evidence that linear models without normally distributed residuals may still provide valid results, given sufficient sample size. The normal log link model appears to fit the data poorly in terms of diagnostic plots model. Based on the diagnostic plots the gamma and log normal models appear to represent the data better.

A possible caveat of using the log normal model for modelling listing prices is that the expected values are on the log scale and back transformation is necessary. Transforming expected values from the log scale back to the original scale by means of exponentiation results in geometric mean estimates and not arithmetic mean estimates (Olivier *et al*, 2008). However, the natural logarithm is monotonic, and the back transformed estimates are equivalent to median estimates if the distribution of  $\log(x)$  is symmetric (Musset, 2006). An appealing feature of the gamma and normal models are that expected values are kept on the original scale where arithmetic mean expected values are computed. For this reason, the log normal models are discounted from the candidate model selection. The gamma models are chosen over the normal models based on the diagnostic plots, lower AICs and lower RMSEs. A discussion of the gamma modelling results ensues where listing price was regressed on the physical and locational attributes.

The property type factor variable included 6 levels namely, apartment, cluster, duplex, house, simplex and townhouse. The property type apartment was used as the reference level, resulting in the other property types being compared to this level. Table 3.2 tabulates the beta coefficient estimates for each covariate along with the corresponding p-values. To make reporting succinct Table 3.2 discounted the factor variable suburb coefficients as there were over 2000 levels present in the data that varied between years. The suburb factor variable was used as a control variable to account for variability amongst listing prices and to account for the spatial dependency in the data.

**Table 3.2:** Gamma model results summary

| Year      | 2013          |         | 2014          |         | 2015          |         | 2016          |         | 2017          |         |
|-----------|---------------|---------|---------------|---------|---------------|---------|---------------|---------|---------------|---------|
|           | $\hat{\beta}$ | p-value | $\hat{\beta}$ | P-value | $\hat{\beta}$ | p-value | $\hat{\beta}$ | p-value | $\hat{\beta}$ | p-value |
| Intercept | 9.913         | 2e-9    | 11.013        | 2e-9    | 10.932        | 2e-9    | 11.305        | 2e-9    | 11.148        | 2e-9    |
| log(Size) | 0.664         | 2e-9    | 0.626         | 2e-9    | 0.558         | 2e-9    | 0.479         | 2e-9    | 0.512         | 2e-9    |
| Bedrooms  | 0.003         | 0.313   | 0.017         | 2e-9    | 0.021         | 2e-9    | 0.034         | 2e-9    | 0.025         | 2e-9    |
| Bathrooms | 0.111         | 2e-9    | 0.096         | 2e-9    | 0.112         | 2e-9    | 0.117         | 2e-9    | 0.112         | 2e-9    |
| Cluster   | 0.090         | 2e-9    | 0.136         | 2e-9    | 0.146         | 2e-9    | 0.187         | 2e-9    | 0.187         | 2e-9    |
| Duplex    | 3e-3          | 0.874   | 0.025         | 0.104   | 0.035         | 0.024   | 0.086         | 2e-9    | 0.079         | 2e-9    |
| House     | 0.027         | 3e-4    | 0.063         | 2e-9    | 0.103         | 2e-9    | 0.158         | 2e-9    | 0.141         | 2e-9    |
| Simplex   | 0.061         | 4-e5    | 0.068         | 5e-7    | 0.078         | 2e-9    | 0.117         | 2e-9    | 0.087         | 2e-9    |
| Townhouse | 0.050         | 0.064   | 0.063         | 2e-9    | 0.077         | 2e-9    | 0.090         | 2e-9    | 0.099         | 2e-9    |

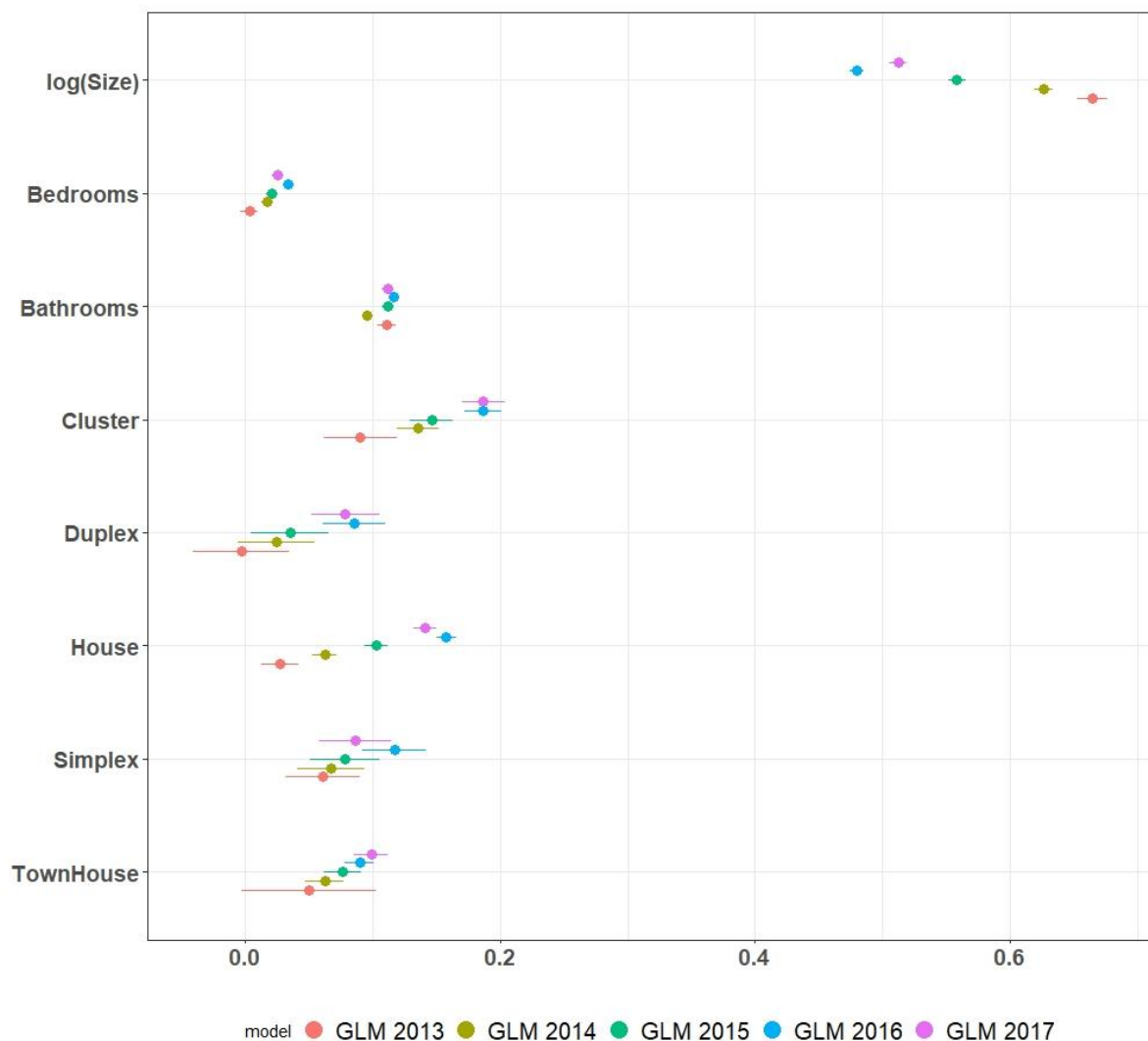
Notes: Numbers are rounded to 3 decimal places and scientific notation was adopted for brevity.

The coefficients, given by  $\hat{\beta}$  are expressed as percentage effects. The covariates  $\log(\text{Size})$  and the number of bathrooms were consistently statistically significant for each year. The natural logarithm was applied to the size covariate to improve linearity. The coefficients can be interpreted as follows:

1. A 1% increase in size (squared meters), on average increased the listing price of a residential property by  $\hat{\beta} \times 100$  (%) for a given year.
2. Each additional bedroom, on average increased the listing price of a residential property by  $\hat{\beta} \times 100$  (%) for a given year.
3. Each additional bathroom, on average increased the listing price of a residential property by  $\hat{\beta} \times 100$  (%) for a given year.
4. The property types in Table 3.2 are percentage difference comparisons between apartments where a property type was  $\hat{\beta} \times 100$  (%) greater than or less than apartments (reference level) depending on the sign in front of the  $\hat{\beta}$ .

Evident from Table 3.2 is that each additional bathroom, on average, contributes more to the listing prices of homes than each additional bedroom. An appealing feature of this parametric framework is the transparency and interpretability of the model coefficients. Property market participants are able to make informed decisions about renovating their homes by examining the marginal utility of different characteristics.

Figure 3.3 facilitates an easy way to assess how the GLM coefficients have changed over the observational periods.

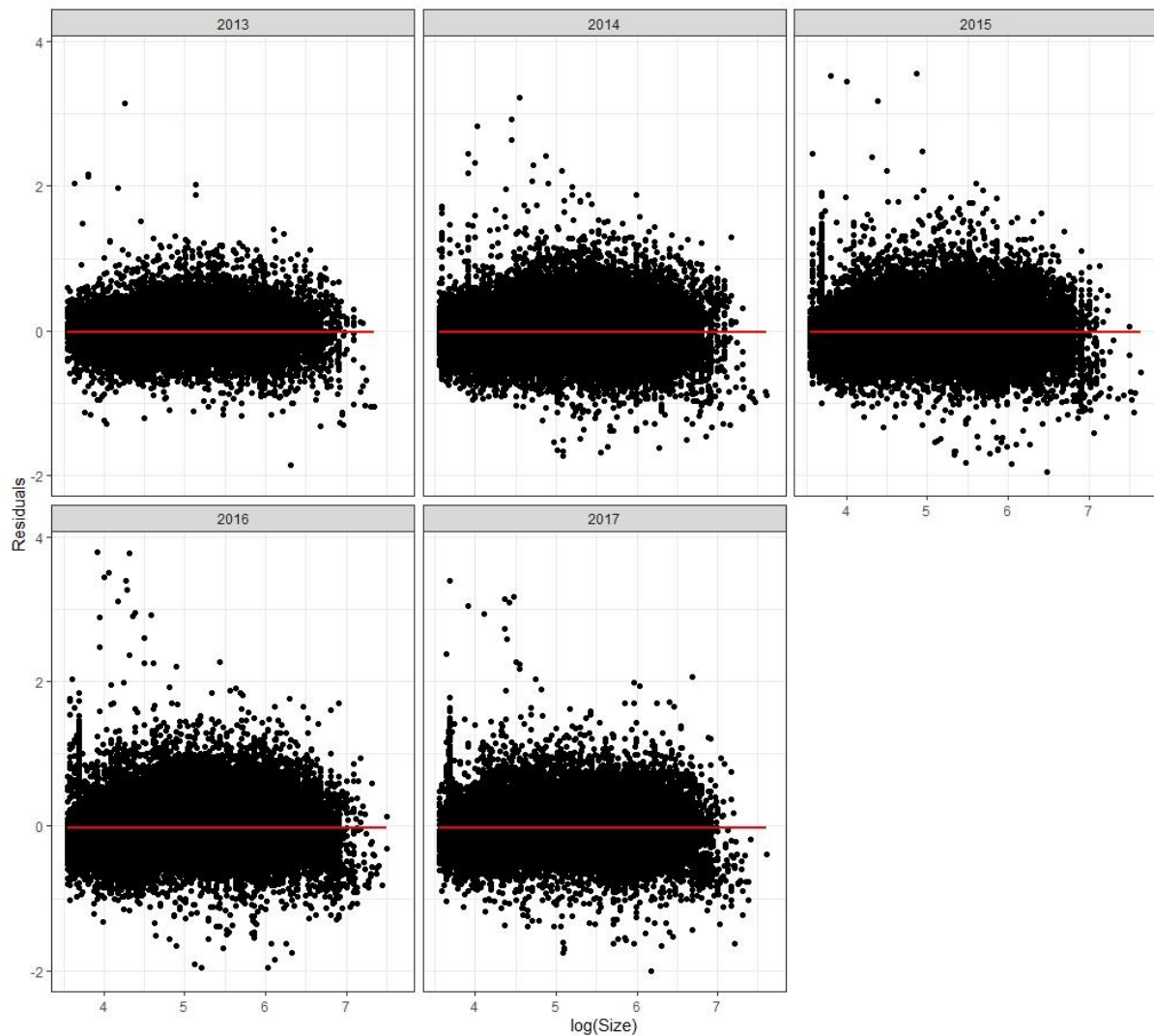


**Figure 3.3:** GLM Coefficients

The yearly coefficients show that townhouses have grown almost linearly year on year compared to the reference level (apartments). Similarly, for simplexes, duplexes, houses, complexes and bedrooms with the exception of 2016 to 2017 where some overlap is evident.

Plotting the residuals against individual covariates of the linear predictor should result in a null pattern, like the residual versus fitted values plot (Mc Cullagh and Nelder, 1989). It is for this reason that the natural logarithm was applied to the size covariate.

Figure 3.4 illustrates the relationship between the deviance residuals and the natural log of the size covariate for each yearly gamma model.



**Figure 3.4:** Gamma model residuals against transformed size covariate

Evident from Figure 3.4 is generally a null pattern, which was achieved by transforming the size covariate.

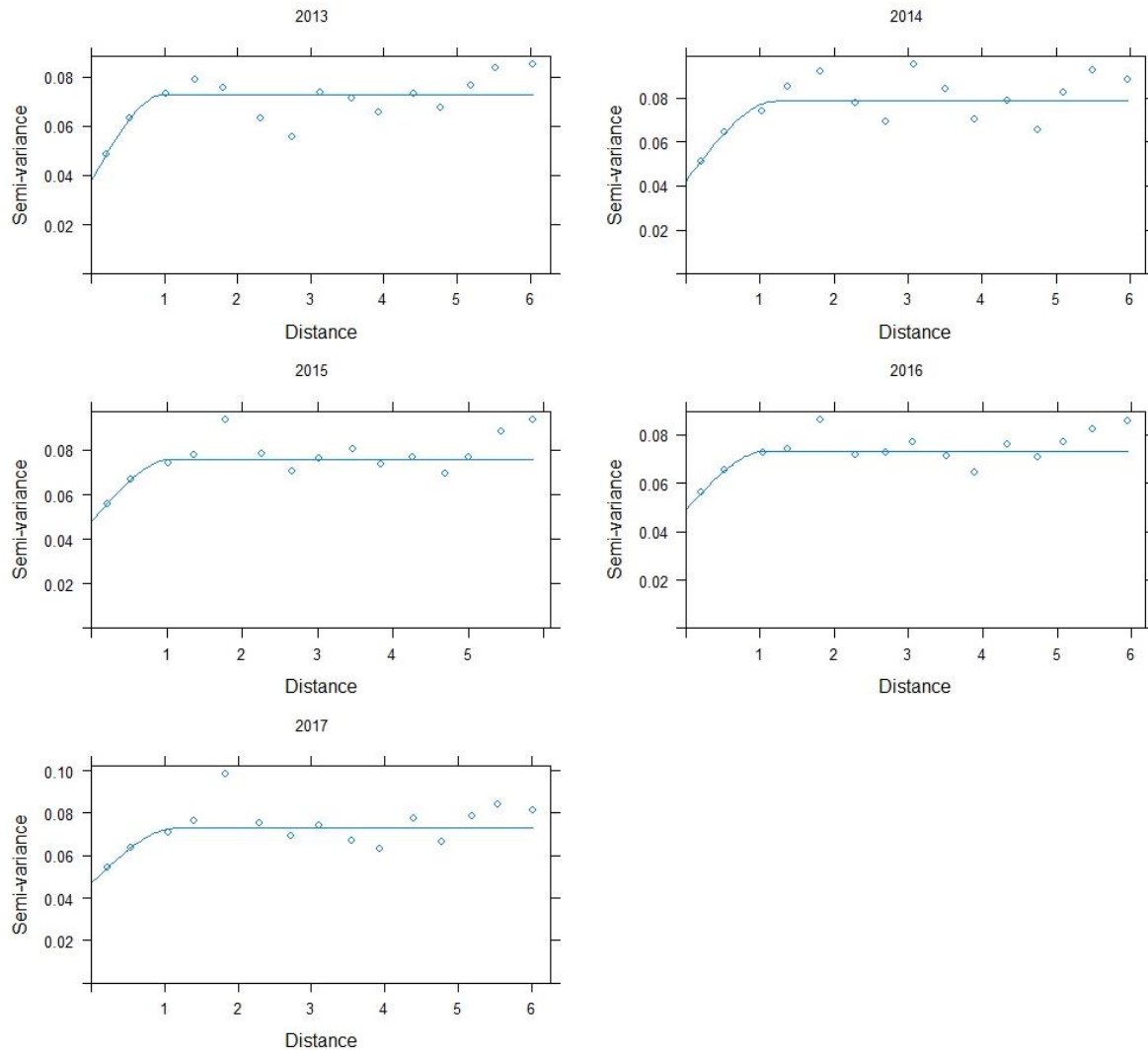
The analysis of deviance presented in Table 3.3 indicates that the residual deviance for each yearly gamma model was consistently lower than the null deviance. This means that the covariates accounted for greater deviance explained than intercept only models and as such, indicates a good fit.

**Table 3.3:** Analysis of deviance

| Year | Residual Deviance | Null Deviance |
|------|-------------------|---------------|
| 2013 | 1470.18           | 13394.60      |
| 2014 | 4056.69           | 31487.73      |
| 2015 | 4448.07           | 33272.34      |
| 2016 | 6030.46           | 45648.77      |
| 2017 | 4008.25           | 32260.69      |

The modelling of spatial data in this study required the assessment of the assumption of independence, which was investigated using several plots and by performing hypothesis tests. Variograms quantify the spatial dependence in data by describing the spatial variance. The yearly gamma models' residuals were plotted using spherical variograms and are presented in Figure 3.4 where similarities were found between all models. The ranges, distances beyond which the data are no longer correlated, are quite long which suggests spatial autocorrelation is not an issue in the modelling results. The nugget effects as a percentage of the total sills are quite large which could indicate some variation at a small scale.





**Figure 3.5:** Gamma model variogram plots

A permutation test for Moran's  $I$  for the given weighting spatial scheme was applied to formally test for the presence of spatial autocorrelation where under the null hypothesis, the data is randomly dispersed. The Moran's  $I$  statistic or correlation coefficient ranges between  $-1$  and  $1$ , where  $-1$  shows perfect negative spatial autocorrelation and  $1$  shows perfect positive spatial autocorrelation. Hundreds of permutations were run, 999 in total, for each yearly gamma model. The results of the tests are presented in Table 3.4 which indicate a weak negative correlation. Formally, at an alpha of  $0.05$  there is not enough evidence to reject the null hypothesis of no spatial autocorrelation for each yearly gamma model. This coincides with the findings of Bourassa *et al* (2007) where the addition of a location dummy variable accounted for spatial dependence adequately.

**Table 3.4:** GLM permutation test for Moran's I

| Year | Statistic | p-value |
|------|-----------|---------|
| 2013 | -0.0312   | 0.999   |
| 2014 | -0.0267   | 0.999   |
| 2015 | -0.0129   | 0.999   |
| 2016 | -0.0207   | 0.999   |
| 2017 | -0.0345   | 0.999   |

### 3.3 Summary

This chapter investigated generalised linear models as an alternative to log linear models to develop hedonic price functions to estimate residential property listing prices in South Africa from January 2013 to August 2017. The gamma generalised linear model provided the best fit and good generalisability whilst keeping the expected values on the original scale, which is an appealing alternative to log linear models. The spatial dependence of residential properties was effectively accounted for by including a suburb factor variable, supported by variograms and Moran's I tests, showing no evidence to reject the null hypothesis of no spatial autocorrelation. This framework provides property market participants with the ability to parsimoniously and transparently quantify the utility derived over the marginal distribution of the physical characteristics of properties. Although linear models are transparent, relaxing functional form assumptions using generalised additive models could provide a better goodness of fit and are investigated in the next chapter. Furthermore, the hierarchical structure of homes is exploited using random effects which is useful to produce smaller standard errors around suburbs with smaller sample sizes.

## Chapter Four

### Hierarchical Generalised Additive Models

This chapter extends the generalised linear model framework presented in chapter three. Relaxing functional form assumptions, hierarchical generalised additive models are introduced, treating covariates as smooth functions along with the treatment of the longitudes and latitudes as isotropic bivariate smooths. Exploiting the hierarchical structure of homes to model the spatial heterogeneity can be accomplished using hierarchical models which is considered as an alternative to modelling the geospatial data as a bivariate smooth function.

#### 4.1 Model Description and Motivation

Hastie and Tibshirani (1986) introduced generalised additive models as a flexible alternative to linear models, the latter taking the form:

$$\eta(\mu) = X\beta,$$

where  $\beta$  is  $p < n$  unknown parameters and the matrix  $X_{n \times p} = [X_1^T, \dots, X_n^T]^T$  is a set of known independent variables (the model matrix), and  $X\beta$  is the linear structure with  $\eta(\cdot)$  as a smooth function of the mean (Lindsey, 1997). Generalised additive models are an extension of generalised linear models, which define smooth functional relationships rather than linear functional relationships, where the smoothness may be established automatically (Maindonald, 2010; Wood, 2017). Generalised additive models can be viewed as semi-parametric generalised linear models, where the linear predictor depends on unknown smooth functions, given by:

$$g(\mu_i) = \beta_0 + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + \dots + f_k(x_{ki}), \quad (4.1)$$

where  $g(\cdot)$  denotes the link function and  $\mu_i \equiv E(y_i)$  with  $y_i$  belonging to an exponential family distribution and  $f_j(\cdot)$  are unknown smooth functions of covariates (Wood, 2006). Therefore, like generalised linear models, generalised additive models can account for different distributions including normal, however, the shape of the relationship between the response and covariates is not specified by some explicit functional form, but rather described through non-parametric smoothers. This makes

generalised additive models' useful for relationships that exhibit complex shapes, that can then be smoothed. Considering a simple model with one smooth function of one covariate:

$$y_i = f(x_i) + \epsilon_i, \quad (4.2)$$

Expressing  $f(x_i)$  so that (4.3) becomes a linear model, is achieved by choosing a basis. The basis defines the space of functions of which  $f$  is an element (Wood, 2006). The smoothing term is a spline function of the covariate. The treatment of covariates as smoothed functions can be extended to interact with parametric terms, taking the form:

$$y_i = f(t_i)x_i,$$

where  $t_i$  represents a covariate and  $x_i$  is a factor variable. Each element of the  $i$ th row of the model matrix for  $f(t_i)$ , is multiplied by  $x_i$  for each  $i$ . This results in a linear regression coefficient for  $x_i$ , varying smoothly by  $t_i$ . Expanding further on how smooths functions are defined and penalised, each smooth function denoted  $f_j$  is delineated by a sum  $K$  simpler, fixed basis functions donated  $b_{j,k}$ , which are then multiplied by estimated corresponding coefficients, leading to the expression:

$$f_j(x_j) = \sum_{k=1}^K \beta_{jk} b_{jk}(x_j),$$

where  $K$  is the basis complexity or size and determines how complex each smoother is. Note that overfitting is voided by imposing a smoothing penalty, which prevents excess wiggleness, through regulating the basis function coefficients (Pedersen *et al*, 2019). Generalised cross validation (GCV) and restricted maximum likelihood (REML) are techniques aimed at controlling the wiggleness of the smooths. GCV selects asymptotically optimal smoothing parameters, with regards to low prediction error (Wood, 2011). However, this procedure comes at the cost of slower convergence of smoothing parameters to their optimal values. Furthermore, given finite sample sizes Reiss and Ogden (2009) show that GCV is more likely to give variable smoothing

parameters, with multiple minima and tend to over fit (smooths are wigglier than they should be). In contrast, they show that REML tends to penalise overfit more sternly, leading to stronger optima and less variability of smoothing parameters. Based on these assertions, the REML procedure is consequently implemented in this study.

One-dimensional smooths models can be extended to two or more dimensions. However, isotropic smooths should be used when covariates are on the same unit of measurement, and scale invariant smooths should be used when this is not the case (Wood, 2006). Isotropic smooths assume that a one-unit change in one variable is equivalent to a one-unit change in another variable. Tensor product smooths are more appropriate when this is not the case, which generalises to using a lattice of bendy strips with different flexibility in different directions. This study consequently uses isotropic smooths to model the effects of structural characteristics of properties on listing prices.

Extending this framework to allow smooth functional relationships between covariates and a response to vary between groups, pooling the functions towards a common shape, can be achieved through a hierarchical generalised linear model (HGAM) (Pedersen *et al*, 2019). Often it is of interest to model between-group variability, hierarchical models impose structure where the relationships between the covariates and the response may differ between groups. This framework allows the intercept or slope, or both, to be subject to grouping. The notation presented in (4.1) extends to include random effects, resulting in

$$g(\mu_i) = \beta_0 + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + \dots + f_k(x_{ki}) + Z_i b_i,$$

where  $Z_i b_i$  relates to the random component. What distinguishes hierarchical models from classical regression, is the modelling of the variation between groups. In fixed effects models, only the error term is random, however hierarchical models, also referred to as mixed effects models or multilevel models, introduces additional sources of random variation, providing a structure of group specific profiles, called random effects. Random effects involve shrinkage, taking data from all the groups to estimate the mean and variance of the global distribution of group means, which can lead to smaller standard errors around group means (Gelman and Hill, 2007). For groups with

small sample sizes, shrinkage is the strongest, where partial pooling by the model improves estimates. This allows the researcher to investigate whether such relationships differ or hold, across groups. The implementation of HGAMS's in this study are constructed using penalised regression splines, proposed by Wood and Augustin (2002), which enjoys several benefits, including lower computational cost, model selection and multi-dimensional smooths.

Two modelling approaches are adopted in this study, both of which use HGAM's, due to the spatially clustered nature of properties. The first model, hereafter referred to as the coordinates HGAM, attempts to deal with the spatial dependence of listing prices, by modelling the suburb latitude and longitude coordinates as an isotropic bivariate function using smooths on the sphere, which is analogous to second order thin plate splines in two-dimensional space, proposed by Wendelberger (1981). The second model, hereafter referred to as the suburb HGAM, does not use the suburb latitude and longitude coordinates, but rather treats the suburb as a random effect with varying intercepts. Individual suburb variation is likely to be present, with some properties having relatively higher prices for their location, and others having relatively lower prices. These individual differences can be modelled by assuming different random intercepts for each suburb. Treating the suburb variable as a random effect is applied to account for variation between suburbs, leveraging partial pooling, and to account for the spatial dependency of listing prices. In the case of this research, these random effects can be considered an approximated weighted average of the mean of the observations in the suburbs and the overall mean of South Africa. The amount of information in a given suburb determines the weight as such, for suburbs with little information, estimates will tend to equal the global mean and large suburbs will tend to be "unpooled", using fewer effective degrees of freedom. Gelman and Hill (2007). Zuur, *et al* (2009) propose hierarchical models to resolve non-independencies in data, making this approach a good candidate to account for the spatial dependency. Goodness of fit diagnostic plots for generalised additive models are like generalised linear models with respect to distributional assumptions (Augustin *et al*, 2012). However, checking the basis dimensions used for smooth terms is important to ensure over smoothing is not present, which can be caused by small basis dimensions. Wood (2017) suggests that the exact choice of the basis dimension is not critical, however,

it should be large enough to ensure that there are sufficient degrees of freedom to represent the underlying data adequately. Though a large basis dimension comes at the cost of increased computation, increasing the basis dimension and looking for important statistically significant changes as a result thereof, could be a useful method for finding a good basis dimension.

Model evaluation is performed by investigating model generalisability to out of sample data, comparing several goodness of fit statistics and assessing residual diagnostic plots, including variograms to investigate possible spatial autocorrelation, similar to how the GLM's were diagnosed.

## 4.2 Partitioning Around the Medoids

Although properties are clustered within suburbs, exploratory analysis may reveal additional spatial groupings, which could add predictive power to modelling efforts. To investigate this, great circle distance, the shortest distance between points on a sphere, was calculated between each suburb for each yearly cross-sectional dataset, respectively. The distance matrices were then passed to the partitioning around medoids (PAM) algorithm, to find spatially homogeneous groups for each respective dataset. Schubert and Rousseeuw (2019) provide a succinct mathematical explanation of how PAM works, which is described in the next few sentences along with the pseudocode. The total deviation (TD), the absolute error criterion, as the objective is given by:

$$TD := \sum_{i=1}^k \sum_{x_j \in C_i} d(x_j, m_i),$$

which is the sum of dissimilarities of each data point  $x_j \in C_i$  to the medoid  $m_i$  of its cluster. The medoid of set  $C$  is the object with the smallest sum of dissimilarities to all other points in the set given by:

$$medoid(C) := \underset{x_j \in C}{argmin} \sum_{x_j \in C_i} d(x_i, x_j),$$

PAM searches for  $k$  representative groups amongst the observations and is robust in the sense that it minimizes the sum dissimilarities instead of the sum of squares. The pseudocode for the PAM algorithm was generated with LaTeX in TeXworks.

---

**Algorithm 1** PAM BUILD: Find Initial Centers

---

```

1:  $(TD, m_1) \leftarrow (\infty, \text{null});$ 
2: foreach  $x_i$  do
3:    $TD_j \leftarrow 0;$  // first medoid
4:   foreach  $x_0 \neq x_j$  do  $TD_j \leftarrow TD_j + d(x_0, x_j);$ 
5:   if  $TD_j < TD$  then  $(TD, m_1) \leftarrow (TD_j, x_j);$  // smallest distance sum
6: for  $i = 1 \dots k - 1$  do // other medoids
7:    $(\Delta TD^*, x_i^*) \leftarrow (\infty, \text{null});$ 
8:   foreach  $x_j \notin \{m_1, \dots, m_i\}$  do
9:      $\Delta TD \leftarrow 0;$ 
10:    foreach  $x_0 \notin \{m_1, \dots, m_i, x_j\}$  do
11:       $\delta \leftarrow d(x_0, x_j) - \min_{0 \in 1, \dots, m_i} d(x_0, 0)$ 
12:      if  $\delta < 0$  then  $\Delta TD \leftarrow \Delta TD + \delta;$ 
13:      if  $\Delta TD < \Delta TD^*$  then  $(\Delta TD^*, x^*) \leftarrow (\Delta TD, x_j);$  // best reduction in TD
13:       $(TD, m_{i+1}) \leftarrow (TD + \Delta TD^*, x^*)$ 
14: return  $TD, \{m_1, \dots, m_k\}$ 

```

---



---

**Algorithm 2** PAM SWAP: Iterative Improvement

---

```

1: repeat
2:    $(TD^*, m^*, x^*) \leftarrow (0, \text{null}, \text{null});$ 
3:   foreach  $m_i \in \{m_1, \dots, m_k\}$  do // each medoid
4:     foreach  $x_j \notin \{m_1, \dots, m_k\}$  do // each non medoid
5:        $\Delta TD \leftarrow 0;$ 
6:       foreach  $x_0 \notin \{m_1, \dots, m_k\} \setminus m_i$  do  $\Delta TD \leftarrow \Delta TD + \Delta(x_0, m_i, x_j);$ 
7:       if  $\Delta TD < \Delta TD^*$  then  $(\Delta TD^*, m^*, x^*) \leftarrow (\Delta TD, m_i, x_j);$ 
8:   break loop if  $\Delta TD^* \geq 0;$ 
9:   swap the roles of medoid  $m^*$  and non-medoid  $x^*$ ; // perform best swap
10:   $TD \leftarrow TD + \Delta TD^*$ 
11: return  $TD, M, C;$ 

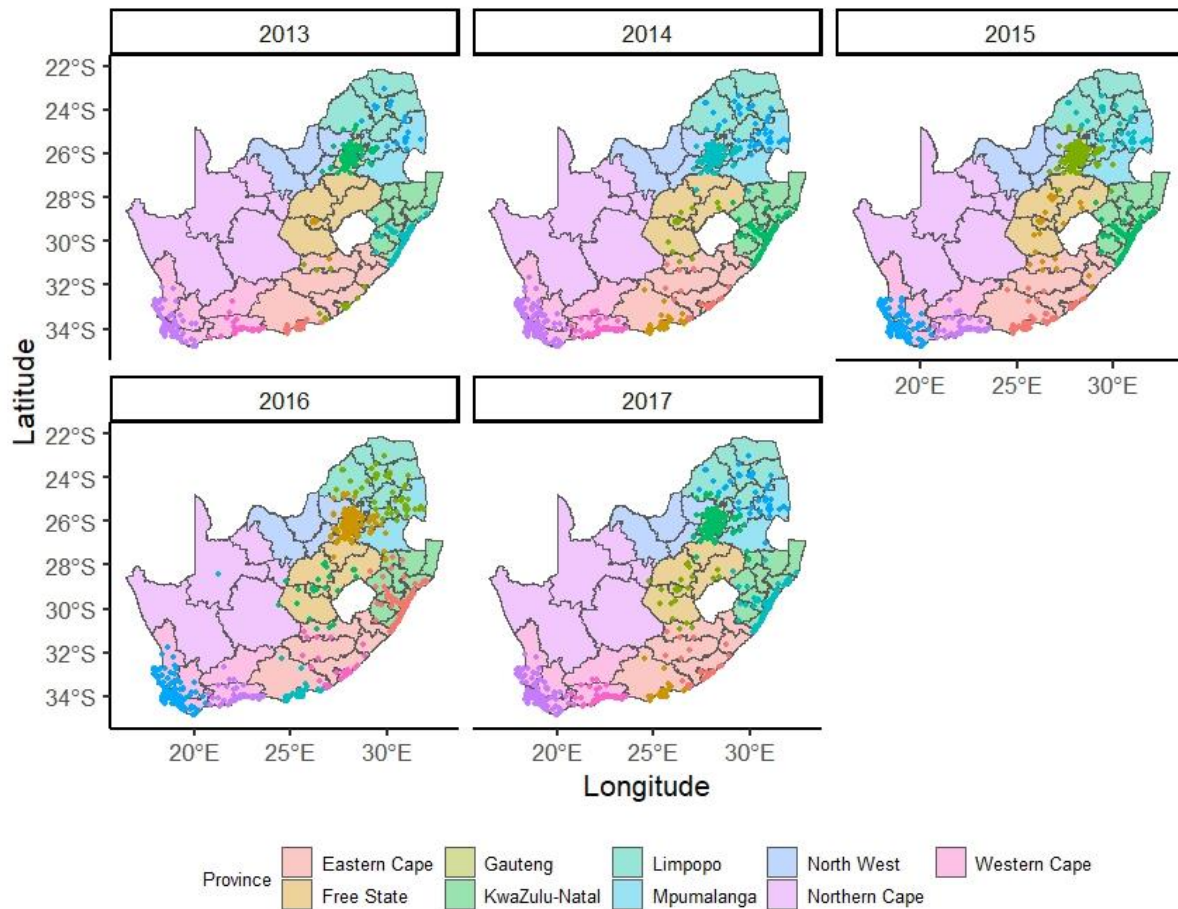
```

---

### 4.3 Results and Discussion

The results of the spatially clustered data are presented in Figure 4.1. Yearly cross-sectional maps of South Africa are illustrated, revealing the distinct or homogenous spatial clusters.

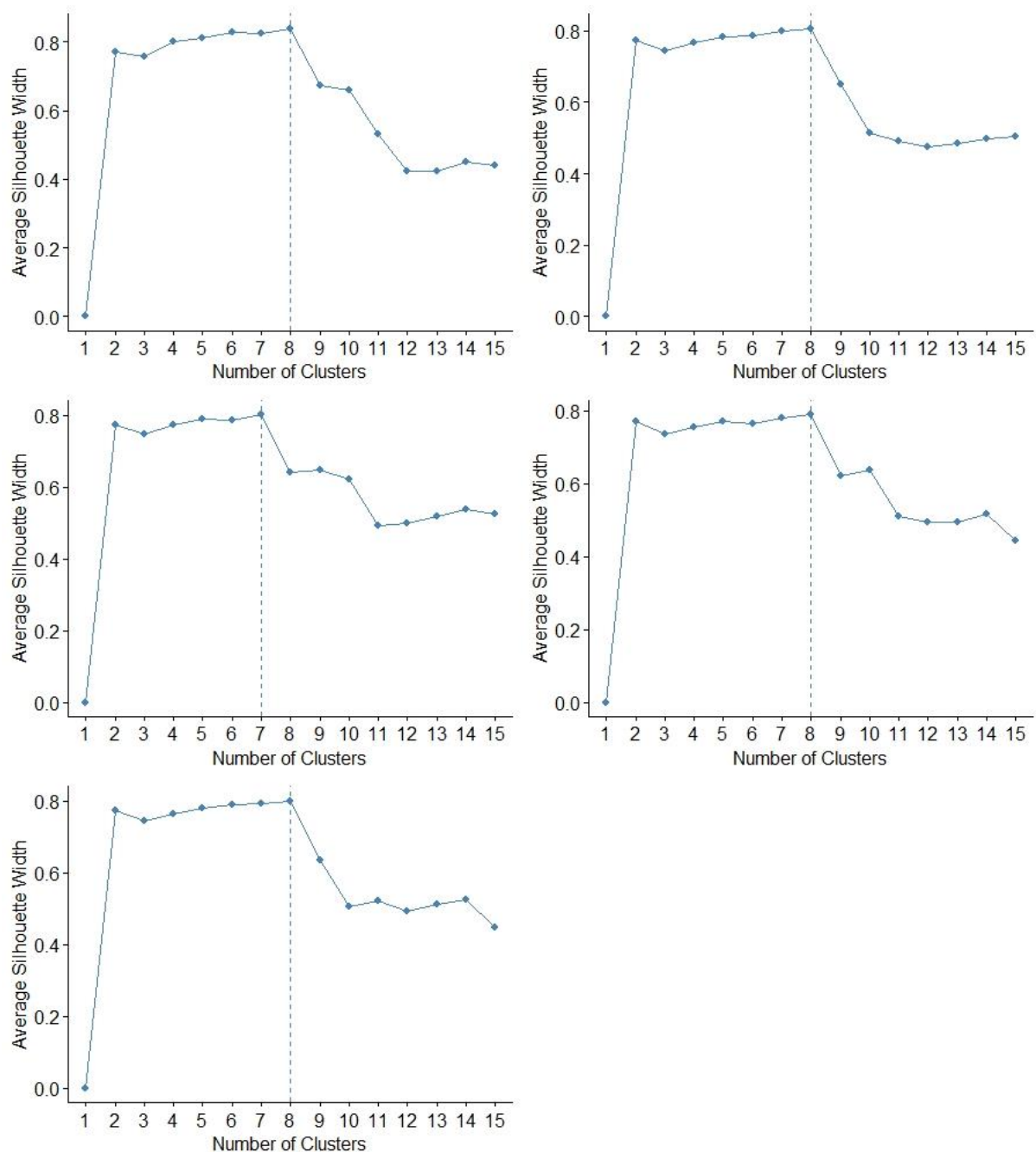




**Figure 4.1:** PAM Spatial clusters

The algorithm consistently split suburbs in the Western Cape province into separate spatial clusters whilst combining suburbs in the Limpopo and Mpumalanga provinces. The average silhouette method, proposed by Rousseeuw (1987), was used to determine the optimal number of clusters which measures how similar a data point is to its membership cluster compared to other clusters. Figure 4.2 depicts the optimal number of clusters for each yearly dataset. The optimal number of clusters seems to be 8 with the exception of 2015 where 7 clusters are optimal. These spatial clusters

will be included as random intercepts in the hedonic models, both the coordinates HGAM's and suburb HGAM's.



**Figure 4.2:** PAM silhouette plots

The basis dimensions for the numeric covariates were kept the same for both the coordinates HGAM's and the suburb HGAM's. An extra penalty was added to each model, meaning the smoothing parameter estimation has the ability to remove terms

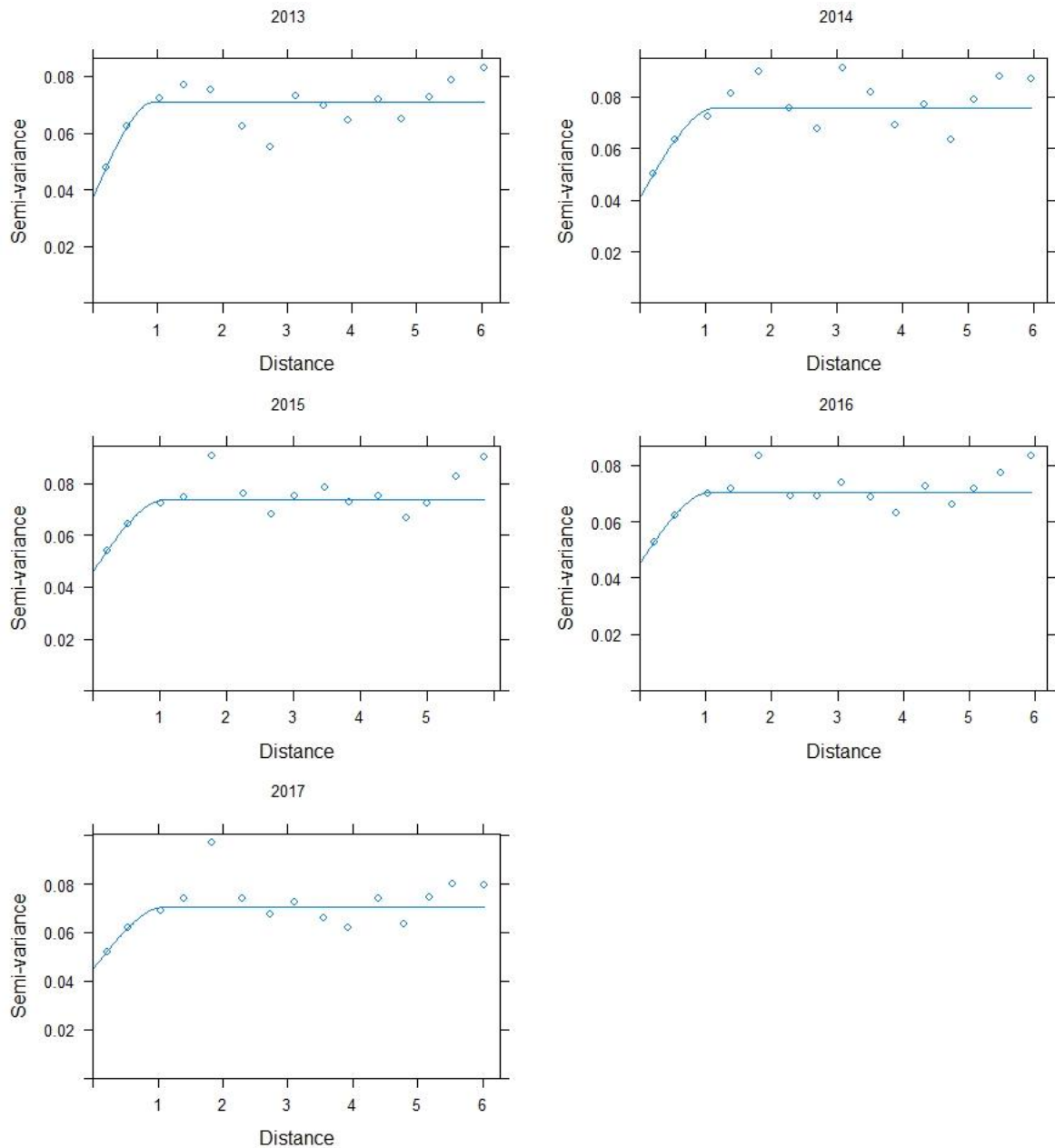
from models. In both approaches, all the covariates were statistically significant, and none were penalised out. A comparison of model performance is presented in Table 4.1.

**Table 4.1:** Results and comparison of HGAM's

| HGAM<br>(Model) | Training RMSE |         | Holdout RMSE |         | $D^2$  |        | AIC     |         |
|-----------------|---------------|---------|--------------|---------|--------|--------|---------|---------|
|                 | Coords        | Suburb  | Coords       | Suburb  | Coords | Suburb | Coords  | Suburb  |
| 2013            | 821 339       | 694 324 | 803 646      | 708 794 | 0.83   | 0.89   | 673527  | 664405  |
| 2014            | 895 816       | 746 601 | 885 449      | 747 555 | 0.80   | 0.87   | 1711345 | 1687867 |
| 2015            | 935 486       | 749 107 | 922 869      | 749 161 | 0.79   | 0.87   | 1834995 | 1806062 |
| 2016            | 894 052       | 702 484 | 900 474      | 720 593 | 0.79   | 0.87   | 2597570 | 2553714 |
| 2017            | 888 836       | 705 273 | 880 368      | 715 071 | 0.80   | 0.88   | 1806229 | 1777514 |

Notes: The deviance explained figures are rounded to two decimal places, all other figures are rounded to the nearest whole number. Coords is the coordinates HGAM.

The suburb HGAM's, which treated the suburb as a random effect, yielded a lower holdout RMSE across each yearly model, in comparison to the coordinates HGAM's. Noticeably, the suburb HGAM's also produced lower AICs and higher deviance explained statistics. The suburb HGAM's outperformed the coordinates HGAM's, based on all goodness of fit statistics. The unexplained spatial heterogeneity was modelled effectively by suburb level random effects. This means that the use of partial pooling produced better estimates in comparison to treating the spatial coordinates as bivariate splines on a sphere. Visualization of the spatial dependency of the suburb HGAM's residuals are presented Figure 4.3, providing a variogram of each model.



**Figure 4.3:** HGAM variogram plots

The distance is calculated using great circle distance, which means that the x-axis units can be interpreted in kilometres. The spatial dependency seems to taper off and flatten after 1km, which is good sign that most of the spatial dependency has been accounted for. The sills are relatively short with long ranges, which suggests that spatial autocorrelation is not an issue.

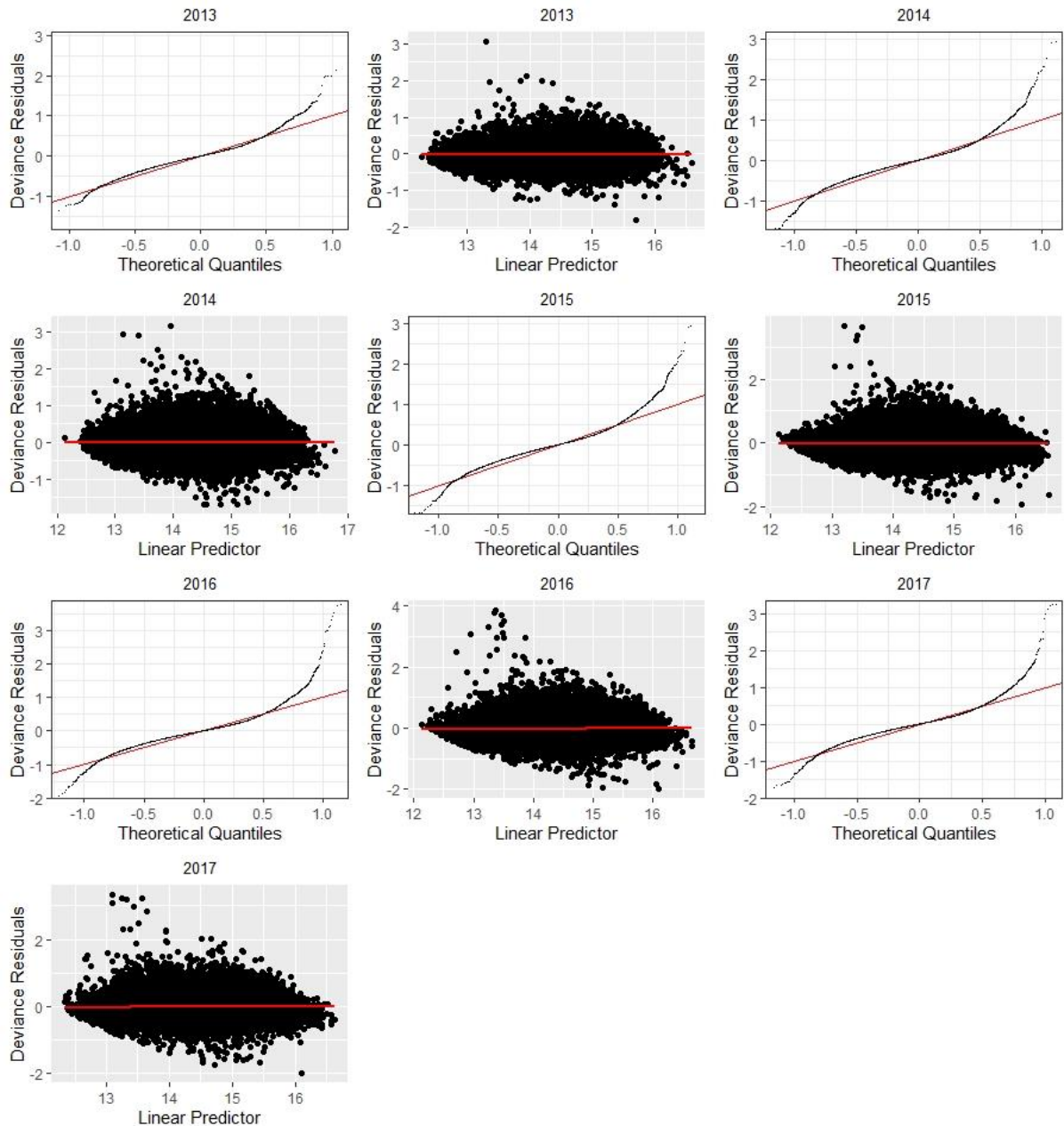
Similar to the GLM's in chapter two, Moran's I tests were performed, and the results are presented in table 4.2.

**Table 4.2:** HGAM permutation test for Moran's I

| Year | Statistic | p-value |
|------|-----------|---------|
| 2013 | -0.0339   | 0.999   |
| 2014 | -0.0275   | 0.999   |
| 2015 | -0.0105   | 0.996   |
| 2016 | -0.0105   | 0.995   |
| 2017 | -0.0323   | 0.999   |

The results of the tests indicate a weak negative correlation. Formally, at an alpha of 0.05 there is not enough evidence to reject the null hypothesis of no spatial autocorrelation for each yearly suburb HGAM.

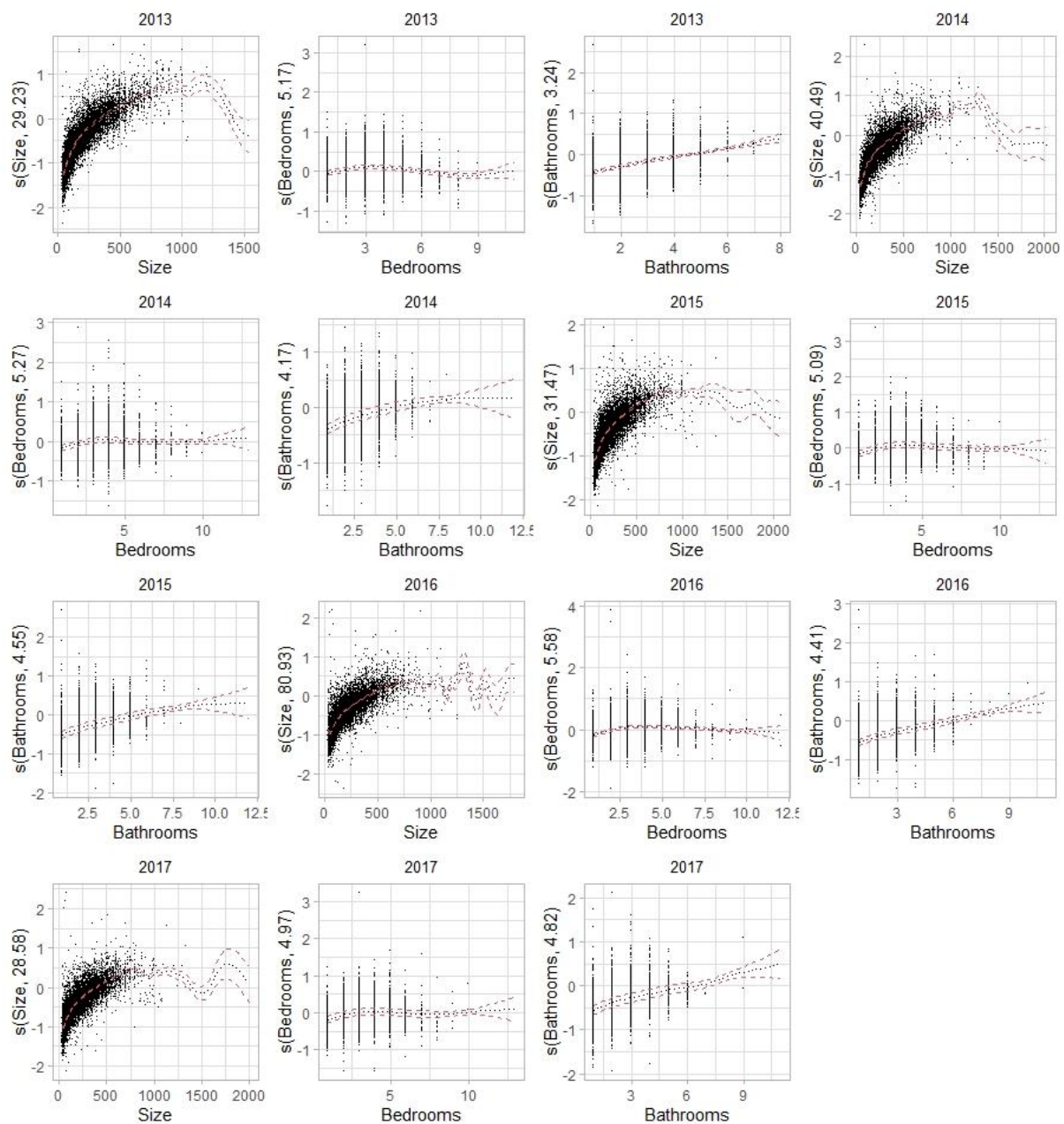
The suburb HGAMs outperformed the coordinates HGAMs based on the deviance explained, AICs and holdout RMSEs. Moreover, the suburb HGAMs account for spatial dependency in the data. Based on these criteria, the suburb HGAMs were selected as the best model fit and discussed further, in more detail. Examining the residual diagnostic plots is an important exercise. Figure 4.4 presents the residuals diagnostic plots for the suburb HGAM's.



**Figure 4.4:** HGAM diagnostic residual plots

No discernible pattern is present in the residual linear predictor plots, indicating that the choice of link function and selection of smooths are appropriate. The Q-Q plots of the yearly models show some departure from normality at the lower and upper quantiles, indicating that the residual distribution is heavy tailed. However, with sufficient sample size this does warrant concern Schmidt and Finan (2018), especially when the departure is not extreme.

Interpreting the smooths requires visualization. The focus of interest is how the covariates are related to the response variable which can be visualized by plotting the component smooth functions that make up the model. Figure 4.5 present the smooth terms visually where smoothed covariates are plotted against the partial residuals on the scale of the linear predictor. The y-axes represent the respective partial residuals, the residuals after removing the effects of other covariates, and the x-axes represent the respective covariates.



**Figure 4.5** Smooth covariate plots



The smooths for each year exhibit non-linear shapes that vary across periods. The marginal utility of the number of bathrooms appears to have a positive effect on listing prices, compared to the number of bedrooms, which is relatively flat. This suggests sellers of homes can command a premium for more bathrooms. The number of bathrooms smooths further shows almost linear relationship with the response. The size covariate exhibits strong non-linearity where the curves increase steeply, then levels off and tapers for larger sized homes. This suggests that the marginal utility does not increase in a linear fashion and sellers cannot command a premium based purely on size, with evidence that a diminishing return exists. Homeowners wishing to determine the listing price that their properties will expeditiously sell for, through the reconciliation of supply demand, can be guided by understanding the implicit value of each marginal characteristic. The smooth functions presented above enable this understanding for each characteristic *ceteris paribus*.

#### **4.4 Summary**

Residential property prices were modelled as a function of physical and locational attributes using HGAMs, a flexible alternative to strictly specified functional form models. Treating the suburb as a random effect with varying random intercepts outperformed modelling the suburb coordinates as a bivariate smooth, using splines on the sphere, whilst accounting for the spatial dependence in the data. Random effects for the suburb covariate captured important group level variation, where partial pooling was useful to capture between suburb variability, improving predictive power. The findings suggest that the hedonic price functions in South Africa are non-linear and that smooth functions are appropriate for estimating the relationship between listing prices and structural property characteristics. The non-linearities between listing prices and property characteristics supports existing economic real estate research where the use of flexible models is better at capturing more realistic hedonic price functions. The goal of the hedonic models is to produce appraisal functions for accurate predictions which is then used to develop the hedonic price index. Given this goal, machine learning methods are investigated in the next chapter to investigate whether lower out of sample errors can be achieved.



## **Chapter Five**

### **Machine Learning Methods**

Previous chapters have investigated statistical models to appraise residential property, these models are designed for inference about the relationships between covariates and a response variable. Machine learning models are not inferential, however, they are designed to make accurate predictions with very little or no assumptions. This is because statistics mathematically models the data generating process to form hypotheses to draw conclusions about a population of interest (Bzdok, *et al*, 2018). Statistics focuses on fitting probabilistic models to compute parameter estimates which describes the population or “true” effect that is unlikely due to noise whereas machine learning focuses on prediction, using learning algorithms to find patterns in data (Bzok, *et al* 2016). The data used for modelling can help determine whether statistical or machine learning methods are more appropriate. Statistics typically deals with long data where the number of observations or subjects is greater than the input variables. In contrast to wide data, machine learning can be useful as it makes very little or no assumptions about the data generating process and can be especially effective without controlled experimental design and when large amounts of complicated non-linear interactions are present (Bzdok, 2017). However, this may come at the cost of having a machine learning model that is harder to interpret or understand but the alternative could be a statistical model with larger uncertainty in parameter estimates and less precision.

Given the large sample size of the data in this study and considering potential non-linear interactions, two tree based machine learning methods and a multi-layer feed forward neural network are investigated in this chapter as hedonic pricing functions to appraise residential property in South Africa. This chapter examines gradient boosted machines (GBM), random forests (RF) and artificial neural networks (ANN) as the candidate methods.

## 5.1 Model Description and Motivation

### 5.1.1 Gradient Boosted Machines

Boosting is a technique of improving a learning algorithm which executes repeated iterations of a weak learner, in the case of GBM, by constructing decision trees sequentially from the residuals (Freund and Schapire, 1996; Friedman, 2001). Therefore, each tree is grown using information from previously grown trees. Boosting seeks to combine performance of iterations of learners, let  $h_1, h_2, \dots, h_T$  represent a set of hypotheses with the composite ensemble hypothesis given by:

$$f(x) = \sum_{i=1}^T \alpha_i h_i(x),$$

where  $\alpha_i$  is the coefficient with which the ensemble  $h_i$  is combined,  $\alpha_i$  and  $h_i$  are learned through the boosting procedure (Meir and Ratsch, 2003). The GBM algorithm learns slowly by fitting a decision tree to the residuals from the model, then adding this new decision tree into the fitted function in order to update the residuals. Importantly, previous trees affect the construction of new trees. The output of the boosted model is given by:

$$\hat{f}(x) = \sum_{b=1}^B \lambda \hat{f}^b(x),$$

where  $\hat{f}$  is improved slowly by fitting more and different shaped trees to the residuals of previous trees using a shrinkage parameter  $\lambda$ , the algorithm is defined below by Hastie, Tibshirani and Friedman (2009).

---

**Algorithm 1** Gradient Tree Boosting Algorithm

---

- 1: Initialise  $f_0(x) = \operatorname{argmin}_{\gamma} \sum_{i=1}^N L(y_i, \gamma)$ .
- 2: For  $m = 1$  to  $M$ :
- 3:     (a) For  $i = 1, 2, \dots, N$  compute

$$r_{im} = \left[ \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_{m-1}}$$

- 3:     (b) Fit a regression tree to the target  $r_{im}$  giving terminal regions

$$R_{jm}, j = 1, 2, \dots, J_m.$$

- 3:     (c) For  $J = 1, 2, \dots, J_m$  compute

$$\gamma_{jm} = \operatorname{argmin}_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma).$$

- 3:     (d) Update  $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$ .

- 4: Output  $\hat{f}(x) = f_M(x)$ .
- 

### 5.1.2 Random Forests

Bagging using bootstrapping to generates  $B$  different training sets, training a model on each training set to get  $\hat{f}^{*b}(x)$  where averaging is applied over the predictions to obtain:

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^{*b}(x),$$

RF employs bagging, however, to decorrelate the trees, at each split in a tree, the algorithm only considers a random sample of  $m$  predictors from the full set of  $p$  predictors (Hastie, Tibshirani and Friedman, 2009). This allows for the trees to use different predictors and strong predictors will not always appear in the top split, meaning the trees will look quite different, reducing the correlation between trees which leads to a reduction in variance. The random forest algorithm in a regression setting from Hastie *et al* (2015) is provided below.

---

**Algorithm 1** Random Forest for Regression

---

1. For  $b = 1$  to  $B$ :
  - (a) Draw a bootstrap sample  $Z^*$  of size  $N$  from the training data.
  - (b) Grow a random-forest tree  $T_B$  to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size  $n_{min}$  is reached.
    - i. Select  $m$  variables at random from the  $p$  variables.
    - ii. Pick the best variable / split-point amongst the  $m$ .
    - iii. Split the node into two daughter nodes.
2. Output the ensemble of trees  $\{T_b\}_1^B$

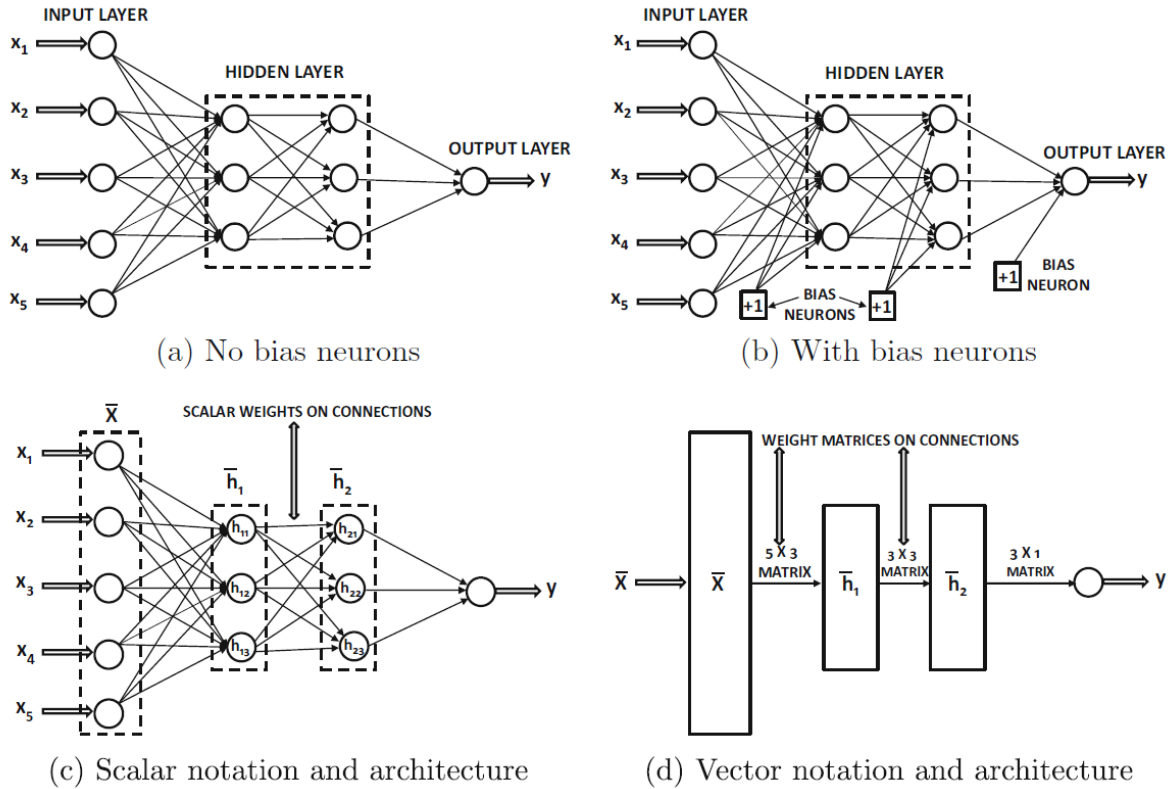
To make a prediction at a new point  $x$ :

$$\hat{f}_{rf}^B(x) = \frac{1}{B} \sum_{b=1}^B T_b(x).$$

---

### 5.1.3 Artificial Neural Networks

Multi-layer feed forward neural networks contain multiple computational (hidden) layers, successively feeding forward into one another from the input layer to the output layer with the typical architecture assuming all nodes in one layer connect to all nodes in the next layer. These computations are not visible to the user which is why they are often referred to as hidden layers. A graphical representation explaining the architecture of a feed-forward ANN network with two hidden layers and one output layer is presented in Figure 5.1 below



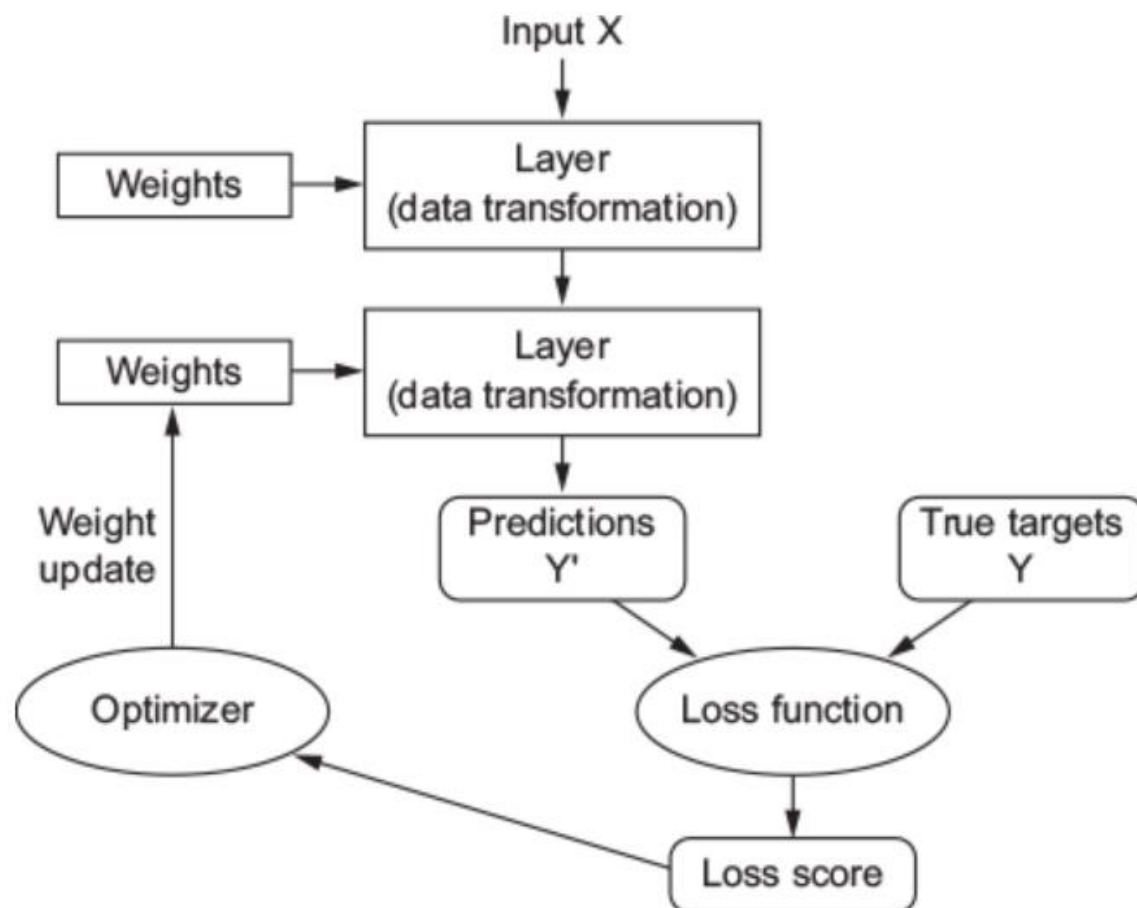
**Figure 5.1.** Example of Feed Forward Artificial Neural Network Architecture

Source: Aggarwal (2018)

Given some values of an input  $x$  the neural network computes an output  $y$ , similar to most other machine learning algorithms. However, ANN use weights, the parameters of a layer, where some transformation at a layer is performed and stored in the weights, simply given by:

$$y = f_w(x),$$

where  $w$  represents the weights. Weights are adjusted using the loss function which measures distance between the outcome variable and predictions which becomes the feedback signal to adjust weights in some direction to lower the loss score. This happens using backpropagation, the optimizer. Figure 5.2 shows the process of how the loss score is used as a feedback signal to adjust the ANN weights.



**Figure 5.2.** Feed Forward Artificial Neural Network Process Diagram

Source: Chollet and Allaire (2018)

Although ensemble, tree-based models often remain better than neural networks for smaller data and / or feature sets such as the case of studies with a small set of predictor variables (Aggarwal, 2018), this chapter explores the use of ANN's to fit residential property appraisal functions.

## 5.2 Model Training and Validation

H2o is highly scalable open-source provider of parallelized machine learning algorithms that are distributed in memory making it a fast and efficient machine learning platform (LeDell *et al*, 2019). GBM, RF and ANN are part of the H2o stack which can be developed using different programming languages such as R and Python or the easy-to-use H2o flow web interface for non-programmers. Gartner (2018), a global research and advisory firm, named H2o a leader amongst 16 vendors in their "Magic Quadrant for Data Science". Key reasons for using the H2o implementation of

GBM, RF and ANN in this study include the ability to fit exponential families of distributions, automatic early stopping based on convergence of a specified metric, the ability to tune many other hyperparameters with cross validation and in the case of GBM, the use of stochastic GBM which improves generalisation through column and row sampling during model training (Click *et al*, 2016; Friedman, 2002). R or Python scripts using the H2o functionality can be embedded into backend on premise or cloud systems for deployment purposes. Alternatively, the final model can be exported as a Java object and embedded into web applications. This makes the H2o implementation of GBM, RF and ANN algorithms portable and interoperable for organisations like property portals.

Cross validation is applied to optimise the hyperparameters, with the aim of reducing the out of sample error. The RMSE is used to test model fit and generalisability. In supervised machine learning problems, model tuning involves finding the optimal hyperparameters for a predictive task. Tuning hyperparameters vary the complexity of models with the aim of finding the values of the tuning parameters that minimise the average prediction error (Hastie, Tibshirani and Friedman, 2001). Searching over a high dimensional hyperparameter space to find the optimal combinations thereof can result in significant computational cost. This is often a drawback of traditional (cartesian) and manual grid searches which can be mitigated by using a random grid search which samples uniformly from the set of all possible hyperparameter value combinations (Bergstra and Bengio, 2012). This study implements a random grid search which allows for early stopping of model building based on convergence of the user supplied training error metric. The findings of Bergstra and Bengio (2012) shows that a random grid search strategy is able to produce models that are at least as good or better than those from manual and traditional grid searches. Zhong *et al* (2018) provide evidence that early stopping is useful in the reduction of the hyperparameter search space in neural network architectures. Early stopping is applied in this study which stops the algorithm if the RMSE does not improve for 25 training rounds based on a moving average of 10000.

Evaluation of model generalisation hyperparameter selection can be achieved using  $k$  fold cross validation. This involves splitting the data into  $k$  roughly equal parts whilst

maintaining the original distribution of the response, Table 5.1 illustrates an example of 5-fold cross validation.

**Table 5.1:** 5-Fold cross validation structure

|                |              |              |              |              |
|----------------|--------------|--------------|--------------|--------------|
| Fold 1         | Fold 2       | Fold 3       | Fold 4       | Fold 5       |
| Validation Set | Training Set | Training Set | Training Set | Training Set |

The procedure involves fitting a model to the training folds and calculating the prediction error on the validation fold which is then repeated for folds  $k = 1, 2, \dots, K$  and finally, combining the  $K$  estimates of prediction error (James *et al*, 2013). Hastie, Tibshirani and Friedman (2001) provide a detailed description which is summarised in the following sentences. Let:  $\kappa : \{1, \dots, N\} \rightarrow \{1, \dots, K\}$  be an indexing function indicating which fold observation  $i$  belongs to from the randomised fold splits. The fitted function is denoted by  $\hat{f}^{-k}(x)$  which is computed with the validation set. This provides a measurement of the cross-validation prediction error, given by:

$$CV(\tilde{f}) = \frac{1}{N} \sum_{i=1}^N L(y_i, \tilde{f}^{-K(i)}(x_i)),$$

Extending this framework to include a set of models  $f(x, \alpha)$  indexed by a tuning parameter  $\alpha$  is given by:

$$CV(\tilde{f}) = \frac{1}{N} \sum_{i=1}^N L(y_i, \tilde{f}^{-K(i)}(x_i, \alpha)),$$

Cross validation can be applied to models with many tuning parameters to search for the combination of hyperparameters that produce the lowest prediction error.  $k$  fold cross-validated gradient boosted machines are built on 80% of the data with 20% withheld as the final validation set, making the generalisation framework robust (LeDell *et al*, 2019). This study implements 5-fold cross validation where yearly property listings from all respective suburbs are randomly blended into the 5-folds, making the cross validation spatially mixed based on the distribution of response.



Typically, when splitting data into training and validation sets or cross validation folds, researchers want validation sets to be independent from training sets, however, spatial data often violates this requirement. The random selection of validation data from the entire spatial domain will result in dependence between training and validation sets because of spatial structure. This leads to overly optimistic error estimates when extrapolating outside the spatial structure. Blocking is an approach designed to remedy this by forcing testing on spatially distant records (Trachsel and Telford 2016). However, if the objective of the model is to interpolate or predict within the same spatial structure, random cross validation or random splitting techniques are reasonable approaches as the model's conditions do not change (Roberts et al., 2017). The models developed in this study are interpolation models, they aim to use the property portals existing data and make predictions on the same spatial structure. Therefore, random data splitting and cross validation techniques are employed.

Partial dependence plots (PDP) and variable importance plots (VIP) are developed for the method with the lowest holdout RMSE to understand the effect of the covariates on the response. PDP are a useful interpretation technique for machine learning algorithms which plot the marginal effect of a covariate on the response holding other covariates constant (Friedman, 2001; Hastie *et al*, 2009), whilst VIP show the relative importance of predictor X (Greenwell and Boehmke, 2020).

Similar to chapters two and three, variograms and Moran I tests are applied to the residuals of the method with the lowest holdout RMSE to test for the presence of spatial autocorrelation.

### **5.3 Results and Discussion**

The results of the 5-fold cross validation yearly machine learning algorithms are presented in Table 5.2 showing the goodness of fit measures.

**Table 5.2:** Cross validation model summaries

|            |      | Fold 1    |                | Fold 2    |                | Fold 3    |                | Fold 4    |                | Fold 5    |                |
|------------|------|-----------|----------------|-----------|----------------|-----------|----------------|-----------|----------------|-----------|----------------|
|            | Year | RMSE      | R <sup>2</sup> | RMSE      | R <sup>2</sup> | RMSE      | R <sup>2</sup> | RMSE      | R <sup>2</sup> | RMSE      | R <sup>2</sup> |
| <b>GBM</b> | 2013 | 668 796   | 0.81           | 736 971   | 0.80           | 696 821   | 0.81           | 688 292   | 0.82           | 711 756   | 0.80           |
|            | 2014 | 716 860   | 0.83           | 705 882   | 0.83           | 744 746   | 0.82           | 710 181   | 0.82           | 708 496   | 0.83           |
|            | 2015 | 700 373   | 0.83           | 734 539   | 0.82           | 712 935   | 0.83           | 706 369   | 0.83           | 702 167   | 0.84           |
|            | 2016 | 664 084   | 0.84           | 659 409   | 0.84           | 651 150   | 0.85           | 665 760   | 0.84           | 659 930   | 0.84           |
|            | 2017 | 670 688   | 0.84           | 680 752   | 0.84           | 673 863   | 0.84           | 686 535   | 0.84           | 661 416   | 0.84           |
| <b>RF</b>  | 2013 | 760 832   | 0.78           | 783 753   | 0.76           | 819 370   | 0.76           | 793 293   | 0.76           | 781 754   | 0.76           |
|            | 2014 | 770 309   | 0.80           | 812 452   | 0.78           | 774 280   | 0.79           | 774 803   | 0.79           | 777 907   | 0.79           |
|            | 2015 | 773 291   | 0.80           | 778 359   | 0.80           | 795 367   | 0.79           | 777 530   | 0.80           | 781 512   | 0.80           |
|            | 2016 | 723 671   | 0.81           | 732 294   | 0.81           | 730 617   | 0.81           | 719 279   | 0.81           | 734 013   | 0.80           |
|            | 2017 | 737 617   | 0.80           | 769 744   | 0.80           | 757 650   | 0.80           | 734 927   | 0.81           | 758 682   | 0.80           |
| <b>ANN</b> | 2013 | 749 092   | 0.78           | 710 762   | 0.79           | 760 460   | 0.79           | 723 314   | 0.79           | 718 852   | 0.79           |
|            | 2014 | 1 090 722 | 0.59           | 1 100 098 | 0.59           | 1 081 611 | 0.60           | 1 057 546 | 0.60           | 1 091 519 | 0.59           |
|            | 2015 | 725 386   | 0.81           | 746 959   | 0.81           | 746 278   | 0.81           | 772 433   | 0.80           | 763 138   | 0.80           |
|            | 2016 | 847 486   | 0.74           | 835 268   | 0.74           | 826 365   | 0.75           | 846 515   | 0.74           | 1 080 526 | 0.58           |
|            | 2017 | 721 689   | 0.81           | 714 548   | 0.82           | 776 842   | 0.80           | 761 511   | 0.8            | 776 171   | 0.79           |

Notes: The R<sup>2</sup> figures are rounded to two decimal places, all other figures are rounded to the nearest whole number.

The out of sample errors in each fold for the yearly GBM and RF respectively are quite consistent with the ANN showing the largest variance, producing the largest out of sample errors in 2014 and 2016. However, the GBM slightly outperform the RF. Combining the holdout predictions to gauge an unbiased overall average fit is presented in Table 5.3.

**Table 5.3:** Combined holdout goodness of fit summaries

| Year | GBM     |                | RF      |                | ANN       |                |
|------|---------|----------------|---------|----------------|-----------|----------------|
|      | RMSE    | R <sup>2</sup> | RMSE    | R <sup>2</sup> | RMSE      | R <sup>2</sup> |
| 2013 | 704 527 | 0.81           | 787 800 | 0.76           | 732 496   | 0.79           |
| 2014 | 717 398 | 0.82           | 781 950 | 0.79           | 1 081 611 | 0.60           |
| 2015 | 711 415 | 0.83           | 781 212 | 0.80           | 772 433   | 0.80           |
| 2016 | 660 072 | 0.84           | 727 975 | 0.81           | 887 232   | 0.71           |
| 2017 | 674 687 | 0.84           | 751 724 | 0.80           | 750 152   | 0.80           |

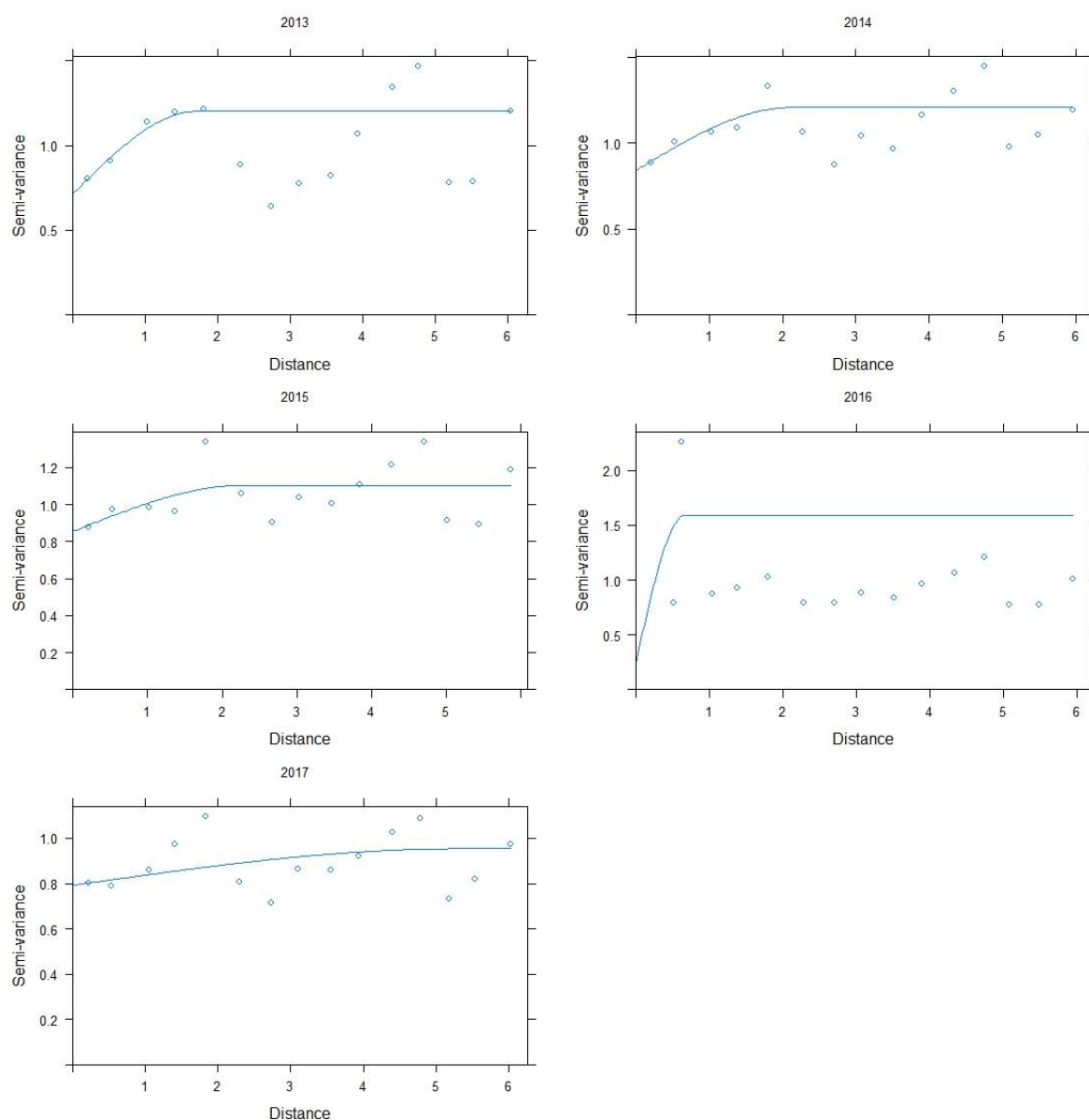
The holdout errors for the tree-based models are consistent for each yearly model respectively with 2016/7 producing the lowest generalisation errors for both GBM and RF. However, the GBM produce the lowest holdout errors for each year and are examined in further detail as the best method. Similar to the findings of Mayer et al (2019) in their hedonic study of Swedish family homes, GBM produced the lowest out of sample error whilst neural networks can be quite erratic.

Table 5.4 examines the spatial dependency in the data by applying the Moran's I test to the residuals of the yearly GBM.

**Table 5.4:** GBM permutation test for Moran's I

| Year | Moran's I Statistic | Moran's I p-value |
|------|---------------------|-------------------|
| 2013 | -0.03534            | 0.999             |
| 2014 | -0.01447            | 0.098             |
| 2015 | -0.01114            | 0.999             |
| 2016 | -0.00784            | 0.992             |
| 2017 | -0.01216            | 0.999             |

The Moran's I test shows that GBM account for the spatial dependency. Formally, at an alpha of 0.05 there is not enough evidence to reject the null hypothesis of no spatial autocorrelation for each yearly GBM. Figure 5.3 presents the scaled variograms of the GBM's.



**Figure 5.3:** GBM Variogram plots

The variograms are calculated using great circle distance, the same as the earlier chapters. The sills are relatively short with varying gradients and long ranges, which suggests that spatial autocorrelation is not an issue.

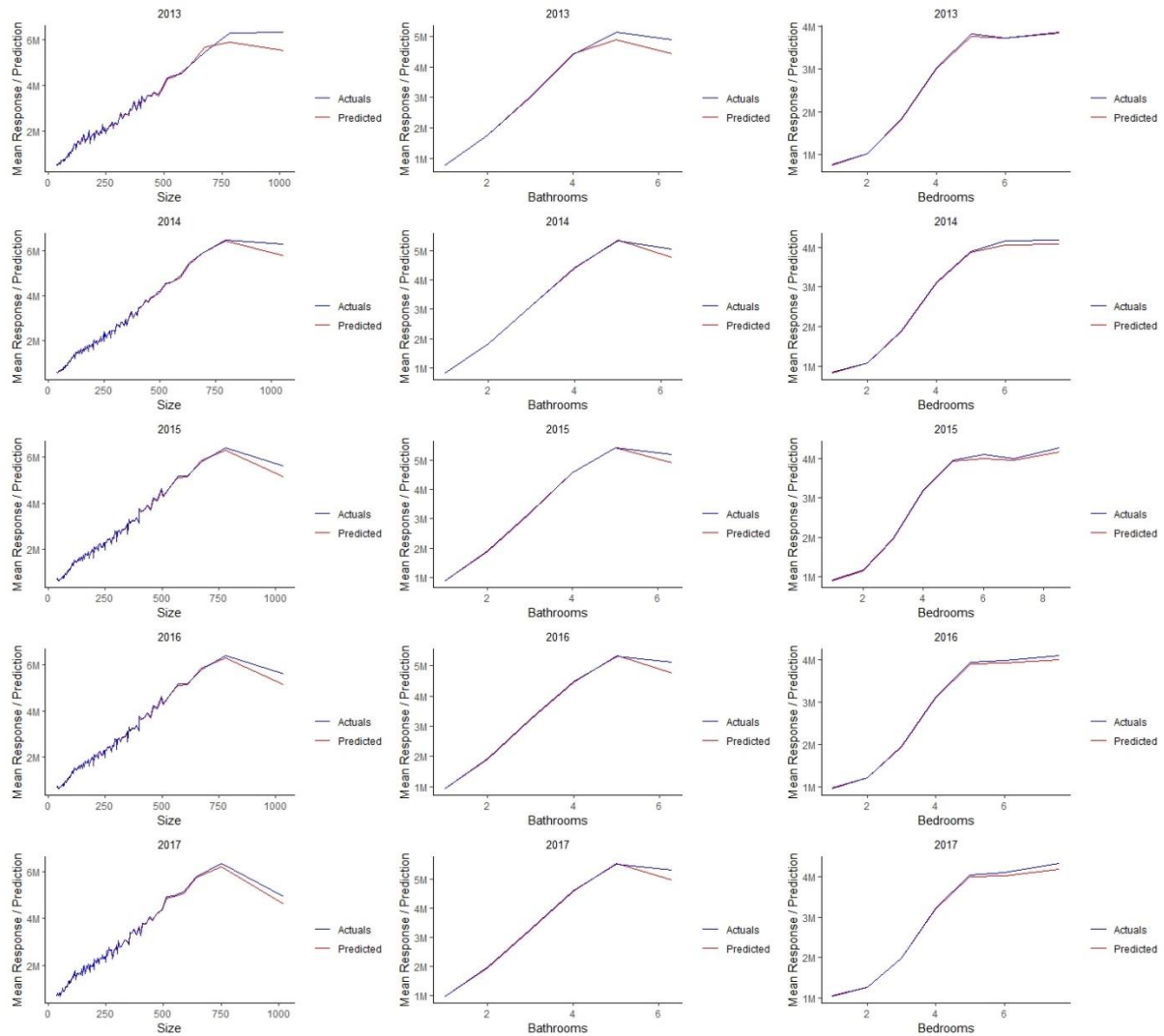
A random grid search applied to each yearly GBM allowed for different hyperparameters to be selected for different the models. Table 5.5 details the hyperparameters chosen in the final models with summary statistics about each model.

**Table 5.5:** GBM model summaries

| Year | number<br>of trees | sample<br>rate | column<br>sample<br>rate<br>per<br>tree | learn<br>rate | min<br>depth | max<br>depth | mean<br>depth | min<br>leaves | max<br>leaves | mean<br>leaves |
|------|--------------------|----------------|---|---------------|--------------|--------------|---------------|---------------|---------------|----------------|
| 2014 | 809                | 0.6            | 0.77                                    | 0.02          | 3            | 19           | 9.47          | 4             | 55            | 28.19          |
| 2015 | 809                | 0.6            | 0.77                                    | 0.02          | 2            | 19           | 9.85          | 4             | 58            | 29.69          |
| 2016 | 809                | 0.6            | 0.77                                    | 0.02          | 3            | 19           | 10.74         | 5             | 81            | 39.94          |
| 2017 | 809                | 0.6            | 0.77                                    | 0.02          | 2            | 19           | 9.63          | 4             | 58            | 29.33          |

The number of trees, sample rate, column sample rate per tree and learning rate hyperparameters were constant for each yearly model. The difference in model complexity is derived from how the individual trees were grown. On average 2016 had deeper and larger trees grown, 2016 also experienced the lowest holdout RMSE. The deeper trees could be attributed to fact that 2016 had substantially more data than other years.

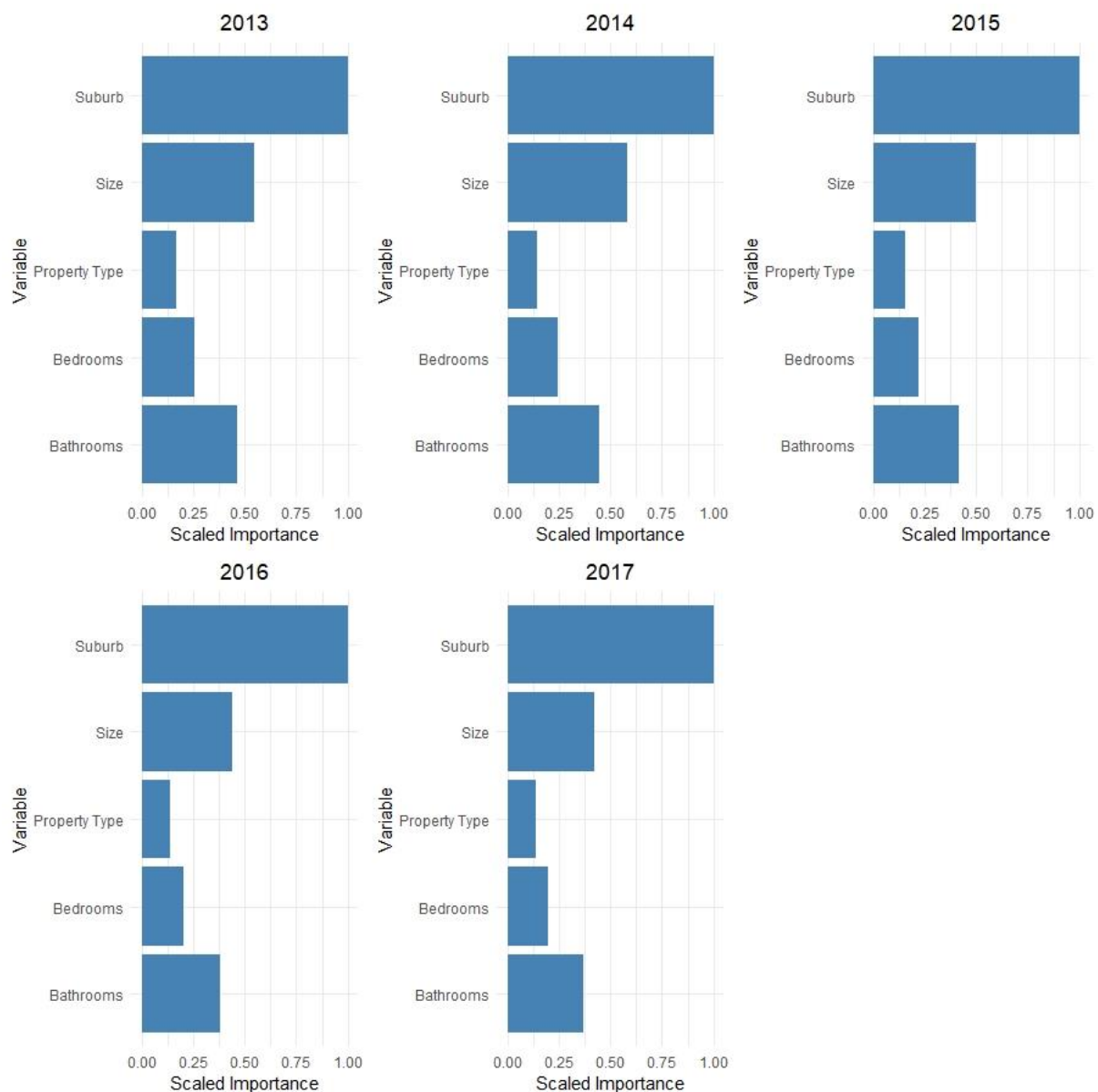
The PDP for each of the numeric covariates are presented next. The implementation of PDP in this study summarise the estimated relationship along with the actual relationship between the response and covariates by showing a calibration curve. A covariate is first grouped into 1% bins where the mean of the predicted outcome and response is calculated holding other covariates constant. Figure 5.4 show how the mean response and predicted outcome changes with a change in the given numeric covariate, namely size, bedrooms and bathrooms.



**Figure 5.4:** Partial dependence calibration plots

The yearly size curves share a similar shape where tapering is evident. The utility increases steeply initially but then drop off over the marginal distribution. This suggests that on average larger sized properties, greater than  $\approx 800$  square meters experience a diminishing return. The marginal utility of bedrooms is positive up to 5 bedrooms. Thereafter, flattening out is evident for properties with an increased number of bedrooms. The number of bathrooms PDP shows that the marginal utility for bathrooms increases steeply up to  $\approx 5$  bathrooms where additional bathrooms add low to no extra marginal utility. The yearly PDP reveal a diminishing return for larger properties, showing that larger homes do not necessarily result in increased prices. Applying the characteristics method proposed by de Haan and Erwin (2011) where separate cross-sectional models were developed provided value in being able to distinguish how the physical characteristics utility curves vary from period to period.

Variable importance is calculated and presented in Figure 5.5. Friedman (2002) applied variable importance to GBM's leveraging the work of Breiman (2001a) who used randomisation of the out-of-bag observations which are observations held back during random forest training and applied before the algorithm has completed.



**Figure 5.5:** Variable importance plots

The suburb a property is located in is the most important predictor of listing price in each yearly GBM. This result coincides with previous hedonic studies which highlight locational effects as statistically significant. The size of the property and number of

bathrooms are consistently deemed the most important physical attributes for each yearly model.

## **5.4 Summary**

This chapter developed and compared yearly hedonic price functions using gradient boosted machines, random forests and multilayer artificial neural networks. An open-source scalable machine learning platform was used to train and tune the 5-fold cross validated where the tree-based algorithms generalised well with consistent out of sample errors and lower variance, compared to the multilayer artificial neural networks. The gradient boosted machines outperformed the random forests and were able to account for the spatial dependency adequately in the data by including a location categorical variable. Developments in making the results of machine learning techniques more transparent has come a long way where the use of partial dependence and variable importance plots reveal the relationships and importance of covariates on the outcome variable. The partial dependence plots showed that the marginal utility for different physical characteristics varied at different quantiles showing that, on average, larger sized properties do not necessarily yield higher prices and can result in diminished returns. The suburb categorical variable was consistently deemed the most important followed by size and the number of bathrooms, reinforcing the old adage about the importance of location of a property. The gradient boosted machines produced the lowest out of sample errors and will be used as the appraisal functions to produce the residential property price index and web application which ensues in the following chapter.



## **Chapter Six**

### **Listing Price Index and Web Application**

Applying index number theory to hedonic models is how property price developments are measured. This chapter deals with the construction of a RPPI for listing prices in South Africa. The listing price index (LPI) and accompanying web application are developed to show how property price developments can be measured and democratised to property market participants.

#### **6.1 Hedonic Model Selection**

Different techniques were applied to develop cross sectional hedonic models for each year in the observation period. Although the log normal and gamma generalised linear models had similar goodness of fit measures, the gamma model was advantageous by keeping the fitted values on the original scale. The gamma distribution was subsequently used in all model fitting. Relaxing functional form assumptions, hierarchical generalised additive models were fitted where penalised regression splines and treating the suburb as a random intercept, improved model generalisability by lowering the out of sample errors in comparison to treating the spatial coordinates as a bivariate smooth. Finally, gradient boosted machines were fitted using 5-fold cross validation, further reducing the out of sample errors and outperforming the random forests and multilayer artificial neural networks. Table 6.1 summarises the performance for the best models of all three methods namely, generalised linear models, hierarchical generalised additive models and machine learning methods.

**Table 6.1:** Hedonic model out of sample errors

| Period | Holdout RMSE (out of sample errors) |             |         |
|--------|-------------------------------------|-------------|---------|
|        | Gamma GLM                           | Suburb HGAM | GBM     |
| 2013   | 719 286                             | 708 794     | 704 527 |
| 2014   | 764 730                             | 747 555     | 717 398 |
| 2015   | 772 905                             | 749 161     | 711 415 |
| 2016   | 746 554                             | 720 593     | 660 072 |
| 2017   | 724 390                             | 715 071     | 674 687 |

Of the three candidate models, gradient boosted machines produced the lowest out of sample errors and are used to develop the LPI and web application.

## 6.2 Index Number Theory

Price indices are widely used in economics and finance as statistical measures of change in a representative set of data points, with several techniques available to calculate hedonic price indices. Typically, properties are not transacted frequently, making price development measurements difficult. However, hedonic models facilitate quality-adjusted price development measurements through the application of index number theory (Hill, 2013). The characteristics method calculates the price change of a typical home from a reference period to a successive period. The counterfactual question being, what is the price change of a set of average home characteristics from a reference period, the first hedonic valuation, to period  $t$ , the second hedonic valuation. In a Laspeyres-type form, the characteristics approach to creating a hedonic RPPI, takes the average characteristics of properties in a reference period and then re-values the same basket of characteristics in successive periods (Silver, 2016). The predicted prices of the average characteristics from the reference period, using a period  $t$  hedonic regression (the numerator), is compared to the predicted prices of the average characteristics from the reference period, using the reference period hedonic regression (the denominator). This results in a constant (reference period) quality price index, answering the question of what the estimated price of a property with the reference period average characteristics would be, if it were on the market in the comparison period  $t$ . A Paasche-type form follows a similar design, except the average characteristics for period  $t$  are used and revalued against the reference period

hedonic regression. The Fisher index would involve taking the geometric mean of the Laspeyres index and Paasche index, which is considered a better approximation (ILO, *et al*, 2004). Medians instead of means are further suggested, where distributions of home characteristics are highly skewed (de Haan and Diewert, 2011). To answer the more granular question of what price property  $i$  in the reference period, given its set of characteristics, would be in period  $t$ , a different technique is necessary. The imputation method achieves this by predicting the reference periods property prices at period  $t$ , using the hedonic regression from time  $t$  (Silver, 2016). An average of these counterfactual predictions is taken and compared with the average of the matched reference period actual prices. Omitted variable bias is likely, which is why dual imputations are recommended (Hill, 2013). The dual imputation method involves running a hedonic regression in the reference period, and revaluing each property and its set of characteristics, using the hedonic regression for period  $t$ . The average of the predicted prices denoted by  $\hat{P}_{i|z_i, k^0}^t$  for period  $t$ , conditioned on the quality-mix in the reference period, is compared to the average of the predicted prices of the reference hedonic regression, denoted by  $\hat{P}_{i|z_i, k^0}^0$ . Dual imputation hedonic price indices to some extent, offset upwards bias, by using predicted values in the numerator and denominator, making it a recommended technique (de Haan, 2004a; de Haan 2009; Diewert *et al*, 2009; Hill and Melser 2008; Hill 2013). Consider a set of properties listed for sale or transacted in the reference period. The dual imputation method facilitates a matched model methodology, where price comparisons are made with the same matched properties in period  $t$ . This results in the measurement of pure price changes by controlling for the quality-mix of properties effectively. The problem of quality-mix is resolved, by imputing the period  $t$  prices for properties in the reference period, answering the counterfactual question, of what a property in the reference period would be valued at, for period  $t$ . An implicit weighting is assigned to each properties price change, using the dual imputation method, namely its relative price in the reference period. A commensurate weight is given to properties relative prices under the imputation method, which is equal to the relative expenditure, an apt measure of the relative weight to attach to property price changes (Silver, 2016). Both the imputation and characteristics techniques are quite intuitive. The imputation method can be thought of as a ratio of average quality adjusted prices of matched homes, whilst the characteristics method can be thought of as a ratio of prices of a constant

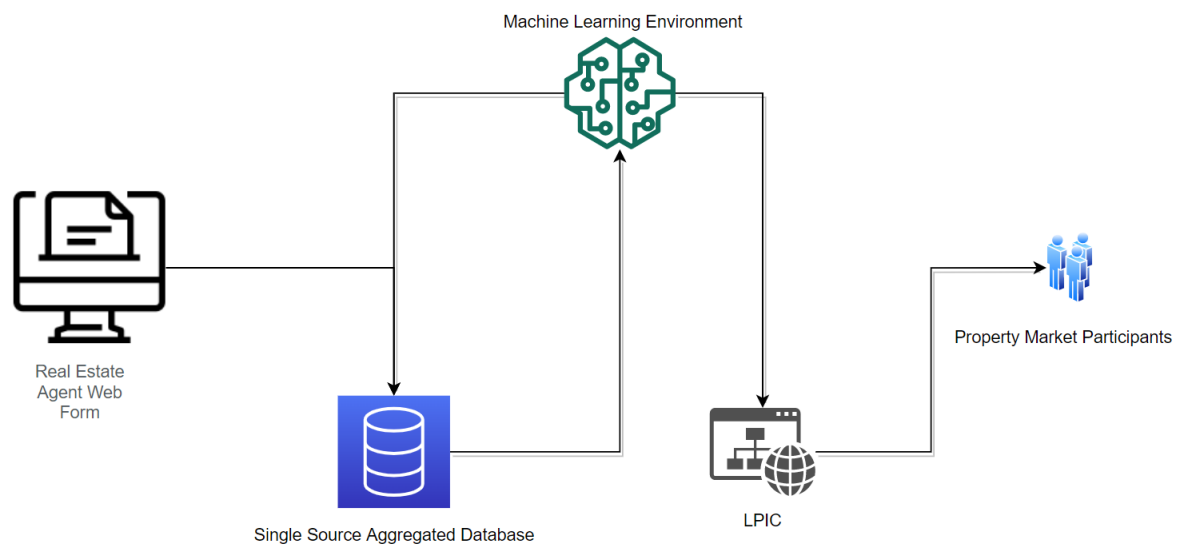
quality bundle of characteristics. Extending price comparisons to multiple periods, beyond the bilateral examples discussed so far, is achievable through various techniques. In a fixed base scenario, an index is measured as the constant quality price changes between each successive period  $t$  and the reference period. Periodic updating and linking can be beneficial when constructing a residential property price index, adjacent period indices use a rolling window, which are rebased each period (Hill, 2013). The window in this approach requires only two periods, the current period and the period prior to it. The window is then shifted to adjacent periods and chained together (Triplett, 2006). A chained index is closer to a theoretical index than a fixed version, and may alleviate substitution bias (Balk, 2008). Dual imputation hedonic models that are updated via chaining, enjoy the benefit of circumventing revisions, a desirable feature in developing price indices. Dual imputations provide a measure of constant quality price changes, weighted by their predicted prices in the reference period for a Laspeyres-type index and predicted prices in period  $t$  for a Paasche-type index. A Fisher type dual imputation index, the geometric mean of the Laspeyres-type and Paasche-type indices, is expressed by:

$$\sqrt{\prod_{h=1}^{H_{t+1}} \left[ \left( \frac{\hat{p}_{t+1,h}(z_{t+1,h})}{\hat{p}_{t,h}(z_{t+1,h})} \right)^{1/H_{t+1}} \right] \times \prod_{h=1}^{H_t} \left[ \left( \frac{\hat{p}_{t+1,h}(z_{t,h})}{\hat{p}_{t,h}(z_{t,h})} \right)^{1/H_t} \right]},$$

where  $\hat{p}_{t+1,h}(z_{t,h})$  denotes the imputed price in the comparison period for individual homes. The LPI developed in this study adopts a chained, dual imputation Fisher index for the gradient boosted hedonic models and compared to a simple median mix-adjusted index and the ABSA global property guide index for South Africa.

### 6.3 Project Process

An architectural diagram of the proposed deployment methodology is presented in Figure 6.1 which shows the flow of data from different environments in order to provide end users with market insights.

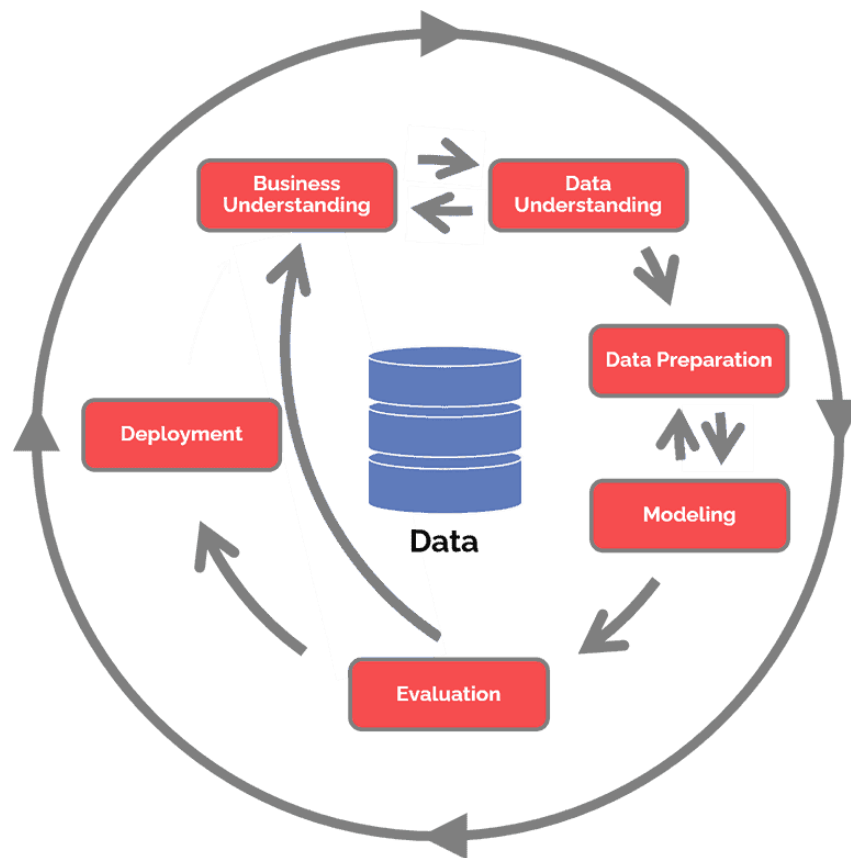


**Figure 6.1:** Deployment architecture

Estate agents populate listings into a web form which constrains inputs to the correct data type, helping ensure data integrity. These individual listings are persisted to a database and ingested into machine learning environment where gradient boosted machines are applied to obtain the hedonic models. The hedonic models are then consumed by the Listing Price Index Calculator (LPIC), a front-end web application, to derive the hedonic prices and index numbers, democratising the insights to property market participants. To make the methodology computationally inexpensive, the machine learning algorithms can be built daily or monthly depending on how often new listing data is persisted to the database. This would only happen for the current years' data and previous models would be serialised or the predictions of those models persisted back to the database. This schematic can apply to property portals or individual real estate agencies. In the case of this study Microsoft SQL sever was used to store and persist data. An SSIS package was developed for the ETL (extract, transform and load) process which consisted of 28 comma separated value files. The open-source statistical programming language R and Shiny package were used to develop the source code for the models and web application. The source code for the web application can be found in the appendix.

The practical implementation of this study, resulting in the proposed architecture in Figure 6.1 would follow the Cross Industry Standard Process for data Mining (CRISP-

DM), first proposed and published by a consortium of companies in 1999 (Chapman, 2000). Figure 6.2 illustrates the stages of CRISP-DM.



**Figure 6.2:** CRISP-DM framework

Source: Hotz (2022)

The first stage involves understanding the business question or what the business wants to achieve. In the case of this research, how Private Property (Pty) Ltd can leverage their data to make residential property appraisal easy, given some characteristics about the home and provide insights about price developments over time to help users of their website make more informed decisions. The data understanding stage involves investigating what data is needed and the suitability thereof. This would have involved the exploratory data analysis performed in chapter two and the ETL process discussed above. The modelling stage asks what modelling methods should be applied based on the exploratory data analysis and business question, and then applies these models. This stage was illustrated by chapters three, four and five where various models were fit to the data. The evaluation stage involves

selecting the best model based on the business objectives. In the case of this study, the model with the lowest-out of sample error which proved to be gradient boosted machines. Finally, the deployment stage which is illustrated in Figure 6.1 would take the form of the LPIC where the users can access the results of the model and methodology.

## 6.4 Results and Discussion

Firstly, a median mix adjusted index is constructed and presented in Table 6.2 using equation (1.1). The stratification was done on the property type and size of the property. Homes were grouped by property type and the 25<sup>th</sup> and 75<sup>th</sup> quantiles were calculated, denoting small and large homes respectively while medium sized homes fell in between. Further post stratification was not possible due to very low sample sizes in each stratum.

**Table 6.2** Median mix adjusted index results

| Median Mix<br>Adjusted<br>Index | 2013-2014 | 2014-2015 | 2015-2016 | 2016-2017 |
|---------------------------------|-----------|-----------|-----------|-----------|
|                                 | 11.74     | 5.52      | -2.02     | 3.01      |

Notes: Figures are nominal and stated as percentages

Compared to the figures in Table 6.3 below, the simple median mix adjusted index is noisy producing larger yearly changes which is why is not favoured in literature.

Nominal or non-inflationary adjusted annual listing price developments are calculated for South Africa and presented in Table 6.3 using the GBM hedonic models.

**Table 6.3:** Nominal South African LPI results

| Fisher Index | 2013-2014 | 2014-2015 | 2015-2016 | 2016-2017 |
|--------------|-----------|-----------|-----------|-----------|
|              | 7.35      | 5.48      | 3.25      | 3.39      |

Notes: Figures are stated as percentages

Although listing prices appear to experience growth for each year in the observation period, a diminishing return is evident. Yearly inflation rate comparisons are presented in Table 6.4 to facilitate an understanding of real listing price developments.

**Table 6.4:** Annual inflation rate comparisons

| Annual Inflation | 2013-2014 | 2014-2015 | 2015-2016 | 2016-2017 |
|------------------|-----------|-----------|-----------|-----------|
|                  | 6.09      | 4.58      | 6.34      | 5.27      |

Notes: Figures are stated as percentages

Source: Inflation rate: South Africa | Statista (2020)

Factoring in the above inflation rates, Table 6.5 provides real listing price developments for South Africa over the observation period.

**Table 6.5:** Inflation adjusted South African LPI results

| Fisher Index | 2013-2014 | 2014-2015 | 2015-2016 | 2016-2017 |
|--------------|-----------|-----------|-----------|-----------|
|              | 1.19      | 0.86      | -2.91     | -1.79     |

Notes: Figures are stated as percentages

After accounting for inflation, annual listing price developments were lackluster from 2013 to 2015, showing negative growth in latter periods. 2016 experienced the largest contraction in listing prices which was a function of low growth and high inflation. No other South African listing price indices have been identified to compare against the one developed in this study. However, Table 6.6 presents the ABSA Global Property Guide annual house price changes on transacted properties.



**Table 6.6:** ABSA global property guide annual house price changes

| Year | Annual House Prices Changes |                    |
|------|-----------------------------|--------------------|
|      | Nominal                     | Inflation-adjusted |
| 2014 | 6.31                        | 0.92               |
| 2015 | 6.24                        | 1.01               |
| 2016 | 4.58                        | -2.33              |
| 2017 | 3.86                        | -0.61              |

Notes: Figures stated as percentages

Source: Delmendo (2020)

Although Table 6.6 uses transaction prices and not listing prices and only considers houses, similarities can be observed. Table 6.6 shows a contraction for inflation adjusted house prices in 2016 and 2017, these results are similar to the finding in this study.

Provincial listing price developments over the observation period shows growth disparity between the nine South African provinces where prices developments are more robust in certain parts of the country. Table 6.7 presents nominal listing price developments broken down to a provincial level.

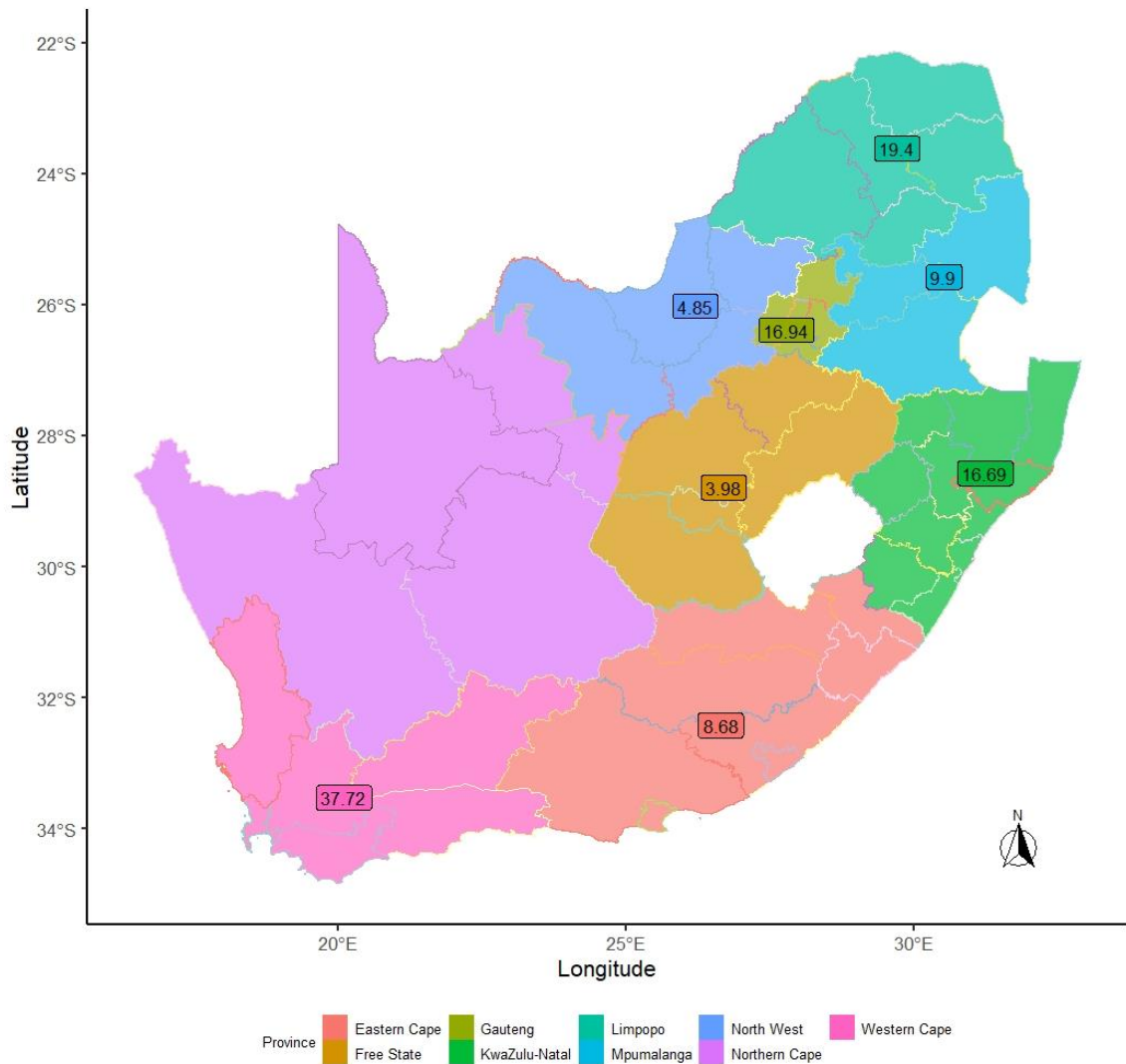
**Table 6.7:** Provincial LPI results

| Province      | 2013-2014 | 2014-2015 | 2015-2016 | 2016-2017 |
|---------------|-----------|-----------|-----------|-----------|
| Eastern Cape  | 1.43      | 3.54      | 0.83      | 1.99      |
| Free State    | -0.42     | 3         | -0.49     | 1.07      |
| Gauteng       | 7.59      | 4.35      | 1.49      | 1.6       |
| KwaZulu-Natal | 3.87      | 4.5       | 3.32      | 3.18      |
| Limpopo       | 6.34      | 2.65      | -0.19     | 6.49      |
| Mpumalanga    | 5.58      | 0.1       | -0.01     | 1.54      |
| North-West    | 2.16      | 3.29      | 2.91      | -2.25     |
| Northern Cape | NA        | NA        | NA        | -19.01    |
| Western Cape  | 9.57      | 8.37      | 6.79      | 6.83      |

Notes: Figures are nominal and stated as percentages

Drilling down into provisional price developments, reveals that the Western Cape experienced the largest growth in listing prices for each year in the observation period. Gauteng showed consistent tapering in listing price developments. The Free State, Limpopo, Mpumalanga, North-West and Northern Cape contracted in various periods with 2015-2016, on average, experiencing more negative effects provincially. This is echoed by examining Table 6.5 where listing price developments were at their lowest for 2015-2016 throughout the observation period.

Figure 6.3 displays the nominal provincial price developments using 2013 as the reference period and 2017 as the comparison period. This is synonymous with a fixed base index where 2013 is the base period.



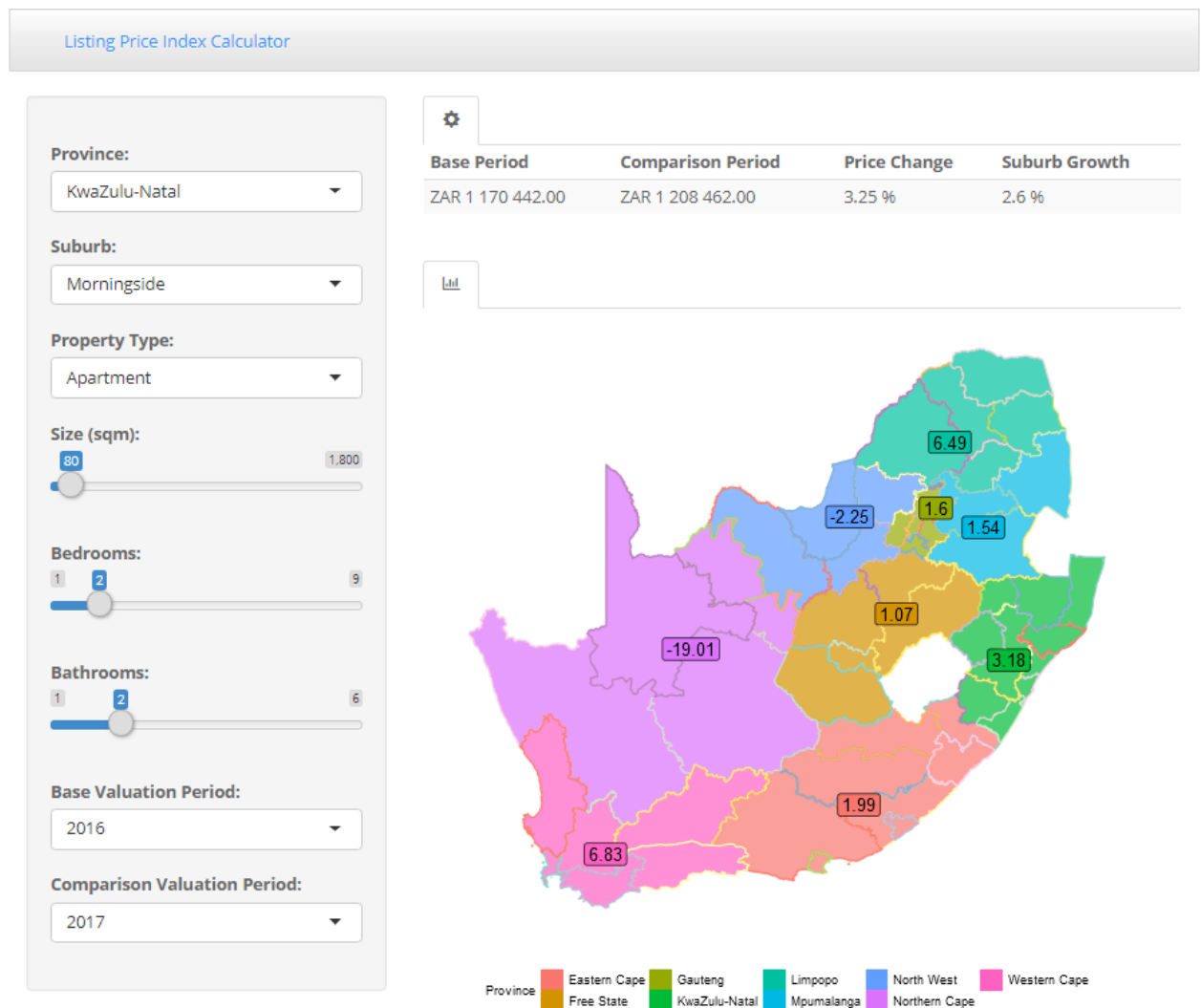
**Figure 6.3:** Provincial listing price growth 2013-2017

The Western Cape achieved the largest growth in listing prices, outperforming Gauteng and KwaZulu-Natal by more than double growth. Limpopo achieved the second highest performance. The Northern Cape only had data in 2016 and 2017 which is why it does not feature in Figure 6.3.

## 6.5 Listing Price Index Web Application

Making listing price development insights consumable to households can be achieved through the creation of a data product. Data products are products that use data and statistical models or machine learning algorithms to democratise insights and add value (Sands, 2018). Some examples include Google search and the Netflix movie

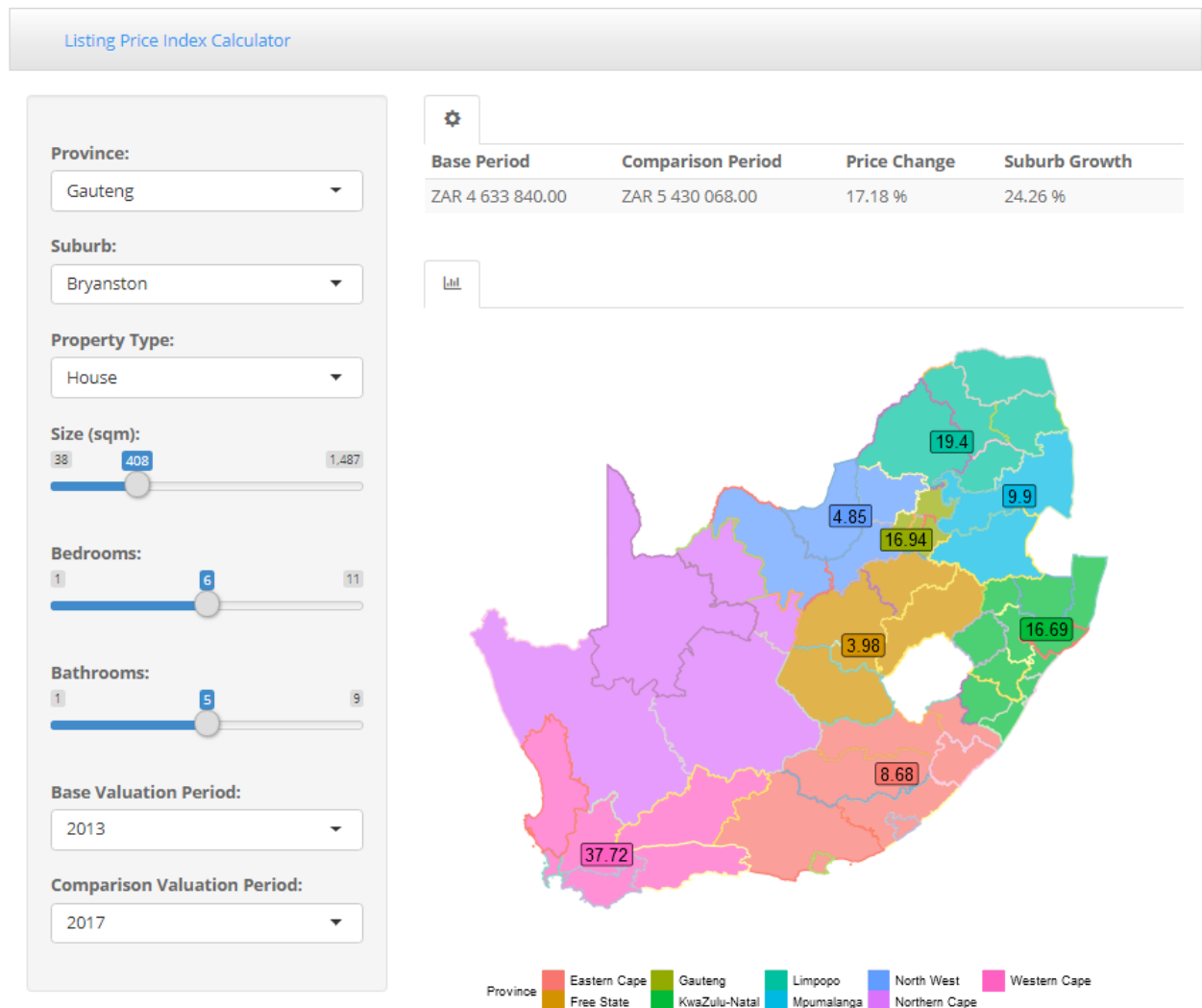
and Amazon product recommenders. The Listing Price Index Calculator (LPIC) is a data product presented in the form of a web application that allows users to easily ascertain listing price developments of individual properties and make comparisons at a suburban and provincial level. The LPIC presents the application of the methodology proposed in this study. Figure 6.4 illustrates an example of a 2-bedroom, 2-bathroom, 80-square meter flat in KwaZulu-Natal Morningside.



**Figure 6.4:** LPIC example one

The LPIC shows that from 2016 to 2017, a property with the above set of characteristics grew by 3.25% whilst properties in the same suburb grew on average by 2.6% and properties in KwaZulu-Natal grew 3.18% on average. This indicates that on average, this property achieved higher listing price growth relative to the suburb and province. The LPIC has a dynamic user interface that allows users to easily update

their selection to obtain the information they want. The LPIC facilitates either a fixed base or chained approach. A second example is presented in 6.5 for a 6-bedroom, 5-bathroom, 408-square meter house in Bryanston, Gauteng.



**Figure 6.5:** LPIC example two

In this case, price developments are examined from 2013 to 2017 where a property with the above set of characteristics, on average experienced a 17.18% increase in listing price whilst properties in Bryanston, on average grew by 24.26% and 16.94% on average in Gauteng. This indicates that although the listing price for this particular property, on average grew more than properties in Gauteng, it did not keep up with the average growth experienced in Bryanston.

For property market participants wishing to sell their homes, the LPIC data product can be a useful first assessment of what to list their homes for on the market, making it a complimentary tool to estate agent valuations. Potential buyers are able to ascertain the typical asking price for a home with the desired set of characteristics in the suburb of interest. Real estate agents and property portals are well positioned to leverage their extensive property market data to develop a similar data product that provides property market participants with a simple, data driven tool to assess and understand price developments.

## **6.6 Summary**

Monitoring residential property price developments is important for various property market participants and policy makers, however, controlling for the quality-mix is critical to ensure pure price changes are measured and not simply changes in the sample composition. Hedonic models are well suited to counter the quality-mix problem, coupled with index number theory, they provide a consistent way to measure quality adjusted price changes. This chapter presented how gradient boosted machines can be used to construct a listing price index for South African homes from January 2013 to August 2017. Although positive nominal growth was experienced, a diminishing return was evident. After accounting for inflation, listing price developments were lackluster and even negative for latter years. This is congruent to real estate publications produced by ABSA. The Western Cape experienced the largest growth throughout the observation period. The framework presented was developed into a data product, the Listing Price Web Application, in the form of a web application, making obtaining listing price developments at a property, suburb and provincial level, a simple selection on a user interface. Such an application is an example of how real estate agents and property portals can leverage the plethora of their available data, to present property market participant with an interactive tool to understand listing price developments. Sellers of homes can use data products like the Listing Price Web Application to get an understanding of what price to list their homes for on the market, making it a data driven solution to compliment real estate listing price valuations.

## **Chapter Seven**

### **Conclusion**

This chapter presents a summary of the study, highlighting the key findings and value achieved from the work. The limitations of the research are also addressed.

#### **7.1 Conclusion**

The aim of this study was to apply statistical and machine learning methods to appraise residential property and measure price developments through the creation of a residential property price index using index number theory. Furthermore, an important objective was to create a web application, the Listing Price Index Calculator, illustrating how this methodology can be used by property portals and real estate agencies to help market participants make informed decisions regarding the sale or purchase of a home of interest.

Using hedonic pricing theory, the starting point was appraising different residential property types throughout South Africa over from January 2013 to August 2017 in a cross-sectional manner. Different models were considered, including generalised linear models, hierarchical additive models and machine learning algorithms, namely random forests, gradient boosted machines and neural networks. However, due to data integrity issues, the data firstly had to be “scrubbed”, removing duplicate property entries, missing data and outliers. An autoencoder proved useful to identify anomalous data showing a more realistic spread of the data.

Chapters three to five fit hedonic models as appraisal functions to build the residential property price index.

Chapter three investigated generalised linear models as a candidate methodology. The gamma model showed better overall goodness of fit compared to the normal model and had the advantage of keeping estimates on the original scale, unlike the log normal model similar to the findings of Bax and Chasomeris (2019).

All the covariates were statistically significant and easily interpretable using this transparent framework, similar to Bourassa et al (2007) where treating location as a fixed effect accounted for the spatial dependence in the data. The findings provide useful insights of how the marginal utility of different property characteristics influence listing prices *ceteris paribus*. The marginal utility of bathrooms was consistently higher than bedrooms and the growth of townhouses in comparison to apartments was positive linear growth over the observation period.

Chapter four implemented hierarchical generalised additive models to investigate partial pooling and smooth relationships between property characteristics and listing prices where linear functional form was relaxed. An additional spatial covariate was introduced through the partitioning around the medoids clustering algorithm which was treated as a random effect. The hierarchical structure of homes was modelled through random intercepts where partial pooling of the suburb captured important group level variation and showed to be more effective than treating the longitude and latitude as an isotropic bivariate function using splines on the sphere. The findings suggest that the hedonic price functions in South Africa are non-linear and that smooth functions are appropriate for estimating the relationship between listing prices and structural property characteristics. These findings are congruent to the research of Pace (1998) in a study estimating residential property prices in Memphis, USA where smooth functions of property characteristics outperformed log linear models. Furthermore, an improvement in model performance was experienced when combining random effects to model the hierarchical structure of properties, nested at multiple locations, similar to the findings of Tan *et al* (2019) who also demonstrated the effectiveness of this approach over non-hierarchical models with explicit functional form.

Chapter five presented a machine learning approach to developing hedonic models comparing random forests, gradient boosted machines and neural networks as hedonic functions. 5-Fold cross validation was applied with a random grid search to search over the hyperparameter space. The gradient boosted machines generalised well to the out of sample data and produced the lowest generalisation errors compared to the random forests and neural networks. The results of this study are similar to the findings of the research done on single family homes in Switzerland by Wezel *et al* (2005) where gradient boosted machines provided the lowest out-of-sample errors in



comparison to random forests and neural networks where the latter provided the most inconsistent. The partial dependence plots showed that the marginal utility for different physical characteristics varied at different quantiles showing that, on average, larger sized properties do not necessarily yield higher prices and can result in diminished returns. The suburb categorical variable was consistently deemed the most important followed by size and the number of bathrooms.

All three-of the best candidate modelling approaches, namely the gamma generalised linear models, suburb hierarchical generalised additive models and gradient boosted machines accounted for the spatial dependency in the data, however, the gradient boosted machines produced the lowest out of sample error overall.

Chapter six applied index number theory using the gradient boosted machines and developed a residential property price index, comparing it to a simple median mix-adjusted index and the ABSA global property guide. The gradient boosted machine listing price index showed a similar trend of annual home price changes to the one published by ABSA bank in the ABSA global property guide. The findings show the greatest contraction for inflation adjusted home prices was experienced from 2015-2016 as a function of low growth and high inflation which was also in line with the ABSA global property guide. Provincial price changes were also measured with the Western Cape experiencing the largest growth in listing prices for each year over the observation period. The Listing Price Index Calculator presents the application of the methodology proposed in this study. For property market participants wishing to sell their homes, the Listing Price Index Calculator could be a useful first assessment of what to list their homes for on the market, making it a complimentary tool to estate agent valuations. Potential buyers are able to ascertain the typical asking price for a home with the desired set of characteristics in the suburb of interest.

Real estate agents and property portals are well positioned to leverage their extensive property market data to develop a similar data product that provides property market participants with a simple, data driven tool to assess and understand price developments.

## **7.2 Limitations and Future Work**

Limitations of this research include not having the property level spatial data (the latitude and longitude coordinates) which could have improved the study and modelling efforts, facilitating different spatial models and distance measures like proximity to schools, central business district etc. This information is considered private data and was not provided for the study. Not having this information may have cancelled out the neighbourhood effect where large differences of home prices may be present in a suburb. Additionally, more property characteristic data could have augmented modelling efforts. de Haan and Diewert, (2011) cite the age of a property as an important property attribute in hedonic models. Future work could include overcoming these limitations and comparing improved models and different models taking cognisance of this additional data.

## References

- Aggarwal, C. (2018). *Neural Networks and Deep Learning*. Springer, Yorktown Heights, New York.
- Anglin, P. and Gençay, R. (1996). Semiparametric estimation of a hedonic price function. *Journal of Applied Econometrics*, 11(6), pp.633-648.
- Ando, A., and F. Modigliani., (1963). The Life-Cycle Hypothesis of Saving: Aggregate Implications and Tests. *American Economic Review*, 103. pp.55–84.
- Anselin, L. (2006). *Spatial Econometrics*, in T. C. Mills (ed.) and K. Patterson(ed.),Palgrave Handbook of Econometrics, New York: Palgrave Macmillon.
- Augustin, N.H., Sauleaub, E.A., Wood, S.N. (2012). On quantile quantile plots for generalised linear models. *Computational Statistics & Data Analysis*. Vol 56(8), pp. 2404-3409.
- Balk, B.M. (2008), Price and Quantity Index Numbers; Models for Measuring Aggregate Change and Difference, New York: Cambridge University Press.
- Bax, D., and Chasomeris, M. (2019). Listing price estimation of apartments: A generalised linear model. *Journal of Economic and Financial Sciences*, 12(1).
- Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1). pp. 281–305.
- Bin, O. (2004). A prediction comparison of housing sales prices by parametric versus semi-parametric regressions. *Journal of Housing Economics*, 13(1), pp.68-84.
- Blum, A., Kalai, A. and Langford, J. (1999). Beating the Hold-Out: Bounds for K-fold and Progressive Cross-Validation. *In Computational Learning Theory*. pp. 202-208.
- Bordo, M.D., and Jeanne, O. (2002). Boom-Busts in Asset Prices, Economic Instability, and Monetary Policy. CEPR Discussion Paper 3398. Centre for Economic Policy Research, London. Available at: <https://www.nber.org/papers/w8966>.
- Bourassa, S. C., Cantoni, E. and Hoesli, M. (2007). Spatial Dependence, Housing Submarkets, and House Price Prediction, *Journal of Real Estate Finance and Economics*, 35(1), pp. 142-160.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), pp.5-32.

Brunauer, W., Feilmayr, W. and Wagner, K., (2012). A New Residential Property Price Index for Austria, *Statistiken Daten & Analysen*, Q3/12, pp.90–102.

Bzdok, D. and Eickhoff, S.B. (2016) Statistical learning of the Neurobiology of Schizophrenia, *The Neurobiology of Schizophrenia*, pp. 337–350. Available at: <https://doi.org/10.1016/b978-0-12-801829-3.00027-6>.

Bzdok, D. (2017) Classical statistics and statistical learning in Imaging Neuroscience, *Frontiers in Neuroscience*, 11. Available at: <https://doi.org/10.3389/fnins.2017.00543>.

Bzdok, D., Altman, N. and Krzywinski, M. (2018) Statistics Versus Machine Learning, *Nature Methods*, 15(4), pp. 233–234. Available at: <https://doi.org/10.1038/nmeth.4642>.

Candel, A., LeDell, E., Parmar, V. and Arora, A. (2017). Deep Learning with H2O, H2O.ai Inc., California, from <http://docs.h2o.ai/h2o/latest-stable/h2o-docs/booklets/DeepLearningBooklet.pdf>.

Case, K. and Shiller, R. (1987). Prices of single-family homes since 1970: new indexes for four cities. Available at: <https://doi.org/10.3386/w2393>.

Case, K. and Shiller, R. (1987). The Efficacy of the market for Single-Family Homes. *American Economic Review*, 79(1), pp.125-137.

Case, K., Quigley, J. and Shiller, R., (2005). Comparing Wealth Effects: The Stock Market Versus the Housing Market. *Advances in Macroeconomics* 5(1), pp.1235-1235.

Chapman, P. (2000) *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS.

Chollet, F. and Allaire, J. (2018). *Deep Learning with R*. Manning Publications, New York.

Chiles, J. and Delfiner, P. (1999). *Geostatistics: Modeling spatial uncertainty*. New York, John Wiley & Sons, p. 695.

Chiles, J. and Delfiner, P. (1999). *Geostatistics: Modeling spatial uncertainty*. New York, John Wiley & Sons, p. 695.

Clark, I. (2010). Statistics or geostatistics? Sampling error or nugget effect? Fourth World Conference on Sampling and Blending, 110, pp. 13-18.

Click, C., Malohlava, M., Parmar, V., Roark, H., and Candel, A. (2016). Gradient Boosted Models with H2O. <http://h2o.ai/resources/>.

Davidson, A. C. and Snell, E. J. (1991). Residuals and diagnostics. Chapter 4 of Hinkley et al.

Day, B. (2003). Submarket Identification in Property Markets: A Hedonic Housing Price Model for Glasgow. Working Paper - Centre for Social and Economic Research on the Global Environment.

Diewert, W.E. (2007), The Paris OECD-IMF Workshop on Real Estate Price Indexes: Conclusions and Future Directions, Discussion Paper 07-01, Department of Economics, University of British Columbia, Vancouver, British Columbia, Canada, V6T 1Z1.

Diewert, W.E., S. Heravi and M. Silver (2009), Hedonic Imputation Versus Time Dummy Hedonic Indexes, pp. 161-196 in Price Index Concepts and Measurement, W.E. Diewert, J. Greenlees and C. Hulten (eds.), *NBER Studies in Income and Wealth*, Chicago: University of Chicago Press

Dobson, A. and Barnett, A. (2008). *An Introduction to Generalized Linear Models*, Third Edition. Texts in Statistical Science, 77. Chapman & Hall/CRC Press, Boca Raton, Florida.

Du Preez, M., Lee, D. and Sale, M. (2013). Nonparametric estimation of a hedonic price model: A South African case study. *Journal for Studies in Economics and Econometrics*, 37, pp.41-62.

Els, M. and Von Fintel, D., 2010. Residential property prices in a submarket of South Africa: Separating real returns from attribute growth. *South African Journal of Economics*, 78(4), pp.418-436.

Fox, J. and Weisberg, S. (2018). Visualizing Fit and Lack of Fit in Complex Regression Models with Predictor Effect Plots and Partial Residuals. *Journal of Statistical Software*, 87(9).

Freund, Y. and Schapire, R.E. (1996). Experiments with a New Boosting Algorithm. Proceedings from *ICML '96: The 13<sup>th</sup> International Conference on Machine Learning, Bari, Italy: Morgan Kaufmann*, 148-156.

Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5). pp. 1189–1232.

Friedman, J. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4), pp.367-378.

LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Aboyoun, P., Kurka M., and Malohlava, M. (2019). h2o: R Interface for H2O. R package version 3.22.1.1. <https://CRAN.R-project.org/package=h2o>.

Gartner. (2018). *gartner magic quadrant - Open Source Leader in AI and ML*. [online] Available at: <https://www.h2o.ai/gartner-magic-quadrant/> [Accessed 4 Sep. 2019].

Gelman A, Hill J. (2007). *Data Analysis Using Regression and Hierarchical/Multilevel Models*. New York: Cambridge University Press.

Girouard, N. and Blöndal, S., 2001. House prices and economic activity. OECD Economics Department Working Papers No. 279, ECO/WKP(2001)5, OECD, France.

Goodhart, C. and Hofmann, B. (2008). House prices, money, credit, and the macroeconomy. *Oxford Review of Economic Policy*, 24(1), pp.180-205.

Goodman, A.C. (1978). Hedonic prices, price indices and housing markets. *Journal of Urban Economics*, 5(4), pp.471-484.

Greene, W.H. (2003). *Econometric Analysis*. 5th Edition, Prentice Hall, Upper Saddle River.

Greenwell, B. and Boehmke, B., 2020. Variable Importance Plots—An Introduction to the vip Package. *The R Journal*, 12(1), p.343.

Guisan, A. and Zimmermann, N.E. (2000.) Predictive habitat distribution models in ecology. *Ecological Modelling*, 135. pp.147-186.

Gujarati, D.N. (2004). *Basic Econometrics*. 4th Edition, Tata McGraw-Hill, New York.

de Haan, J. (2004), Direct and Indirect Time Dummy Approaches to Hedonic Price Measurement, *Journal of Social and Economic Measurement*, 29, pp.427-443.

de Haan, J. (2009), Comment on Hedonic Imputation versus Time Dummy Hedonic Indexes, pp. 196-200 in Price Index Concepts and Measurement, W.E. Diewert, J.S. Greenlees and C.R. Hulten (eds.), *Studies in Income and Wealth*, 70, Chicago: University of Chicago Press.

de Haan, J. and Diewert, E. (2011). Handbook on residential property indices, Eurostat European Commission, viewed 12 December 2019, from <https://ec.europa.eu/eurostat/documents/3859598/5925925/KS-RA-12-022-EN.PDF>.

Harvard Business Review. 2020. *How To Build Great Data Products*. [online] Available at: <https://hbr.org/2018/10/how-to-build-great-data-products> [Accessed 20 May 2020].

Hastie, T. and Tibshirani, R. (1986). Generalised Additive Models. *Statistical Science*, 1(3), pp.297-310.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. Boca Raton, FL, USA: CRC Press.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *Elements of Statistical Learning*. Springer Series in Statistics Springer New York Inc., New York.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd Edn. Springer, New York.

Hill, R. and Scholz, M. (2018), Can Geospatial Data Improve House Price Indexes? A Hedonic Imputation Approach with Spline., *Review of Income and Wealth*, 64(4), 737–756.

Hill, R.J. and D. Melser (2008), Hedonic Imputation the Price Index Problem: An Application to Housing, *Economic Inquiry*, 46(4), pp.593-609.

Hill, R. J. (2013). Hedonic Price Indexes for Residential Housing: A Survey, Evaluation and Taxonomy. *Journal of Economic Surveys*, 27(5), pp. 879-914.

Hotz, N. (2022) *What is CRISP DM?*, *Data Science Process Alliance*. Available at: <https://www.datascience-pm.com/crisp-dm-2/> (Accessed: November 22, 2022).

ILO, IMF, OECD, Eurostat, United Nations, World Bank (2004), *Consumer Price Index Manual: Theory and Practice*, Geneva: ILO.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R*. New York: Springer.

Kuhn, M. and Johnson, K. (2018). *Applied Predictive Modeling*. Springer, New York.

Jiang, L., Phillips, P.C. and Yu, J. (2015), New methodology for constructing real estate price indices applied to the Singapore residential market, *Journal of Banking & Finance*, 61, pp.121-131.

Lindsey, J.K. (1997). *Applying generalized linear models*. Springer Science & Business Media.

LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Aboyoun, P., Kurka M. and Malohlava M. (2019). h2o: R Interface for 'H2O'. R package version 3.22.1.1. <https://CRAN.R-project.org/package=h2o>.

Lindsey, J.K. (1997). *Applying generalised linear models*. New York: Springer Science & Business Media.

Lisi, G. (2013). On the Functional Form of the Hedonic Price Function: A Matching-theoretic Model and Empirical Evidence. *International Real Estate Review*, 16(2), pp.189-207.

Luus, C. (2002). The ABSA Residential Property Market Database for South Africa—Key Data Trends and Implications. BIS papers no 21.

Lyons, R.C. (2015). Measuring house prices in the long run: Insights from Dublin, 1900-2015, viewed 29 April 2018, from <http://eh.net/eha/wp-content/uploads/2015/05/Lyons.pdf>.

Maindonald, J. (2010). Computations for Linear and Generalised Additive Models. *Research Gate*. 2. pp.1-2.

Mayer, M., Bourassa, S., Hoesli, M. and Scognamiglio, D., 2019. Estimation and updating methods for hedonic valuation. *Journal of European Real Estate Research*, 12(1), pp.134-150.

Mccullagh, P. and Nelder, J. (1989), *Generalised linear models*, vol. 37, CRC Press, London.

Meir, R. and Ratsch, G. (2003). An introduction to boosting and leveraging. In: Mendelson S., Smola, A. J. (eds.), *Advanced Lectures on Machine Learning: Machine. LNCS (LNAI)*, 2000, pp.118-183. Springer, Heidelberg.

Musset, L. (2006). OECD Environment Health and Safety Publications Series on Testing and



Assessment' No. 54. [pdf]. Available at:  
<[http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mo  
no\(2006\)18&doclanguage=en](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mo<br/>no(2006)18&doclanguage=en) [Accessed 12 January 2019].

Muthalaly, R.S. (2017). Using Deep Learning to predict the mortality of Leukemia patients', Queen's University (Kingston, Ont.). <http://qspace.library.queensu.ca/handle/1974/15929> (accessed 12 February 2019).

Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A*, 135, pp.370-384.

Olivier, J., Johnson, W. and Marshall, G. (2008). The logarithmic transformation and the geometric mean in reporting experimental IgE results: What are they and when and why to use them? *Annals of Allergy Asthma Immunology*, 100 pp. 333-338 and pp.625-626.

Pace, R. K. (1998). Appraisal using generalized additive models. *Journal of Real Estate Research*, 15, 77–99.

Pedersen, E., Miller, D., Simpson, G. and Ross, N. (2019). Hierarchical generalised additive models in ecology: an introduction with mgcv. *PeerJ*, 7, p.e6876.

Ploner, A. (1999.) The use of the variogram cloud in geostatistical modelling. *Environmetrics*, 10(4) pp.413-437.

R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

Reiss, P.T. and Ogden, R.T. (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society, Series B*. 71, pp. 505-524.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy*, 82(1), pp.34-55.

Rousseeuw, P. Silhouettes. (1987). A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Shimizu, C., Nishimura, K, and Watanabe, T. (2010). Housing Prices in Tokyo: A Comparison of Hedonic and Repeat Sales Measures. *Journal of Economics and Statistics*, 230. pp.792-813.

Silver, M. (2016). How to better measure hedonic residential property price indexes. IMF Working Paper, WP/16/213, IMF, Washington DC.

Schmidt, A. and Finan, C. (2018). Linear regression and the normality assumption. *Journal of Clinical Epidemiology*, 98, pp.146-151.

Schubert, E and Rousseeuw, P. Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. *Lecture Notes in Computer Science* (2019): pp.171–187.

Shimizu, C., Nishimura, K.G. and Watanabe, T. (2010). Housing prices in Tokyo: A comparison of hedonic and repeat sales measures. *Jahrbücher für Nationalökonomie und Statistik*, 230(6), pp. 792–813. Available at: <https://doi.org/10.1515/jbnst-2010-0612>.

Shukur, G. and Mantalos, P. (2004). Size and Power of the RESET Test as Applied to Systems of Equations: A Bootstrap Approach. *Journal of Modern Applied Statistical Methods*, 3(2), pp.370-385.

Tan, R., He, Q., Zhou, K., Song, Y. and Xu., H. (2019). Administrative hierarchy, housing market inequality, and multilevel determinants: a cross-level analysis of housing prices in China. *Journal of Housing and the Built Environment*. 34. Pp.845-868.

Triplett, J.E. (2006), *Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes; Special Application to Information and Technology Products*, Directorate for Science, Technology and Industry, Paris: OECD.

Uyar, B., & Brown, K. (2007). Neighborhood affluence, school-achievement scores, and housing prices: cross-classified hierarchies and HLM. *Journal of Housing Research*, 16(2), pp. 97-116.

van Wezel, M, M Kagie, and R Potharst. (2005). Boosting the Accuracy of Hedonic Pricing Models. *Econometric Institute Research Papers*. No EI 2005-50. Rotterdam: Erasmus University, Erasmus School of Economics (ESE).

Wickham et al., (2019). Welcome to the tidyverse. *Journal of Open-Source Software*, 4(43), 1686, <https://doi.org/10.21105/joss.01686>.

Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2020). shiny: Web Application Framework for R. R package version 1.5.0. <https://CRAN.R-project.org/package=shiny>

Wood, S. N. (2006). *Generalised additive models: An Introduction with R*. Boca Raton, FL, Chapman & Hall/CRC.

Wood, S.N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalised linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 73, pp. 3–36.

Wood, S.N. (2017). *Generalised Additive Models: An Introduction with R* (2nd edition). Boca Raton, FL, Chapman and Hall/CRC.

Wood, S.N. (2019). Package 'mgcv'. [online]. Available from <http://cran.r-project.org/web/packages/mgcv/index.html>.

Wendelberger, J. (1981). PhD Thesis, University of Wisconsin.

Zhong, Z., Yan, J., Wu, W., Shao, J. and Liu, C.L (2018). Practical Block-Wise Neural Network Architecture Generation. IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 2423-2432.

Zillow, (2019). What is a Zestimate? Zillow's Zestimate Accuracy | Zillow. [online] Zillow. Available at: <https://www.zillow.com/zestimate/> [Accessed 15 Aug. 2019]

Zuur, A.F., Ieno, E.I., Walker, N.J., Saveliev, A.A., and Smith G.M. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Statistics for Biology and Health New York: Springer.

## Appendix

```
# Setup Environment -----  
  
library(odbc)  
  
library(DBI)  
  
library(shiny)  
  
library(tidyverse)  
  
library(tm)  
  
library(h2o)  
  
library(parallel)  
  
library(DT)  
  
library(shinyjs)  
  
library(shinyWidgets)  
  
library(shinythemes)  
  
library(shinycssloaders)  
  
library(ggrepel)  
  
require(sf)  
  
library(raster)  
  
library(ggspatial)  
  
library(knitr)  
  
library(kableExtra)  
  
setwd("C:/Users/Dane/Documents/gradient-booted-machine-paper")  
  
options(scipen = 999)  
  
  
# Build UI -----  
  
ui <-
```

```

fluidPage(

  theme = shinytheme("spacelab"),

  navbarPage(

    "",

    theme = shinytheme("spacelab"),

    position = "static-top",

    inverse = FALSE,

    collapsible = TRUE,

    fluid = TRUE,

    tabPanel(

      title = "Listing Price Index Calculator",

      sidebarLayout(

        sidebarPanel(

          br(),

          selectInput(

            inputId = "Province",

            label = "Province:",

            choices = "...",

            selected = "KwaZulu-Natal"

          ),

          selectInput(

            inputId = "Suburb",

            label = "Suburb:",

            choices = "...",

```

```

        selected = "Margate"
    ),
    selectInput(
        inputId = "PropertyType",
        label = "Property type:",
        choices = "...",
        selected = "Apartment"
    ),
    sliderInput(
        inputId = "Size",
        label = "Size (sqm):",
        min = 0,
        max = 0,
        value = 0,
        step = 1,
        round = 0
    ),
    sliderInput(
        inputId = "Bedrooms",
        label = "Bedrooms:",
        min = 0,
        max = 0,
        value = 0,
        step = 1,

```

```

        round = 0
    ),
    sliderInput(
        inputId = "Bathrooms",
        label = "Bedrooms:",
        min = 0,
        max = 0,
        value = 0,
        step = 1,
        round = 0
    ),
    selectInput(
        inputId = "firstValuation",
        label = "Base Valuation Period:",
        choices = "...",
    ),
    selectInput(
        inputId = "secondValuation",
        label = "Comparison Valuation Period:",
        choices = "...",
    ),
),

```

```

mainPanel(tabsetPanel(
  tabPanel(
    icon("cog", lib = "glyphicon"),
    tableOutput("resultsTable") %>%
      withSpinner(color = "#0dc5c1")
  )
),

br(),
tabsetPanel(
  tabPanel(
    icon("bar-chart-o"),
    plotOutput("map",
      height = "600px",
      width = "600px") %>%
      withSpinner(color = "#0dc5c1")
    )
  )
)
)

```



```

    )

    )

    )

# Build Server -----
server <- function(session, input, output) {

# Ingest Data -----

connection <-

odbc::dbConnect(

  odbc::odbc(),

  Driver = "SQL Server",

  Server = "DESKTOP-DDGOMB4\\SQLEXPRESS",

  Database = "PhD",

  trusted_connection = "yes",

  Port = 1433

)

appData <-

DBI::dbGetQuery(conn = connection,

  statement =

    "SELECT * FROM [dbo].[tb_spatiallyClusteredData]") %>%

```

```

dplyr::rename(Province = ProvinceName,
              Size = `[lnSize]`,
              Suburb = Area) %>%

dplyr::mutate(Suburb = trimws(Suburb, which = "left")) %>%

dplyr::rename(Area = ProvinceArea) %>%

dplyr::select(-Cluster, -Data)

```

# Update UI Dynamically -----

# Evaluation period 1:

```

observe({
  x <- appData %>%

  dplyr::filter(ListingYear < 2017) %>%

  dplyr::select(ListingYear) %>%

  dplyr::distinct(.) %>%

  dplyr::arrange(ListingYear)

  updateSelectInput(
    session = session,
    inputId = "firstValuation",
    label = "Base Valuation Period:",
    choices = c(x$ListingYear)[1:length(x$ListingYear)]
  )
}

```

```
})
```

```
# Evaluation period 2:
```

```
observe({  
  x <- appData %>%  
    dplyr::filter(!ListingYear <= input$firstValuation) %>%  
    dplyr::select(ListingYear) %>%  
    dplyr::distinct(.) %>%  
    dplyr::arrange(ListingYear)  
  
  updateSelectInput(  
    session = session,  
    inputId = "secondValuation",  
    label = "Comparison Valuation Period:",  
    choices = c(x$ListingYear)[1:length(x$ListingYear)]  
  )  
  
})
```

```
# Province:
```

```
observe({  
  x <- appData %>%  
    dplyr::filter(ListingYear == input$firstValuation) %>%  
    dplyr::select(Province) %>%
```

```

dplyr::distinct(.)

updateSelectInput(
  session = session,
  inputId = "Province",
  label = "Province:",
  choices = c(x$Province)[1:length(x$Province)]
)

})

# Suburb:
observe({
  x <- appData %>%
    dplyr::filter(ListingYear %in% input$firstValuation &
      Province %in% input$Province) %>%
    dplyr::select(Suburb) %>%
    dplyr::distinct(.)

  updateSelectInput(
    session = session,
    inputId = "Suburb",
    label = "Suburb:",
    choices = c(x$Suburb)[1:length(x$Suburb)]
  )
})

```

```

    )

  })

# Property type:
observe({
  x <- appData %>%
    dplyr::filter(ListingYear == input$firstValuation) %>%
    dplyr::select(PropertyType) %>%
    dplyr::distinct(.)

  updateSelectInput(
    session = session,
    inputId = "PropertyType",
    label = "Property Type:",
    choices = c(x$PropertyType)[1:length(x$PropertyType)]
  )

})

# Size:
observe({
  x <- appData %>%
    dplyr::filter(ListingYear %in% input$firstValuation &

```

```

        PropertyType %in% input$PropertyType
    ) %>%

    dplyr::mutate(Size = exp(Size)) %>%

    dplyr::select(Size) %>%

    dplyr::distinct(.)

suppressWarnings(
  updateSliderInput(
    session = session,
    inputId = "Size",
    label = "Size (sqm):",
    min = min(x$Size),
    value = median(x$Size),
    max = max(x$Size),
    step = 1
  )
)

})

# Bedrooms:

observe({
  x <- appData %>%

  dplyr::filter(ListingYear %in% input$firstValuation &

```

```

        PropertyType %in% input$PropertyType
    ) %>%
    dplyr::select(Bedrooms) %>%
    dplyr::distinct(.)

suppressWarnings(
  updateSliderInput(
    session = session,
    inputId = "Bedrooms",
    label = "Bedrooms:",
    min = min(x$Bedrooms),
    value = median(x$Bedrooms),
    max = max(x$Bedrooms),
    step = 1
  )
)

})

# Bathrooms:

observe({
  x <- appData %>%

  dplyr::filter(ListingYear %in% input$firstValuation &
    PropertyType %in% input$PropertyType

```

```

    ) %>%

    dplyr::select(Bathrooms) %>%

    dplyr::distinct(.)

suppressWarnings(

  updateSliderInput(

    session = session,

    inputId = "Bathrooms",

    label = "Bathrooms:",

    min = min(x$Bathrooms),

    value = median(x$Bathrooms),

    max = max(x$Bathrooms),

    step = 1

  )

)

})

# h2o Setup -----

# Fire up h2o cluster:

suppressWarnings(

  h2o.init(

    ip = "localhost",

    nthreads = parallel::detectCores() - 1,

```



```

    max_mem_size = "12G",
    min_mem_size = "10G"
  ))

# Load h2o learners:

h2o2013 <- h2o.loadModel(paste0(getwd(),
                                "/", "final_grid_2013_model_10"))

h2o2014 <- h2o.loadModel(paste0(getwd(),
                                "/", "final_grid_2014_model_10"))

h2o2015 <- h2o.loadModel(paste0(getwd(),
                                "/", "final_grid_2015_model_10"))

h2o2016 <- h2o.loadModel(paste0(getwd(),
                                "/", "final_grid_2016_model_10"))

h2o2017 <- h2o.loadModel(paste0(getwd(),
                                "/", "final_grid_2017_model_10"))

# Property Predictions -----

# Base model:

baseEvalModel <-
  reactive({
    if (input$firstValuation == 2013) {
      baseModel <- h2o2013
    } else {

```

```

if (input$firstValuation == 2014) {
  baseModel <- h2o2014
} else {
  if (input$firstValuation == 2015) {
    baseModel <- h2o2015
  } else {
    baseModel <- h2o2016
  }
}
}
})

```

# Comparative model:

```

compEvalModel <-
reactive({
  if (input$secondValuation == 2014) {
    compModel <- h2o2014
  } else {
    if (input$secondValuation == 2015) {
      compModel <- h2o2015
    } else {
      if (input$secondValuation == 2016) {
        compModel <- h2o2016
      } else {

```

```

      compModel <- h2o2017
    }
  }
}
})

```

# Get input data for predictions:

```
h2oData <-
```

```

reactive({
  cbind.data.frame(
    "Size" = input$Size,
    "Bedrooms" = input$Bedrooms,
    "Bathrooms" = input$Bathrooms,
    "PropertyType" = input$PropertyType,
    "Province" = input$Province,
    "Suburb" = input$Suburb
  ) %>% dplyr::inner_join(.,
    appData %>%
      dplyr::select(Area, Province, Suburb) %>%
      dplyr::distinct(),
    by = c("Province" = "Province",
      "Suburb" = "Suburb")
  )
}

```

```

) %>%

dplyr::mutate(Size = log(Size)) %>%

dplyr::select(-Suburb, -Province) %>%

as.h2o()

})

```

# Get predictions:

```
fitResults <-
```

```

  reactive({

    baseFit <-

      h2o.predict(

        baseEvalModel(),

        newdata = h2oData()) %>%

        as.data.frame()

```

```
compFit <-
```

```

  h2o.predict(

    compEvalModel(),

    newdata = h2oData()) %>%

    as.data.frame()

```

```
TTR <- (compFit$predict - baseFit$predict) / baseFit$predict
```

```
fitResults <- cbind.data.frame("Base Period" = baseFit$predict,
```

```

        "Comparison Period" = compFit$predict,
        "Price Change" = TTR * 100)
    })

# Build Index -----

# Province index:

# Laspeyres:
indexBaseData <-
  reactive({
    appData %>%
      dplyr::filter(ListingYear == input$firstValuation) %>%
      as.h2o()
  })

indexBaseDataFrame <-
  reactive({
    appData %>%
      dplyr::filter(ListingYear == input$firstValuation) %>%
      as.data.frame()
  })

lspBase <- reactive({

```

```

lspBaseData <- h2o.predict(baseEvalModel(), newdata = indexBaseData()) %>%
  as.data.frame() %>%
  dplyr::mutate(
    Province = indexBaseDataFrame()$Province,
    Suburb = indexBaseDataFrame()$Suburb
  )
})

```

```

lspCF <- reactive({
  lspCFData <- h2o.predict(compEvalModel(), newdata = indexBaseData()) %>%
    as.data.frame() %>%
    dplyr::mutate(
      Province = indexBaseDataFrame()$Province,
      Suburb = indexBaseDataFrame()$Suburb
    )
})

```

```

laspeyresProvince <-
  reactive({
    lspBase() %>%
      as.data.frame() %>%
      dplyr::mutate_if(is.factor, as.character) %>%
      dplyr::group_by(Province) %>%
      dplyr::summarise(meanBase = mean(predict)) %>%

```

```

dplyr::inner_join(
  .,
  lspCF() %>%
    as.data.frame() %>%
    dplyr::mutate_if(is.factor, as.character) %>%
    dplyr::group_by(Province) %>%
    dplyr::summarise(meanCF = mean(predict)),
  by = "Province"
) %>%
dplyr::group_by(Province) %>%
dplyr::summarise(laspeyresDelta = (meanCF / meanBase))
})

```

```

laspeyresSuburb <-
reactive({
  lspBase() %>%
    as.data.frame() %>%
    dplyr::mutate_if(is.factor, as.character) %>%
    dplyr::group_by(Suburb) %>%
    dplyr::summarise(meanBase = mean(predict)) %>%
    dplyr::inner_join(
      .,
      lspCF() %>%
        as.data.frame() %>%

```

```

    dplyr::mutate_if(is.factor, as.character) %>%

    dplyr::group_by(Suburb) %>%

    dplyr::summarise(meanCF = mean(predict)),

    by = "Suburb"

  ) %>%

  dplyr::group_by(Suburb) %>%

  dplyr::summarise(laspeyresDelta = (meanCF / meanBase))

})

```

# Paasche:

```
indexCompData <-
```

```

  reactive({

    appData %>%

      dplyr::filter(ListingYear == input$secondValuation) %>%

      as.h2o()

  })

```

```
indexCompDataFrame <-
```

```

  reactive({

    appData %>%

      dplyr::filter(ListingYear == input$secondValuation) %>%

      as.data.frame()

  })

```



```

pscBase <- reactive({
  pscBaseData <- h2o.predict(baseEvalModel(), newdata = indexCompData()) %>%
    as.data.frame() %>%
    dplyr::mutate(
      Province = indexCompDataFrame()$Province,
      Suburb = indexCompDataFrame()$Suburb
    )
})

```

```

pscCF <- reactive({
  pscCFData <- h2o.predict(compEvalModel(), newdata = indexCompData()) %>%
    as.data.frame() %>%
    dplyr::mutate(
      Province = indexCompDataFrame()$Province,
      Suburb = indexCompDataFrame()$Suburb
    )
})

```

```

paascheProvince <-
  reactive({
    pscBase() %>%
      as.data.frame() %>%
      dplyr::mutate_if(is.factor, as.character) %>%
      dplyr::group_by(Province) %>%

```

```

dplyr::summarise(meanBase = mean(predict)) %>%
dplyr::inner_join(
  ,,
  pscCF() %>%
    as.data.frame() %>%
    dplyr::mutate_if(is.factor, as.character) %>%
    dplyr::group_by(Province) %>%
    dplyr::summarise(meanCF = mean(predict)),
  by = "Province"
) %>%
dplyr::group_by(Province) %>%
dplyr::summarise(paascheDelta = (meanCF / meanBase))
})

```

```

paascheSuburb <-
reactive({
  pscBase() %>%
    as.data.frame() %>%
    dplyr::mutate_if(is.factor, as.character) %>%
    dplyr::group_by(Suburb) %>%
    dplyr::summarise(meanBase = mean(predict)) %>%
    dplyr::inner_join(
      ,,
      pscCF() %>%

```

```

as.data.frame() %>%
  dplyr::mutate_if(is.factor, as.character) %>%
  dplyr::group_by(Suburb) %>%
  dplyr::summarise(meanCF = mean(predict)),
  by = "Suburb"
) %>%
  dplyr::group_by(Suburb) %>%
  dplyr::summarise(paascheDelta = (meanCF / meanBase))
})

```

# Geometric mean function:

```

getGeomMean = function(x, na.rm = TRUE) {
  gm <- exp(sum(log(x[x > 0]), na.rm = na.rm) / length(x))
  return(gm)
}

```

# Fisher province:

```

fisherProvince <-
  reactive({
    laspeyresProvince() %>%
      dplyr::inner_join(.,
        paascheProvince(),
        by = "Province") %>%
      dplyr::group_by(Province) %>%

```

```

dplyr::summarise(
  Fisher = (getGeomMean(c(laspeyresDelta, paascheDelta))-1)*100
)
})

# Fisher suburb:
fisherSuburb <-
  reactive({
    laspeyresSuburb() %>%
      dplyr::inner_join(.,
        paascheSuburb(),
        by = "Suburb") %>%
      dplyr::mutate_if(is.factor, as.character) %>%
      dplyr::group_by(Suburb) %>%
      dplyr::summarise(
        Fisher = (getGeomMean(c(laspeyresDelta, paascheDelta))-1)*100
      ) %>%
      dplyr::filter(
        Suburb == input$Suburb
      )
  })

# Render Table -----
output$resultsTable <-

```

```

function() {
  fitResults() %>%
  dplyr::mutate(
    `Base Period` = paste("ZAR",
                          format(round(`Base Period`, 0L),
                                big.mark = " ",
                                nsmall = 2L)),
    `Comparison Period` = paste("ZAR",
                                format(round(`Comparison Period`, 0L),
                                      nsmall = 2L,
                                      big.mark = " ")),
    `Price Change` = paste(format(round(`Price Change`, 2L),
                                nsmall = 2L),
                            " %"),
    `Suburb Growth` = paste(format(round(fisherSuburb()$Fisher, 2L)), " %")
  ) %>%
  knitr::kable(format.args = list(big.mark = " "),
               format = "html") %>%
  kable_styling(bootstrap_options = c("striped",
                                       "hover",
                                       "condensed",
                                       "responsive",
                                       full_width = FALSE,
                                       position = "left"))

```

```
}
```

```
# Build Map -----
```

```
# Map data:
```

```
suppressWarnings(  
  
```

```
    saData <- shapefile("C:/Users/Dane/Desktop/Shapefiles/SOU.shp"))  
  
```

```
saData$id <- row.names(saData)
```

```
saFort <- fortify(saData) %>%  
  
```

```
  dplyr::filter(long < 35)
```

```
sa <- left_join(saFort, saData@data, by = "id")
```

```
centroids <-
```

```
  sa %>%  
    
```

```
    dplyr::group_by(ADM1) %>%  
      
```

```
      dplyr::summarise(lon = mean(long),  
        
```

```
        lat = mean(lat)) %>%  
      
```

```
      dplyr::rename(Province = ADM1)
```

```
multiPolygonFile <- readRDS("/Users/Dane/Downloads/gadm36_ZAF_2_sf.rds")
```

```

plotData <- multiPolygonFile %>%
  dplyr::select(NAME_1, geometry)

# Build map:
polyMmap <-
  ggplot(plotData) +
  geom_sf(data = plotData,
    aes(fill = NAME_1),
    col = sf.colors(52, categorical = TRUE),
    alpha = 0.7) +
  theme_classic(base_size = 16) +
  theme(plot.title = element_text(hjust = 0.5),
    legend.title = element_text(color = "black", size = 10),
    legend.text = element_text(color = "black", size = 10),
    axis.title.x = element_blank(),
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank(),
    axis.title.y = element_blank(),
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank()) +
  labs(fill = "Province")

# Render Map -----
output$map <-

```

```

renderPlot({

  polyMmap +

  geom_label_repel(
    data = centroids %>%
      dplyr::mutate_if(is.factor, as.character) %>%
      dplyr::inner_join(
        ..
        fisherProvince() %>%
          dplyr::mutate_if(is.factor, as.character) %>%
          dplyr::mutate(Fisher = round(Fisher, digits = 2)),
        by = "Province"
      ),
    aes(
      x = lon,
      y = lat,
      label = Fisher,
      fill = factor(Province)
    ),
    color = "black",
    size = 5,
  ) +

  labs(fill = "Province") +

```



```

guides(
  fill = guide_legend(
    override.aes = aes(label = "")
  )
) +
theme_classic(base_size = 16) +
theme(legend.position = "bottom",
  legend.title = element_text(color = "black", size = 10),
  legend.text = element_text(color = "black", size = 10),
  axis.title.x = element_blank(),
  axis.line.x = element_blank(),
  axis.text.x = element_blank(),
  axis.ticks.x = element_blank(),
  axis.title.y = element_blank(),
  axis.line.y = element_blank(),
  axis.text.y = element_blank(),
  axis.ticks.y = element_blank(),
  axis.line = element_blank())
}))

}

# Build Application -----
shinyApp(ui = ui, server = server)

```

# END -----