Journal of African Real Estate Research

Volume 4, Issue 2

www.journals.uct.ac.za/index.php/JARER/index

A Machine Learning Web Application to Estimate Listing Prices of South African Homes

Dane Bax[1], Temesgen Zewotir[2], and Delia North[3]

[1-3] School of Mathematics, Statistics and Computer Science, University of Kwa-Zulu Natal.

To cite this article: Bax, D., Zewotir, T. & North, D. (2019). A Machine Learning Web Application to Estimate Listing Prices of South African Homes. *Journal of African Real Estate Research, 4*(2), pp.1-23. DOI: 10.15641/jarer.v4i2.802.

Abstract

Due to the heterogeneous nature of residential properties, determining selling prices which will reconcile supply and demand is difficult. Establishing realistic listing prices is vitally important for sellers to prevent prolonged time on the market. Sellers have several resources available to assist in this endeavour, all of which involve understanding current market dynamics through analysing recent sales and listing data. Property portals which aggregate real estate agencies' data, hosting it on online platforms, are one such resource, along with individual real estate agencies. Leveraging this data to develop solutions that could aid sellers in listing price decision making is a potential business objective that could not only add value to sellers but create a competitive advantage by increasing traffic to an online real estate platform. Using data provided by a South African online property portal, this paper creates a web application using machine learning to estimate listing prices for different types of homes throughout South Africa. This study compared log linear and gradient boosted models, estimating residential listing prices over a four-year period. The results indicate that although log linear models are suitable to account for spatial dependency in the data through the inclusion of a fixed location effect, the assumption of linear functional form was not satisfied. The gradient boosted models do not impose explicit functional form requirements, making them flexible candidates. Similarly, these models were able to handle the spatial dependency adequately. The gradient boosted models also achieved a lower out of sample error compared to the log linear models. The findings show that over the observation period, larger properties consistently experience a diminishing return at some point over the marginal distribution of physical characteristics. The web application details how sellers are easily able to obtain mean listing price estimates and gauge the growth thereof, by simply inputting their property interest criteria.

Keywords: Machine Learning; H2o.ai; Hedonic Modelling; Residential Property Valuation

[1] danebax@gmail.com
[2] zewotir@ukzn.ac.za
[3] northd@ukzn.ac.za

1. Introduction

People wishing to sell their homes are faced with the challenging question of what price to list their property for on the market. They have several resources available to help determine this themselves, namely 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 which often results in an expensive sales commission when the property is sold. A comparative market analysis is a frequently used and recommended method for estate agents to use when valuing homes. This method examines several sources of information, including what similar properties have sold for recently, initial listing prices and duration on the market (Private Property, 2019). Real estate agents may also consider important home and neighbourhood characteristics in their estimates. PropertyFox, an online South Africa real estate agent, does not send agents to examine properties, choosing rather to use technology to combine traditional sales index methods with real time listing prices of similar homes for sale in order to derive estimates (PropertyFox, 2019). In real estate economics literature, hedonic pricing is a common approach employed to estimate home prices conditional on a property's set of characteristics, such as: size, number of bedrooms, number of bathrooms and location, amongst others. Similarly to estate agents, a hedonic model performs a comparative market analysis, however, it does so with a mathematical model. 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.

This study compares traditional log linear models to ensemble tree-based, machine learning models, namely gradient boosting, thereby developing hedonic listing price functions for the South African property market. A web application is developed to present how the proposed framework can be democratised to sellers by property portals or real estate agencies as a service offering. The web application allows users to quickly obtain property market listing price estimates and gauge historic growth.

2. Background and Objective

Hedonic pricing is a popular quality adjusted technique used in estimating property prices and constructing residential price indices (Jiang, Phillips & Yu, 2015; Shimizu et al., 2016). Hill (2013), in an extensive literature survey on various residential property index techniques, concluded that hedonic

indices have been favoured over other methods. A hedonic pricing function describes the price of a heterogeneous product through its utility bearing attributes (Rosen, 1974). De Haan and Erwin (2011) outline linear regression as a prominent hedonic pricing technique to estimate the marginal contributions of each property's attribute, taking the form of the full linear model (1) or the logarithmic linear model (2) given by:

$$p_n^t = \beta_0^t + \sum_{k=0}^{n} \beta_k^t \; z_{nk}^t \qquad (1)$$

$$ln\, p_n^t = \beta_0^t + \sum_{k=0}^{n} \beta_k^t \; z_{nk}^t \qquad (2)$$

The assumption that the price p_n^t of property n in period t is a function of a fixed number of parameters. β_0^t and β_k^t are the intercept and characteristic coefficients. Two main approaches exist using this technique. Firstly, the time dummy approach where a single regression 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 & Erwin, 2011). A disadvantage of this approach is the problem of temporal fixity which means that adding new periods to the data will result in changes to the coefficient estimates, resulting in revision estimates (Hill, 2011). The second main approach is the characteristics approach where separate regressions are run for the respective periods allowing the characteristic coefficients to vary from period to period. This is far more reasonable than the fixed time dummy approach (de Haan & Erwin, 2011). The characteristics method deals with temporal fixity and is more popular for computing residential price indices used by statistical agencies and government bureaus (Hill, 2011). The estimation of the 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. This study focuses on the starting point, estimating hedonic price functions for the South African property market.

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 price distribution. Du Preez, et al. (2013) developed a hedonic price function for houses in Walmer, South Africa, using the local constant estimator where the

direct estimate of $E(y|x)$ is produced with a kernel function that produces a smooth estimate of the densities. The researchers found that this non-parametric technique outperformed the parametric linear model. Bax and Chasomeris (2019) developed a hedonic price function for apartments listed for sale in coastal submarkets in KwaZulu-Natal, South Africa, using a gamma generalised linear model. The findings showed that the gamma distribution was appropriate and that treating the location as a fixed effect accounted for the spatial dependency effectively. These studies investigated certain property types and submarkets in isolation. This study aims to bridge this gap in South African real estate pricing literature by extending the scope to different property types and submarkets across South Africa.

The assumption that residential property prices depends linearly on a set of property coefficients makes the use of models, given in equations 1 and 2, attractive techniques to estimate hedonic functions with the added benefit of model transparency. However, Rosen (1974) suggests that this relationship is unlikely to be linear as the marginal cost of characteristics increase, coupled with the inability to unbundle characteristics. Lisi (2013) points out that the non-linear relationships between housing prices and housing characteristics is a key feature in developing hedonic pricing functions, although the specific functional form is not known *a priori*. Parametric hedonic models often suffer from misspecification due to the assumption of an explicit functional form, however, semi-parametric and non-parametric models have flexible functional forms which are capable of capturing more meaningful relationships. Pace (1998), Anglin and Gençay (1996) and Bin (2004) conducted different studies comparing several semi-parametric hedonic price functions to traditional parametric techniques where they showed an improvement in out of sample errors using approaches like generalized additive models. Van Wezel et al. (2005) applied gradient boosting, a non-parametric machine learning algorithm and stepwise linear models to develop hedonic price functions for three different datasets, two of which were US and UK housing datasets. The findings showed that the gradient boosting algorithm achieved a reduction in the out of sample errors in comparison to the stepwise linear models.

The ubiquity of hedonic pricing in real estate economics is evident where models that assume explicit functional form, such as log linear, are often used to map property characteristics to property prices. Although several studies exist that explore the use of semi-parametric and non-parametric techniques, there appears to be a lack of extensive research conducted in South Africa using contemporary machine learning algorithms to derive hedonic price functions for the residential property market. Furthermore, previous South African real estate pricing studies have focused on specific segments of the property market. This study contributes to South African real estate economics literature by comparing gradient boosting to traditional log linear models, developing yearly cross-sectional hedonic listing price functions for different residential property types throughout South Africa over a four-year period. An important feature of this study is the ability to visualize the gradient boosted hedonic price functions in an interpretable way, leveraging recent developments in machine learning literature. This paper presents an

algorithmic solution that could be used as an alternative to, or in conjunction with, manual comparative market analyses. The hedonic price functions are delivered through a web application which has practical implications for real estate agents, sellers and property portals. Sellers and real estate agents simply input the characteristics and location of the property of interest into an online user interface and easily obtain the expected listing price for each year in the data. The application allows users to gauge listing price growth over the study period, which can be informative in pricing decision making. Property portals and real estate agencies are well positioned to leverage their data, developing similar solutions, using potentially richer data.

3. Data and Design Framework

The open source statistical programming language R was used in this study. The dataset comprised of residential property listings spanning January 2014 to August 2017. These were obtained from an online South African property portal, Private Property (Pty Ltd). Table 1 describes the variables used in this study.

Table 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 e.g. apartment
Suburb	The suburb the property is located
Province	The province the property is located
Area	Concatenation of suburb and province
Listing Date	The advertisement date of the property on the portal
Latitude	The latitude coordinates of the area the property is located
Longitude	The longitude coordinates of the area the property is located

The longitude and latitude coordinates were collected via a geocoding API which was necessary for testing for spatial autocorrelation. Duplicate listings were identified and removed using row-wise matching along with incomplete observations. The initial data summary statistics are presented in Table 2.

Table 2. Data Summary Statistics

	Listing Price	Size	Bedrooms	Bathrooms
Minimum	R1,000	2	0	0
1st Quartile	R950,000	98	2	2
Median	R700,000	200	3	2
Mean	R2,461,210	259.8	3.135	2.252
3rd Quartile	R2,950,000	330	4	3
Maximum	R200,000,000	85,102	78	78

The data could be subject to incorrect data capturing arising from human error as real estate agents manually capture the information before it is disseminated via automatic feeds to the property portal. Examining Table 2, the maximum and minimum values seem improbable, therefore incorrect data capturing is a fair assumption. An autoencoder, which is a deep learning neural network, was developed to identify anomalous data points. Autoencoders generalize 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 indicate non-matching of the learned pattern (Candel et al., 2018). Reasonable lower limits were set on some variables using the ABSA bank property price index, the oldest price index in South Africa, as a guideline (Luus, 2002). Listing price was set to ≥R200,000 and size was set to ≥35m2. The autoencoder produced more feasible data by discounting data with a high reconstruction error. Table 3 presents the summary statistics of the final dataset which comprised of 382,826 properties.

Table 3. Final Data Summary Statistics

	Listing Price	**Size**	**Bedrooms**	**Bathrooms**
Minimum	R200,000	35	1	1
1st Quartile	R958,000	100	2	2
Median	R1,690,000	200	3	2
Mean	R2,159,173	231.3	3.1	2.16
3rd Quartile	R2,799,000	316	4	3
Maximum	R19,700,000	2,080	13	12

South Africa is approximately 1.2 million squared kilometres, comprising of nine provinces (Luus, 2002). The distribution of listings throughout the nine provinces is presented in Table 4. Gauteng represents the largest market share of listings over the period. Gauteng is also the smallest province, yet has the largest population (Statistics South Africa, 2019).

Table 4: Spatial and Temporal Distribution of Listings

Province	**Listing Year**			
	2014	**2015**	**2016**	**2017**
Eastern Cape	4,086	4,380	5,899	3,973
Free State	763	1,171	1,770	1,210
Gauteng	37,420	41,105	63,852	46,341
KwaZulu-Natal	9,602	10,731	13,519	9,946
Limpopo	737	748	1,164	770
Mpumalanga	3,079	3,397	4,697	2,067
North West	371	228	324	199
Western Cape	26,422	26,342	34,103	22,395
Northern Cape	0	0	10	5
Total	**82,480**	**88,102**	**125,338**	**86,906**

The period 2016 saw a large increase in listings, likely due to mechanisms of data collection peculiar to the property portal, nevertheless it was not discounted.

This study adopts the characteristics method proposed by de Haan and Erwin (2011) because of the advantage of avoiding revision estimates which would be beneficial in a production environment, making it the practical choice for property portals. This means that yearly cross-sectional models are developed. All statistical hypothesis tests used had a level of significance of 0.05.

3.1 Gradiant Boosting and H2o

Statistical learning is a recent development in the field of statistics. It leverages 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 & Johnson, 2018). Supervised statistical learning involves developing predictive models on training data that generalize to unseen holdout data (Hastie, Tibshirani & Friedman, 2005). Boosting is an example of supervised statistical learning where decision trees are grown sequentially using information from previous trees. Boosting is a technique of improving a learning algorithm which executes repeated iterations of a weak learner by constructing decision trees sequentially from the residuals (Freund & 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:

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \qquad (4)$$

Where α_t is the coefficient with which the ensemble h_t is combined, α_t and h_t are learned through the boosting procedure (Meir & Ratsch, 2003). The boosting 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.

H2o.ai is a 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). Gradient Boosting Models (GBMs) are part of the H2o.ai stack that can be developed using different programming languages such as R and Python, or the easy to use H2o.ai flow web interface for non-programmers. Gartner (2018), a global research and advisory firm, named H2o.ai a leader amongst 16 vendors in their '*Magic Quadrant for Data Science*'. H2o.ai describes GBMs as forward-learning, non-linear ensembles of tree-based models where weak trees are sequentially grown from the incrementally changed data, resulting in an ensemble of weak prediction models that gradually improve estimations

of a response variable iteratively. Key features for using the H2o.ai implementation of GBM in this study include the ability to fit exponential families of distributions, automatic early stopping based on convergence of a specified metric and the use of stochastic GBM which improves generalization through column and row sampling during model building (Friedman, 2002; Click et al., 2016). R or Python scripts using the H2o.ai functionality can be embedded into backend 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.ai implementation of the GBM algorithm portable and interoperable for organisations like property portals.

4. Model Evaluation and Selection

Two model validation approaches are adopted in this study, firstly splitting the data into training and validation sets, and secondly, the use of cross validation. The log linear models and GBMs are built using the training data and evaluated on the holdout (validation) data. Cross validation is then applied to the GBMs to optimise the hyperparameters, with the aim of reducing the out of sample error. Cross validation is not applied to the log linear models as hyperparameters are not applicable, coefficients are estimated through minimizing the sum of squared residuals (Greene, 2003). In both approaches, the root mean squared error (RMSE) is used to test model fit and generalizability. RMSE measures the closeness of model estimates to the observed data (Gujarati, 2004).

The data splitting procedure involved splitting the data into training and holdout sets for each respective year, where 70% of data was used for training and the remaining 30% used to test model generalizability as unseen holdout data. The holdout error provides a robust estimate of model generalizability (Blum, Kalai & Langford, 1999). A function was written to ensure that the assignment of data to the yearly splits was random and that distribution of the response was similar for each split and to the original sample. Finally, for each year, the function kept each area present in each split.

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 minimize the average prediction error (Hastie, Tibshirani & Friedman, 2001). Searching over a high dimensional hyperparameter space to find the optimal combinations thereof can be computationally expensive. 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 & 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 root mean squared error (RMSE) does not improve for 25 training rounds based on a moving average of 10,000.

Evaluation of model generalization 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 illustrates an example of 5-fold cross validation.

Table 5: 5-Fold cross validation structure

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Training Set	Validation 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 summarized 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:

$$CV(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}^{-\kappa(i)}(x_i)\right) \qquad (5)$$

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

$$CV(\hat{f}, \alpha) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}^{-\kappa(i)}(x_i, \alpha)\right) \qquad (6)$$

Cross validation can be applied to models with many tuning parameters to search for the combination of hyperparameters that produce the lowest prediction error. The 5-fold cross-validated GBMs are built on 80% of the data with 20% withheld as the final validation set, making the generalization framework robust (LeDell et al., 2019). Yearly property listings from all respective areas 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 & 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, meaning that 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.

The RESET test, proposed by Ramsey (1969) is applied to the log linear models, designed to detect inappropriate functional form (Shukur & Mantalo, 2004). Under the alternative hypothesis, a model generated by taking powers of the covariates has significant influence (Ramsey, 1969). GBMs do not make any explicit functional form, with the aim of keeping estimates on the original scale, the gamma distribution is used to estimate listing prices, assuming the canonical link function. This will result in arithmetic mean estimates where no back transformation is necessary. Linear models assume that the coefficients combine linearly with the covariates, the best way to investigate these relationships is through a graphical assessment using partial residual plots. A partial residual plot results in a bivariate scatter plot which removes the effect of other covariates except the one of interest and describes its relationship to the response through model residuals (Fox & Weisberg, 2018). These diagnostic plots are used to evaluate the fit of the log linear models. Similarly, Partial Dependence Plots (PDP) are developed for the GBMs to understand the effect of the covariates on the response. PDP's are a useful interpretation tool for 'blackbox' machine learning algorithms which plot the marginal effect of a covariate on the response holding other covariates constant (Friedman, 2001; Hastie et al., 2009).

The estimation of residential hedonic price functions often suffers from spatial autocorrelation, manifesting in correlation of the residuals in regression models (Bourassa et al., 2007). This violates the assumption of independence and needs to be checked. Model fit and validation techniques using random data splits are only reliable for situations where assumptions of independence are checked and in non-extrapolation cases (Roberts et al., 2017). The Moran I test (1950) is used to test for spatial autocorrelation of the residuals which simply measures how the residuals behave in two dimensional space (Anselin 2006). The coefficient ranges from -1 to 1 which shows the strength and direction of spatial autocorrelation. In the case of this study the alternative hypothesis is that positive spatial autocorrelation exists. Positive spatial autocorrelation is when high or low value properties tend to cluster together in space.

5. Results and Discussion

The goodness of fit and diagnostics of the log linear models are presented in Table 6. The RMSE and R_2 statistics are reported as the measures of goodness of fit. The R_2 indicates how much variation in listing price is explained by the variation in the physical and locational characteristics.

Table 6: Log Linear Model Summary and Diagnostics

Year	R_2	Training RMSE	Holdout RMSE	RESET Test p-value	Moran's I Statistic	Moran's I p-value
2014	0.87	766,117	762,544	1.38e-54	-0.029487	0.99
2015	0.87	767,251	774,243	1.22e-56	-0.024477	0.99
2016	0.87	714,015	741,349	1.4e-49	-0.025836	0.99
2017	0.88	724,097	726,167	3.34e-38	-0.030789	0.99

Notes: R_2 has been rounded to two decimal places and RMSEs to the nearest whole number.

The R_2 measures are very high, showing that 87% to 88% of the variation in listing prices is explained by the variation in the explanatory variables. Overall, the log linear models appear to generalize to the unseen, holdout data well, showing robustness. The RESET tests indicate that there is sufficient evidence to reject the null hypothesis of correct specification of linear functional form. These results are congruent to the findings of Els and Von Fintel (2010) who turned to quantile regression after log liner models failed to satisfy the functional form requirement. Unfortunately, the test provides no direction on how to proceed if the model is rejected, however, the partial residual plots shown shortly may provide some guidance. The Moran I specification test shows that there is not enough evidence to reject the null hypothesis in favour of the alternative, positive spatial autocorrelation. The test statistic shows a weak negative correlation between residuals in space. This means that including a fixed effect location variable accounted for the spatial dependency in the data, a similar finding to Bourassa et al. (2007) and Bax and Chasomeris (2019). Examining the residual diagnostic plots is vitally important for parametric models where the assumptions are checked. Figure 1 illustrates the fitted versus residual and quantile-quantile plots for each yearly model.

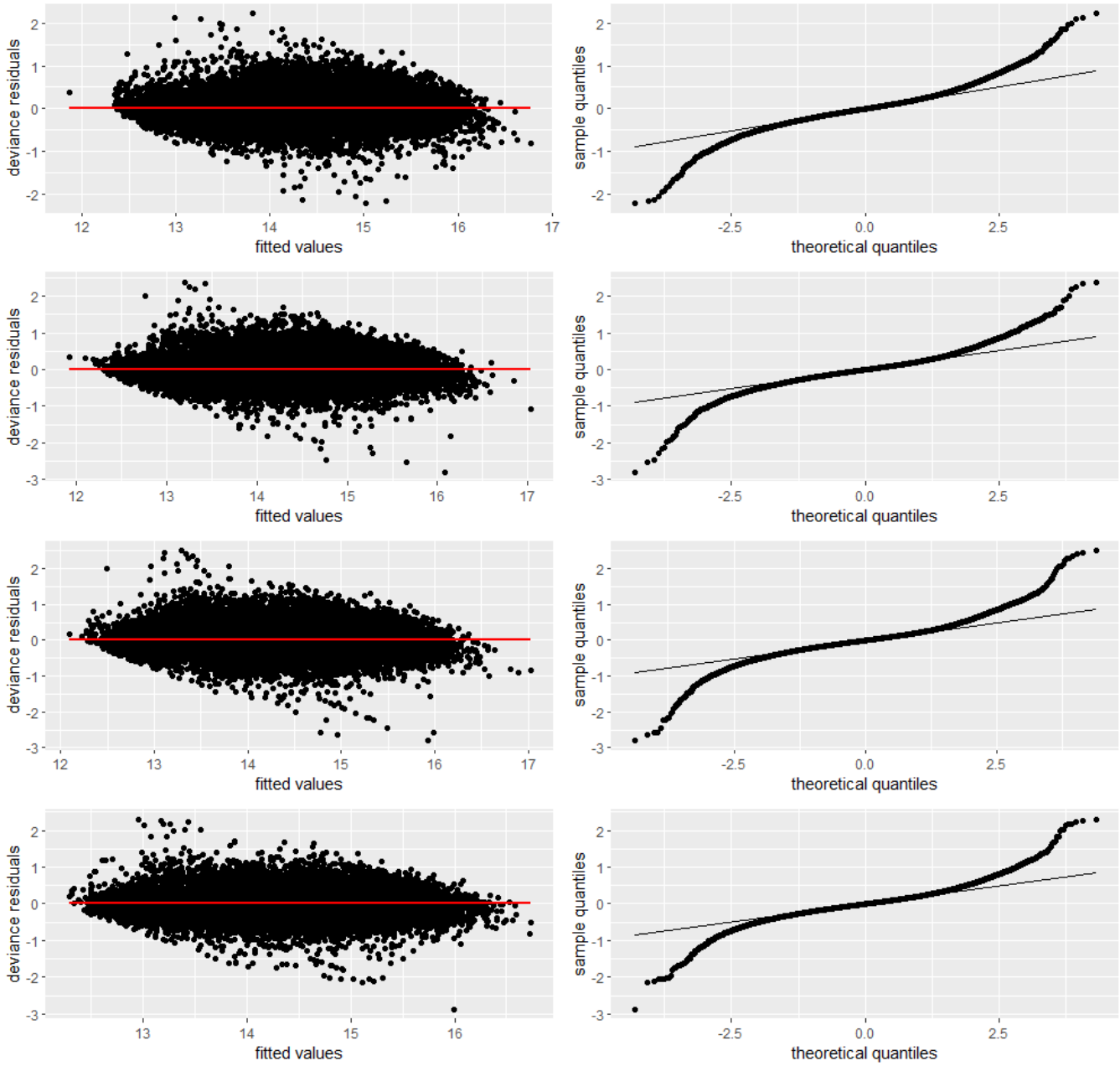

Figure 1: Log Linear Goodness of Fit Plots

Notes: Each row represents a yearly model beginning at 2014 and ending at 2017.

The fitted versus residuals appear homoscedastic, meeting the assumption of constant variance, though the quantile-quantile plots show deviation from normality at the upper and lower quantiles indicating the residual distribution is heavy tailed. However, Schmidt and Finan (2018) provide empirical evidence that linear models without normally distributed residuals may still provide valid results, given sufficient sample size. Figure 2 details the partial residual plots for the yearly log linear models.

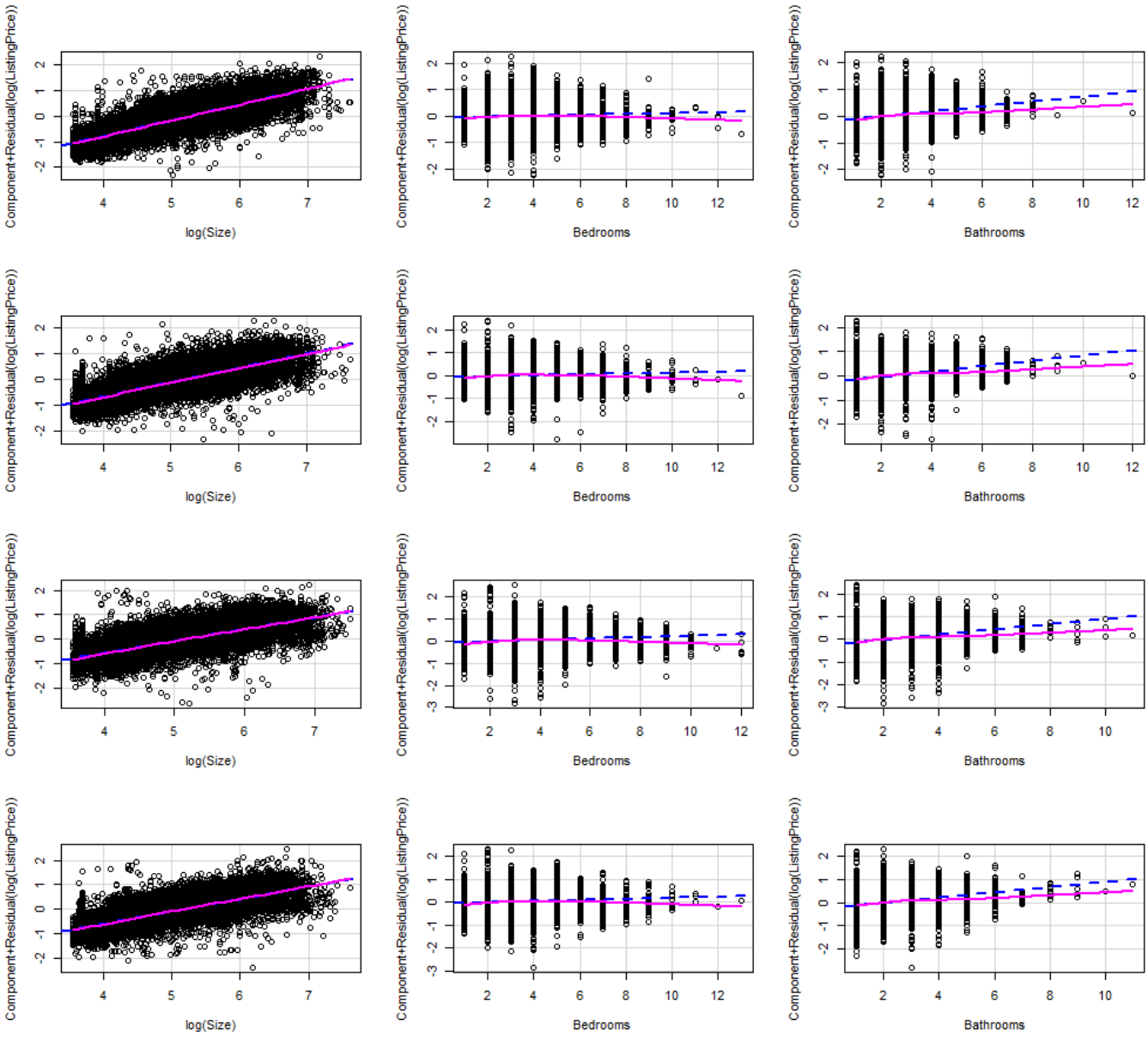

Figure 2: Log Linear Partial Residual Plots

The plots show a positive linear relationship between the log of listing prices and the log of size. The natural logarithm was applied to the size variable to improve linearity which was an appropriate choice given the partial residual plots above. The number of bathrooms shows greater utility over the marginal distribution compared to the number of bedrooms. This means, on average, sellers can expect greater utility from additional bathrooms. The number of bedrooms and number of bathrooms are not linearly related to listing prices suggesting a transformation may be appropriate.

The GBMs using the default hyperparameters are presented next, Table 7 shows the goodness of fit for the training and holdout sets without applying cross validation.

Table 7: GBM Default Hyperparameter Summary

Year	Training RMSE	Holdout RMSE
2014	750,247	779,314
2015	747,383	790,662
2016	733,249	762,224
2017	717,520	753,224

Although the GBMs generalize to the unseen data, the log linear models achieve a lower holdout RMSE for each respective year, indicating the GBMs require tuning. A 5-fold cross validation is applied next using a random grid search and early stopping to find the optimal combination of hyperparameter to reduce the holdout error. The results of the 5-fold cross validation yearly GBMs are presented in Table 8.

Table 8: GBM Cross Validation Summary

	Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
Year	RMSE	R_2	RMSE	R_2	RMSE	R_2	RMSE	R_2	RMSE	R_2
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	680,751	0.84	673,863	0.84	686,535	0.84	661,416	0.84

The goodness of fit measures in each fold for each respective yearly GBM are quite consistent showing that the models generalize well. Combining the holdout predictions to gauge an unbiased overall average fit is presented in Table 9.

Table 9: GBM Combined Holdout Error Summary

Year	Combined holdout prediction RMSE	Improvement from log linear model	Moran's I Statistic	Moran's I p-value
2014	717,398	6%	0.00789	0.031
2015	711,415	8%	-0.00775	0.970
2016	660,072	11%	-0.00242	0.706
2017	674,687	7%	-0.00520	0.903

The holdout errors are fairly consistent for each yearly model with 2016/7 producing the lowest generalization errors. The holdout errors are lower using cross validated gradient boosting with a random hyperparameter sweep, showing that this framework has a lower prediction error benefit. The holdout RMSE for the cross validated GBMs are slightly lower compared to the log linear models, shown in the improvement from log linear model column. These findings are similar to those of van Wezel et al. (2005). The Moran's I test shows that GBMs account for the spatial dependency in the data with the exception of 2014 where we observe statistically significant positive spatial autocorrelation, although the strength of this dependency in the residuals is very weak.

The random grid search applied to each yearly GBM allowed for different hyperparameters to be selected for different the models. Table 10 details the hyperparameters chosen in the final models with summary statistics about each tree.

Table 10: GBM Model Summaries

Year	Number of Trees	Sample Rate	Column Sample Rate per Tree	Learn Rate	Min Depth	Max Depth	Mean Depth	Min Leaves	Max Leaves	Mean 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. The year 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's in this study summarises 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 3 shows how the mean response changes with a change in the given numeric covariate, namely: log size, bedrooms and bathrooms.

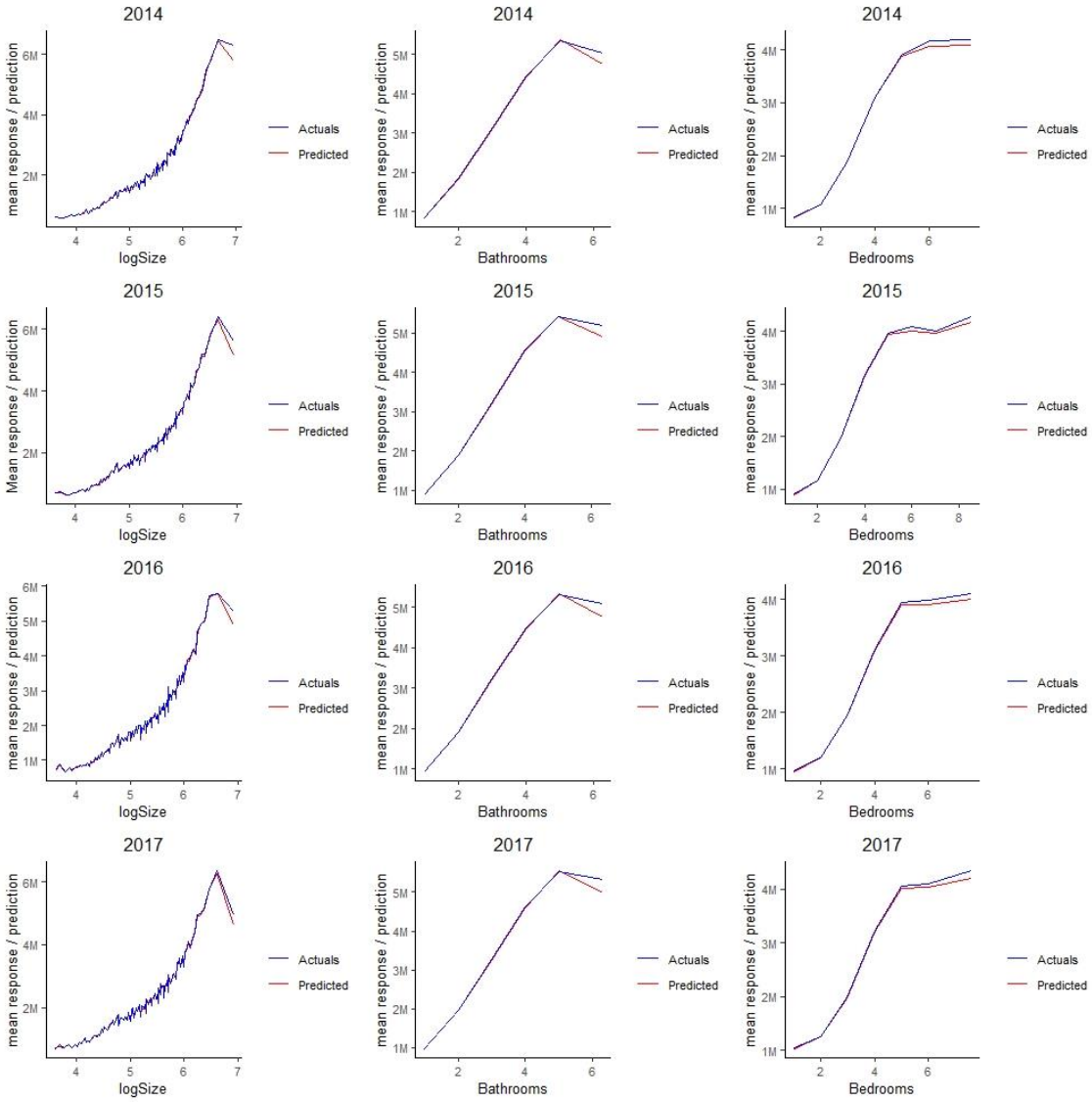

Figure 3: Partial Dependence Calibration Plots

The yearly log size curves share a similar shape where tapering is evident. The utility increases steeply initially but then drops over the marginal distribution. This suggests that larger sized properties, greater than ≈800m2 experience a diminishing return. The marginal utility of bedrooms is positive up to 5-6 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 up to 5 bathrooms where additional bathrooms added no extra value. The yearly PDP plots 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 4. Friedman (2002) applied variable importance to GBMs leveraging the work of Breiman (2001) who used randomization of the out-of-bag observations which are

observations held back during random forest training and applied before the algorithm has completed.

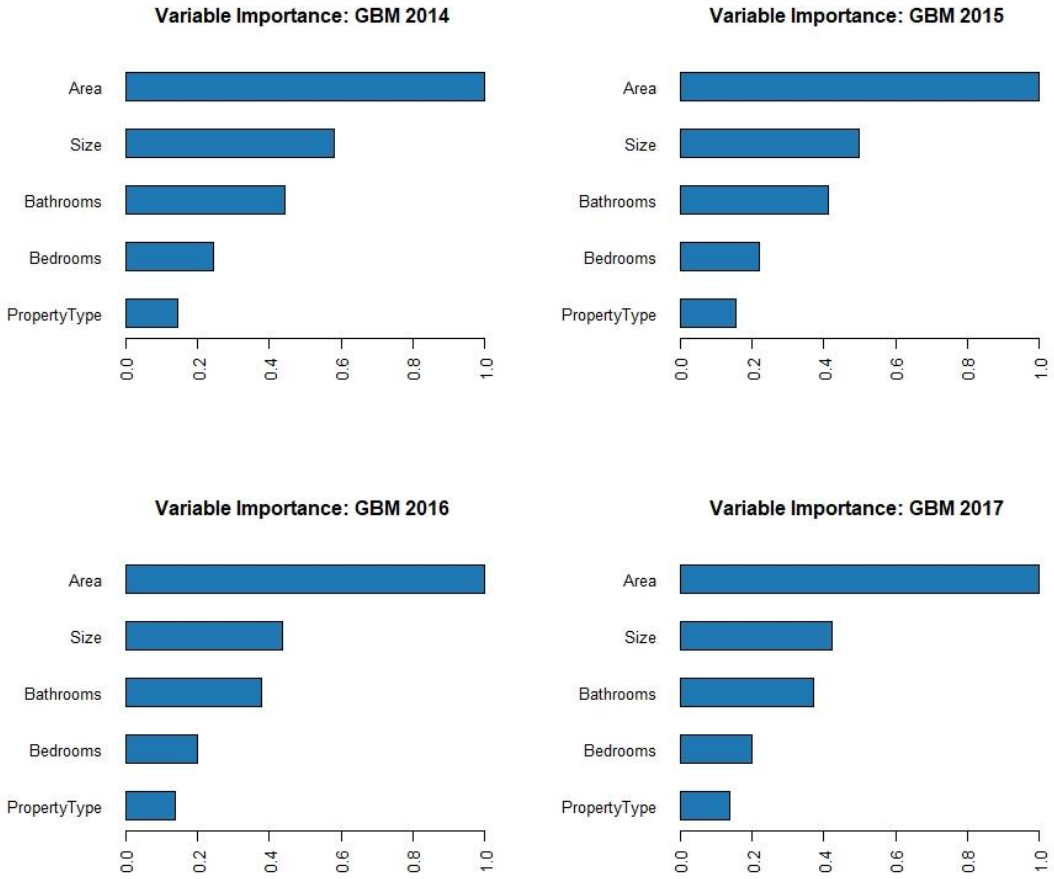

Figure 4: Variable Importance Plots

The area 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.

Zillow, a US based property portal, provides its users with a simple online interface to obtain property valuations using a proprietary algorithm (Zillow 2019). Private Property (Pty) Ltd could leverage the data they collect and store and provide a similar service, using the framework proposed in this study. To demonstrate this, a web application was developed which provides the ability to obtain mean listing prices and gauge listing price growth. Figures 5 and 6 present the listing price and growth calculator application created for this study.

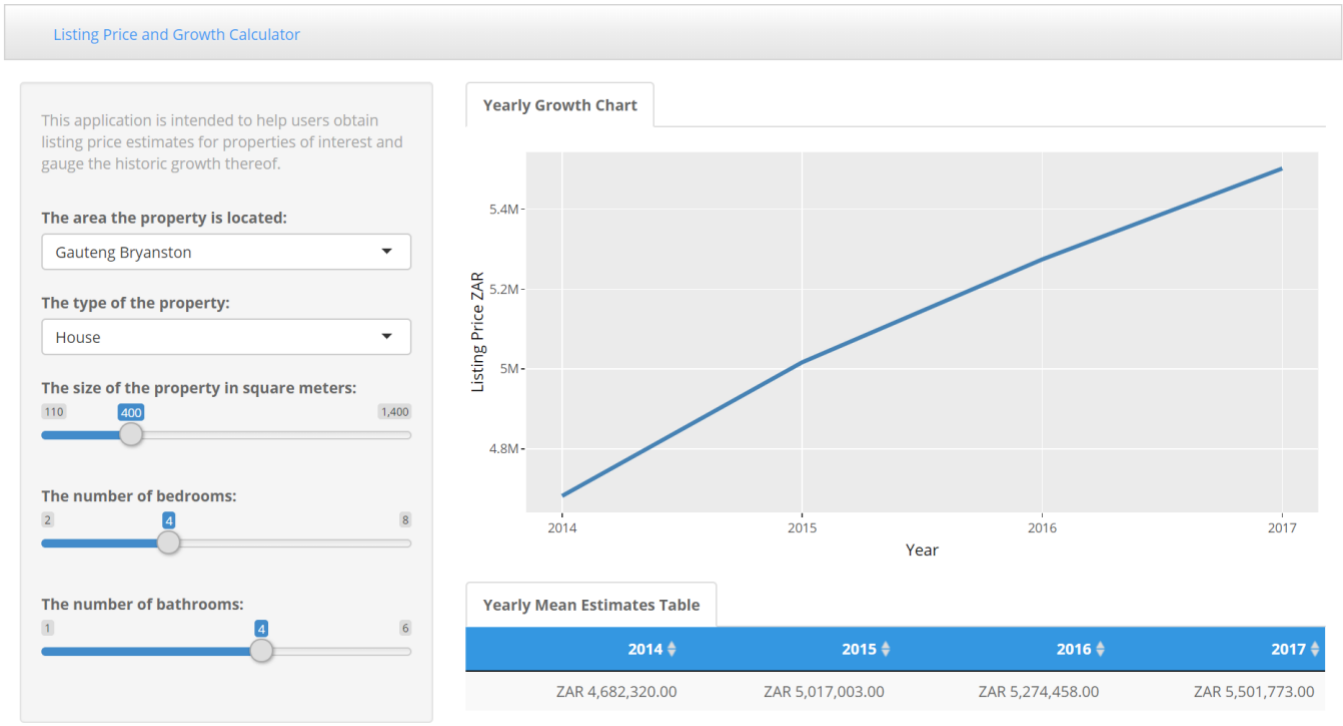

Figure 5: Listing Price and Growth Calculator (example one)

This example shows the mean price estimates for a house in Gauteng Bryanston that that is 400m2 in size with 4 bedrooms and 4 bathrooms. The mean estimates are plotted in a chart over time, and tabulated.

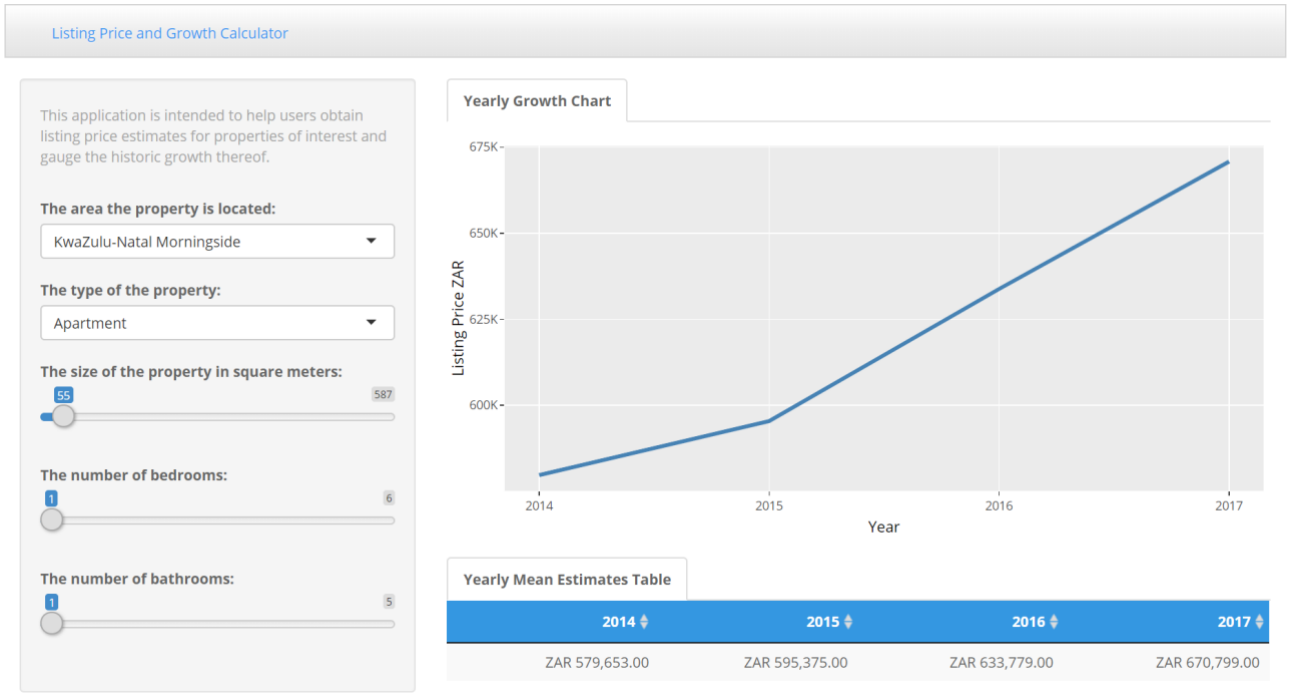

Figure 6: Listing Price and Growth Calculator (example two)

The second example illustrates the mean price estimates for an apartment in KwaZulu-Natal Morningside that is 55m2 with 1 bedroom and 1 bathroom.

The application was built using the GBMs and provides yearly mean estimates to a user given the location and physical characteristics of interest. This means that someone wishing to sell their home can quickly gauge what listing price to set by simply inputting where the property is located and some of the properties physical characteristics. This is can be a convenient alternative or first source of information for a potential seller, compared to trawling a plethora of listings or using the services of an estate agency or paid service.

6. Summary

Determining what price a home will sell for on the market through the reconciliation of supply and demand is challenging and is further compounded for sellers by the multitude of available sources of information. Typically, the services of real estate agents are employed where comparative market analyses are used to produce listing price estimates. This study proposed an algorithmic solution which can be used as an alternative to, or in conjunction with, traditional comparative market analysis methods.

Various studies exist that explore the use of parametric, semi-parametric and non-parametric techniques to estimate residential property prices for different segments of the South African market. This study contributes to the existing real estate pricing literature by developing parametric and non-parametric hedonic price models for different property types throughout South Africa. Traditionally, log linear models have been widely used, both globally and locally, to estimate residential property prices, measuring the utility over the marginal distribution of physical attributes and the effects of categorical variables. Although the framework provides transparency, it often suffers from misspecification of functional form. This study developed and compared yearly hedonic price functions using log linear and Gradient Boosted Models (GBMs). The log linear models seemed to provide a good fit, however, testing whether the functional form was correctly specified resulted in the violation of this assumption, which is congruent to previous South African studies. GBMs were chosen as a flexible alternative. The 5-fold cross validated GBMs outperformed the log linear models, providing a lower out of sample error. Both approaches were able to account for the spatial dependency adequately in the data by including a location categorical variable.

Developments in making the results of ‘blackbox’ 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 don’t necessarily yield higher prices and result in diminished returns. The area location variable was consistently deemed the most important followed by size and the number of bathrooms, reinforcing the old adage about the importance of location and property. A key feature of this study was to

develop a framework to democratise the proposed methodology, showing how property portals or real estate agencies could leverage their data to guide home owners on what price to sell their homes for. A web application was developed that allows a user to simply select the location and physical characteristics of the property of interest and easily obtain mean price estimates and the growth thereof.

Future work could involve the construction of a price index extrapolating to new spatial structures with spatial models or blocking cross validation.

References

Anglin, P. & Gençay, R. (1996). Semiparametric estimation of a hedonic price function. *Journal of Applied Econometrics*, *11*(6), pp.633-648.

Anselin L. (2006). Spatial econometrics. In: Mills T, Patterson K (eds) *Palgrave handbook of econometrics: Volume 1, Econometric Theory*. Basingstoke: Palgrave Macmillan.

Bax, D. & Chasomeris, M. (2019). Listing price estimation of apartments: A generalised linear model. *Journal of Economic and Financial Sciences*, *12*(1). pp.1-11.

Bergstra, J. & 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. (1999). Universal portfolios with and without transaction costs. *Machine Learning*, *35*(3), pp.193-205.

Bourassa, S.C., Cantoni, E. & 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.

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

Click, C., Malohlava, M., Parmar, V., Roark, H. & Candel, A. (2016). *Gradient Boosted Models with H2O*. Available at: http://h2o.ai/resources/.

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.

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

Du Preez, M., Lee, D. & 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. & 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. & 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 13th 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., & Malohlava, M. (2019). h2o: R Interface for H2O. R package version 3.22.1.1. Available at: https://CRAN.R-project.org/package=h2o.

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

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

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

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

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

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

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

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

Jiang, L., Phillips, P. & Yu, J. (2015). New methodology for constructing real estate price indices applied to the Singapore residential market. *Journal of Banking & Finance*, *61*, pp.S121-S131.

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

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

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.

Pace, K. R. (2008). Appraisal Using Generalized Additive Models. *Journal of Real Estate Research*, *(1/2)*, pp.77-99.

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/.

Ramsey, J. B. (1969). Test for Specification error in Classical Linear Least Squares Regression Analysis. *Journal of the Royal Statistical Society*, Series B, *31*, pp.350-371.

Meir, R. & 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.

Moran, P. (1950). A test for the serial independence of residuals. *Biometrika*, *37*(1-2), pp.178-181.

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

Private Property (2019). *Comparative Market Analysis helps you sell faster | Private Property*. [online]. Privateproperty.co.za. Available at: https://www.privateproperty.co.za/advice/property/articles/how-a-comparative-market-analysis-helps-you-sell-faster/4203 [Accessed 28 Nov. 2019].

PropertyFox. (2019). *Determining Best Property Price | Valuations for Selling*. [online]. Available at: https://propertyfox.co.za/determining-best-property-sales-price/ [Accessed 28 Nov. 2019].

Roberts, D., Bahn, V., Ciuti, S., Boyce, M., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J., Schröder, B., Thuiller, W., Warton, D., Wintle, B., Hartig, F. & Dormann, C. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), pp.913-929.

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

Shimizu, C., Nishimura, K, & 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. & Finan, C. (2018). Linear regression and the normality assumption. *Journal of Clinical Epidemiology*, *98*, pp.146-151.

Shukur, G. & 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.

Statistics South Africa. (2019). *Mid-year population estimates 2018 | Statistics South Africa*. [online] Statssa.gov.za. Available at: http://www.statssa.gov.za/?p=11341 [Accessed 16 Aug. 2019].

Trachsel, M. & Telford, R. J. 2016. Technical note: estimating unbiased transfer-function performances in spatially structured environments. – Climate Past 12: pp.1215–1223.

van Wezel, M, M Kagie, & 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).

Zhong, Z., Yan, J., Wu, W., Shao, J. & 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]