School of Accounting, Economics and Finance

Machine learning and residential real estate: Three applications

Zhuoran (Thomas) Zhang

0000-0002-2110-6221

This thesis is presented for the Collaborative Degree of

Doctor of Philosophy

of

Curtin University

and

University of Aberdeen

April, 2024

Declaration

To the best of my knowledge and belief, this thesis contains no material previously published by any other person except where due acknowledgment has been made. This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature:

Date: 3rd April, 2024

Abstract

The applications of machine learning techniques have become popular recently in residential real estate, specifically, the valuation of residential properties. This thesis is a composite of empirical studies for three topics, automated valuation model (*AVM*), residential property price index (*RPPI*), and the analysis of land development. Firstly, machine learning techniques are applied to develop the implementations of *AVM*s, whose purpose is to provide a price estimate of a particular property at a specified time. The main objective is to minimize human intervention in price estimation when the presence of missing values remains a major challenge in the process. Then, the proposed *AVM* implementation is applied for compiling the residential property price index, which tracks the trend of market values, cooperating with the classic indexing approaches. The main objective is to investigate whether more accurate price predictions lead to a better price index and examine how well the machine learning techniques explain the time effects. Thirdly, land development is a "real option" that provides the landowner a right to decide whether and when to develop the vacant land by spending amount of money. The analysis of land development is to examine the real option, including the valuation of the option and the optimal timing to exercise the option. The analysis is conducted using machine learning techniques with the factors on both the investment output (residential buildings) side and the investment cost (construction cost) side, such as the growths and uncertainties of property prices and construction costs.

The empirical results show that the proposed *AVM* implementation in topic one can predict more accurately, meanwhile, the missing values in the data are well handled simultaneously. This is consistent with the results of recent studies. In addition, the residential property price index compiled using the machine learning *AVM* implementation is a competent alternative to the *RPPI* published by the Australian Bureau of Statistics. For land development, the uncertainty of construction costs consistently presents a positive effect on the land value, a risk premium is added on the land price. Meanwhile, the higher price of residential buildings encourages the landowners to develop. On the contrary, the higher construction cost delays the development project.

Keywords: *Machine learning techniques, residential real estate, automated valuation model (AVM), Residential property price index, Land development*

Acknowledgements

The completion of this Ph.D. thesis is only possible with the encouragement and support from my supervisors, my family, and my friends. I would first like to express my sincere gratitude to my supervisory team, Professor *Felix Chan*, Associate Professor *Greg Costello*, and Dr. *Rainer Schulz*, for their continuous guidance, motivation, dedication, and patience during the completion of my Ph.D. study, especially during the COVID-19 pandemic. Not only their contributions to my Ph.D. study are significant, but also their guidance on my academic ethics and life attitude are valuable.

My deep appreciation also goes to my parents, *Yimin Zhang* and *Yuyan Yang* for always encouraging and believing in me during my Ph.D. journey. Their support provides me with strength and courage when I doubt myself. I especially want to thank my friend and my partner Dr. *Ruolin Wang* for her selfless devotion, support, and understanding to me. Please spend your whole life with me.

Finally, I owe a debt of gratitude to my long-time friend Dr. *Zhiyue(Victor) Sun* for the plenty of meals we ate out; and my colleagues: Professor *Steven Rowley*, Associate Professor *Amity James*, Dr. *Joye Khoo* and Dr. *Siwen Song* for their constant help and excellent advice. I gratefully acknowledge the funding received for my Ph.D. from the Aberdeen-Curtin Alliance.

Contents

List of Figures

List of Tables

Abbreviations

ABS	AustraliaN Bureau of Statistics
AFT	Accelerated Failure Model
ALE	Accumulated Local effect
AVM	Automated Valuation Model
CAPM	Capital Asset Pricing Model
GAM	Generalized Additive Model
GARCH	Generalized Auto-Regressive Conditional Heteroskedasticity
GBM	Gradient Boosting Machine
LS	Squared error loss function
LAD	Absolute deviation loss function
LGA	Local Government Area
MAPE	Mean Absolute Percentage Error
MAE	Mean Absolute Error
NA	Not Applicable
OLS	Ordinary Least squares
PER	Percentage Error Range
PDP	Partial Dependence Plot
PDE	Partial Differential Equation
RMSPE	Root Mean Squared Percentage Error
RMSE	Root Mean Squared Error
RPPI	Residential Property Price Index
VAR	Vector Auto-Regression

Chapter 1

Introduction

1.1 Research background

Real estate is defined as the combination of land and any permanent structures built on the land, like a home, a factory, or any other improvements attached to the land, whether natural or man-made. Residential real estate is the most common type of real estate, which is used for residential purposes — that is, developed specifically as a place for people to live. Residential real estate is an important asset class, relevant to households and policymakers. To individual households, purchasing residential real estate is often the largest and most important financial investment made in their lifetime. Some individuals invest in residential real estate as a money-making venture, either by selling it for a profit or renting it out to tenants for a regular passive income. But most people simply live on their property. To policymakers, residential real estate is a critical driver of economic growth and residents' welfare in Australia, especially since Australia is an immigration country. Building approvals, building activities, and construction work done, released by the Australian Bureau of Statistics (*ABS*) monthly or quarterly, are key economic indicators. They include building permits, dwellings commenced, under construction, and construction completion information. These numbers can provide a general sense of economic direction. If building approvals indicate fewer houses in the private sector, it could signal an impending supply shortage for houses shortly, which could drive up house prices.

This thesis comprehensively studies residential real estate by focusing on the valuation of residential properties. Because the values of residential properties are necessarily considered by all the stakeholders in the housing markets.

To individuals, the prices of residential properties mostly are the determinant of whether the households would buy or not. To policymakers, the prices always reflect the market trends. Most importantly, they are the indicator of the demand and supply of residential properties. Screening the housing markets is equivalent to regularizing the prices of residential properties. Such that, this thesis examines three standard questions in real estate research about valuation, automated valuation models, index construction, and land valuation. The automated valuation model (AVM) studies the valuation for individual residential properties, which makes the valuation process much easier for individual households. The price index of residential properties provides market trends to households and policymakers. The price index continuously screens the growth and decline of the local market as a key economic indicator. The value of the vacant land for residential purposes is also necessary to investigate because it is a significant contributor to dwellings. In addition, vacant lands are the origin of real estate.

However, it is difficult to analyze the prices of residential properties and provide reliable information about what goes on in each of them. There are several reasons. Initially, the transactions of residential properties take place in the private markets. Some details of transactions are related to privacy, thus, they are not available for public access. Secondly, residential properties are heterogeneous. There are diverse types of residential properties, such as apartments, home units, detached houses, townhouses etc. In addition, every residential property is unique on this earth, they have diverse designs, locations, and different facilities. Last but not least, the housing markets are local. For example, in Australia, what goes on in Sydney may have no bearing on what goes on in Perth.

These difficulties could be addressed. Naturally, transaction-level data are required for analyzing the local housing market. In such data, normally, the details of transactions and the details of the transacted residential properties are available. In addition, the history of transactions could be chained by the identities of the residential property, such as address or lot ID. Commonly, the privacy of sellers and buyers is hidden. Thereafter, the analysis of such data requires the use of the appropriate methodology. Whereas the linear regression model and the ordinary least squares estimator have been the workhorse for such analysis in residential real estate research for decades. Recently, machine learning techniques come to the public attention, they have added further methods that could be used for diverse analysis instead, which provide more flexibility. They could be more competent than the workhorse model. For the housing market, one relatively closed market could be selected for conducting the analysis. The residents in that market do not have an alternative. Perth is selected, as it is the most “lonely” city in this world. It takes a five-hour flight to the nearest city. Thus, the Perth housing market is relatively closed. Meanwhile, the local housing market of Perth has not been examined in detail that much, compared with Sydney and Melbourne. The data, the appropriate method, and the market, this thesis is conducted after solving these difficulties.

1.2 Research motivation

The motivation to investigate the three standard questions about the valuation of residential properties, automated valuation model, index construction, and land valuation, lies in the following aspects.

Firstly, the automated valuation models are well applied to evaluate individual residential dwellings, which are conducted on the prepared transaction-level data. The accuracy of price predictions is guaranteed, especially when applying machine learning techniques. However, there is a significant issue in academic research and industrial practice, missing values in transaction-level data. In academic research, missing values lead to the hardness of modeling. Some studies choose to simply drop the observations or variables that include missing values, the others have to impute the missing values or estimate multiple regressions with variate combinations of variables, due to missing values (For example, Hill and Scholz 2018, Steurer et al. 2021). In industrial practice, the influence of this issue is more significant. Imagine a situation, where an *AVM* user, such as a valuer or mortgage broker, needs to predict the price of a residential property. If some variables are not observed or available, the *AVM* may not provide a price prediction due to missing characteristics. Such an issue significantly affects the implementation of *AVM*, and its deployment.

Unluckily, the transaction-level data in Perth are as "dirty" as the others. Because local governments or private entities (buyers, sellers or agents) choose not to, or fail to, report critical housing information or because the information is not available. Meanwhile, missing values could be caused by improper data collection or mistakes in data entry. What to do with observations that suffer from missing values? This thesis is motivated to find a method that could overcome this issue. Commonly, there are two available solutions. The first is to avoid all observations or all the variables containing missing values. The process is simple, but, it normally leads to a huge data loss. The second is to impute missing values using the observed information. The data loss is avoided, but the process is complex and time-consuming, especially, when the presence of missing values happens in the forecast problem. Meanwhile, the imputation of missing values has some prerequisites, such as checking the mechanism of missingness. Such that, the best method should be required to avoid a huge data loss and have a simple process.

Secondly, the residential property price index (*RPPI*) has an important use as a macroeconomic indicator of economic growth. Some classic index compilation approaches are well summarized, such as Eurostat (2013). For compiling a price index, some approaches rely on price predictions, such as the hedonic imputation approach. However, it is unsure whether the higher predictive accuracy could benefit the quality of the residential property price index. Recently, it has been demonstrated that machine learning techniques could significantly improve the accuracy of price predictions (Mayer et al. 2019, Schulz and Wersing 2021). However, relatively less literature discussed the potential applications of machine learning techniques for the residential property price index. Additionally, a question remains, whether the highly accurate price predictions will benefit the quality of the residential property price index.

In 2022, the Australian Bureau of Statistics stopped publishing the residential property price index to the public for unknown reasons. To fill the research gaps and find a competent alternative for the *RPPI* published by *ABS*, it is worthwhile to investigate the combinations of machine learning techniques and classic index compilation approaches. Meanwhile, it could be worth examining whether it is possible to interpret the time effects in the indexing implementations applying machine learning techniques when they are well-known as the "black box".

Thirdly, a residential land parcel has its life cycle, from construction to destruction. The beginning of this cycle is land development, the decision, of whether to develop or not, is a real option. At this moment, the valuation of vacant land is similar to the valuation of financial options. There are effects from two sides, the investment cost side and the investment output side, which could affect the value of the vacant land. The past literature has investigated land development theoretically and empirically. In the empirical analysis, the option premium and the effects from the investment output side are well discussed (For example, Quigg 1993, Cunningham 2006). However, the effects from the investment cost side are rarely considered, which may not be appropriate. As proposed in the Australian Bureau of Statistics report (Australian Bureau of Statistics 2020b), the spending on construction is not always the same as planned. In Perth, almost half of newly constructed houses cost more to build than they were approved for, while 25% of newly built houses cost less. Only 31.2% of constructions are with on cost-changes. Landowners prefer that the construction costs are less or equal to their planned budget, and avoid over-spending. Thus, it is necessary to figure out the effects of the factors from the investment cost side.

To address the effects from the investment cost side, the growth and uncertainty of construction costs are considered in the empirical analysis of the land valuation. Such that, the vacant land valuation is comprehensively studied on both sides, this fills the research gap. Meanwhile, it is helpful to understand the factors that could affect the landowner's decisions regarding land development.

1.3 Research objectives

This thesis focuses on the research questions related to the valuation of residential properties. It aims to address three research questions:

1. What to do with observations that suffer from missing values when estimating prices of residential properties using an automated valuation model?
2. Is the automated valuation model with machine learning techniques always better than the linear model (the workhorse mentioned above) when compiling the price index? Whether more accurate price predictions lead to a better price index? How well do the machine learning techniques explain the time effect on valuation?
3. What are the effects of the factors, that are from the investment cost side, on the valuation of vacant land?

These three questions are progressively related. The data issue must be solved first, otherwise, the automated valuation model can not provide any accurate price predictions. Such that, the foundation is built for the second question when the first question is addressed. To investigate the third question, the growth and the uncertainty of house market prices are the necessary factors on the investment output side, which could be derived from the price index compiled in the second question.

Firstly, *to comprehensively study the implementations of AVM with ma-*

chine learning techniques and the linear models when missing values occur in data. In this topic, the combinations of estimation models and missing value strategies are investigated, which have not been paid much attention in the literature. For dealing with missing values, three strategies are examined. Two of them are commonly used solutions, the complete case strategy and the imputation strategy; the rest is only for machine learning techniques, missing value node strategy. Meanwhile, two commonly used loss functions are deployed in the price estimation, the least squares loss and the lease absolution deviation loss. In total, eight implementations are proposed and investigated, and the best one is recommended in this topic.

Secondly, *to examine whether it is achievable for machine learning to be applied for compiling residential property price index, and to investigate whether the accurate estimation benefits the quality of compiled indices.* This topic focuses on constructing the residential property price index using the hedonic regression method. The hedonic regression method relies on price estimation, and it is suitable for transaction-level data (Eurostat 2013). This topic does not only use the hedonic imputation approach, but it also applies a time dummy approach. The time dummy approach requires extraction of the time effects from the price estimation model. The ability to interpret the machine learning model is studied in this topic also. The contemporaneous correlation between compiled indices and the revision of each index are discussed to evaluate the compiled indices. The residential property price index issued by the Australian Bureau of Statistics (*ABS*) is the benchmark. Because it is also necessary to find an alternative, as *ABS* stops publishing the *RPPI*.

Thirdly, *to study the valuation of vacant lands on both the investment output* *side and the investment cost side, the timing of development is also involved in the study.* This empirical study is theoretically supported by the real option theory. As described by the theory, a piece of vacant land is like an American call option with no mature date. The option is exercised, which means the land is developed. The vacant land price and the optimal timing of development could rely on the explicit factors from two sides, the investment output side (dwelling price) and the investment cost side (construction cost), and the other implicit factors, such as the land sizes and locations. In past literature, the investment output side has been well studied, such as Cunningham (2006), and the other effects, for example, monetary policies (Wang et al. 2016). In this empirical analysis, the factors from the investment cost side are taken into account, as the *ABS* report suggests the construction cost is variable during the land development period.

These objectives on three research topics will be conducted and achieved in the Chapter 5 for implementing *AVM* with missing values in data, Chapter 6 for evaluation index compilation methods, and Chapter 7 for analyzing the vacant land valuation and optimal timing of land development respectively.

1.4 Research contributions

By achieving the research objectives, this thesis contributes to the existing residential real estate literature, as a thorough piece of analysis with different topics and diverse applications of machine learning techniques in the residential real estate field.

Firstly, the contributions to the price prediction for residential dwellings comprise the investigations of the different predictive performances of various machine learning implementations with two different loss functions and three missing value strategies. Especially, the solutions for the data issue of missing values are comprehensively studied, and this is the original and initial contribution of this topic. Since this data issue has not been paid much attention in the literature, meanwhile, it is quite common in residential real estate. This topic could be a pioneer that discusses the missing value issue using machine learning with different missing value strategies in the property valuation field. The recommended implementation is the best combination of the price estimation model and the missing value strategy, which has the most accurate predictions and efficient process when missing values occur in transaction-level data. Meanwhile, it provides the foundation for the second topic of this thesis.

Secondly, new compilation methods are provided for the residential property price index, those methods use the machine learning technique proposed previously and classic indexing approaches. Meanwhile, the compilation methods overcome the data issue of missing values, which could complicate the indexing process, for example, the situation faced in (Hill and Scholz 2018). This topic contributes to the literature by investigating the benefits of accurate price predictions on the residential property price index. It answers whether more accurate price predictions lead to a better price index. Uniquely, this study also indirectly studies whether machine learning techniques are interpretable. The temporal marginal effect should be extracted from the machine learning model for achieving a pseudo-time-dummy approach. Such that, this comprehensive study of the residential property price index using machine learning techniques fills in the gaps in the literature, meanwhile, it provides another option for price index compilation rather than using the “workhorse” model. Additionally, this topic presents a complete indexing process for academic research and practical use when missing values occur in transaction-level data.

Thirdly, another prevailing contribution of this thesis is to investigate the beginning of a parcel’s life cycle, land development. The factors that could affect the decision of land development are comprehensively considered on two sides, the investment output side and the investment cost side, depending on the real option theory. Compared with previous studies, this study takes construction costs into account. Usually, the growth rates of house prices and construction costs are assumed as a constant in real option studies. However, this study includes the uncertainty of the growth of prices and construction costs into real options analysis when the assumption can not be obeyed. It studies whether the stability of growth will affect the vacant land value and optimal timing of development when the assumption of constant growth is not true. This topic enhances and complements the empirical findings of vacant land valuation, and comprehensively corroborates the real option theory.

1.5 Structure of the thesis

The rest of the thesis has seven chapters. Chapter 2 summarizes the existing literature related to the three topics interested in conducting *AVM* using data with missing values, constructing residential property price index, and analyzing the value of vacant land and optimal timing of development. It starts with an overview of the literature about the first topic in Section 2.2. It reviews the application of machine learning techniques in the price prediction field for residential properties. Meanwhile, the data issue of missing values is identified in the literature. Section 2.3 is about the literature on the residential property price index. The index compilation methods are summarized, and their drawbacks are discussed. However, the machine learning application in the indexing field is fresh. Meanwhile, it is unsure whether the accurate predictions provided by machine learning would benefit the compiled index. The review for the land development is focused on the land valuation methods and decision-making for development inspired by real options theory in Section 2.4. The review of land development literature points out that the factors from the investment cost side have not been well studied yet. The summaries of the previous research works on these three topics provide the background of this thesis theoretically and empirically. Further, these comprise the current issues and the gaps of residential real estate on these three aspects, such as missing values, which will be solved and filled in this thesis by applying the techniques proposed in Chapter 4.

Chapter 4 introduces the methods used in this thesis. The chapter starts from the forms of tree-based machine learning models and model optimization to interpretation and evaluation, appending with the strategies for solving the data issue of missing values. Initially, it starts with the discussion of decision trees, which form the input to ensemble methods such as random forest and gradient boosting machine (GBM). Then, loss functions, evaluation criteria, and interpreting tools are described, those are the bundle of support tools for the transaction-level data analysis. Additionally, the strategies for the problems associated with the practical issues - missing values - are informed in the second half of this chapter. Chapter 4 addresses one of the mentioned difficulties in conducting residential real estate analysis, the appropriate methods.

Chapter 3 provides an overview of the Perth local housing market and the transaction-level data used in this thesis, which addresses other difficulties of residential real estate analysis, the market and the data. Additionally, the data for other specific purposes are also summarized. For instance, the data contain information about the annual summary of sales, that are cooperated for data preparation purposes. Missing values are not the only data issue. Some other issues are solved in data preparation, such as identifying non-arm length transactions. The detailed treatment process is described in the second half of the chapter. Missing values, the significant problem in transaction-level data, are retained and will be discussed and solved in Chapter 5.

Chapter 5 presents the empirical examination of AVMs that can provide accurate predictions while missing values occur in data. Regarding the statistical modeling, linear model and Gradient Boosting Machine (GBM) are applied for implementing AVMs. In addition, two types of loss functions, the least squares loss function and the least absolute deviation loss function, are combined with three strategies for missing values, complete cases strategy, multiple imputation strategy, and missing value node strategy. Thus, in this thesis, the predictive performance of GBM will be comprehensively investigated using six combinations of loss functions and missing value strategies, comparing with the benchmark, linear model. Three research questions will be answered in this chapter: whether GBM is a competent model that shows the same results as the previous studies; which missing value strategy is recommended and whether loss functions are worthwhile to be discussed in AVMs. If GBM could provide more accurate price predictions as expected, this could benefit the other fields, such as indexing. This gap will be filled in Chapter 6.

Chapter 6 continue the empirical investigation by applying the recommended *AVM* implementation in Chapter 5 for price indexing purpose. Three research questions will be investigated in this chapter. The first is to study whether it is achievable to compile *RPPI* using *GBM* using the hedonic regression method, especially using the time dummy approach. Using the time dummy approach requires that *GBM* should be interpretable, while *GBM* is one of the machine learning models perceived as the "black box". The second is to compare different indexing implementations, which are compiled using different models and different indexing approaches. The indices are examined in two aspects, the accuracy of rough appraisal and the contemporaneous correlation. Thirdly, the revision of indices is discussed, especially for the indices using the time dummy approach. Normally, a price index representing a long-term market trend should be revised routinely, the revision is to correct errors in the index and to update the index values due new information becomes available. Meanwhile, the recommended residential property price index will be used to measure the changes and deviations of building prices in the local market through the investigation period in Chapter 7.

Chapter 7 empirically examines the optimal timing of development and vacant land valuation based on the real option theory. The growth and uncertainty of house prices are measured using the house price index compiled by the same method in Chapter 6. The foundation of this thesis is real option theory that includes all three components of land development, the input (vacant lands), the effort required (construction cost), and the output (residential dwellings). Three questions need to be answered: what factors will affect the optimal timing of development? Will these factors influence the land value? What effect do these factors have on the timing of development and land value? This thesis aims to empirically and comprehensively test the real option theory, such as the effect of construction cost. Overall, the real option theory will be discussed and empirically examined in this chapter.

Chapter 8 summarizes the empirical findings of this research, the research aims, and the research objectives. The strengths and limitations are also mentioned, along with potential directions for future studies.

Chapter 2

Literature review

2.1 Introduction

This chapter provides a review of the literature on three research topics, residential property price prediction, residential property price indices construction, and land price estimation and the optimal timing of land development. The research methods, empirical findings, and potential drawbacks are discussed for each topic. When machine learning techniques (or artificial intelligence) join the competition for the high accuracy of price predictions, the recent literature finds the better performance of various machine learning methods using transaction-level data sets in different regions on the earth. The drawbacks and data issues in transaction-level data and the ability to interpret models need more attention, especially when machine learning methods are applied.

The review starts with an overview of the literature about the price prediction in Section 2.2, and the literature about the price indices in Section 2.3 for residential property. The review for the land price estimation & land development is focused on the land valuation methods inspired by real options theory in Section 2.4. The discussion of prior real estate studies using transaction-level data reveals several overlooked issues in common. The first issue is missing values, which are normally ignored or rarely discussed and handled. Most importantly, another issue of how to interpret the well-performed machine learning model has not been considered by the most recent studies. Thus, these two gaps in real estate research need to be filled.

In summary, the aforementioned gaps addressed in the literature will be discussed in Section 2.5. This thesis intends to discuss the solutions for the missing value issue in the residential property valuation field. Meanwhile, the other applications of machine learning techniques in residential real estate, such as the compilation of residential property price index and vacant land valuation, are also studied, including interpreting the machine learning implementations.

2.2 Valuation for residential property

The valuation model for residential property commonly applies the hedonic pricing theory Rosen (1974), it claims that the value of the item being researched could be decomposed into the constituent characteristics and observed market value or obtained estimates of the contributory value for each characteristic. The theory does not place many restrictions on the price function. An individual residential property could be the item being researched. In the recent literature, the automated valuation model (AVM) is commonly applied following the idea of hedonic pricing theory, which uses one or more mathematical techniques to provide an estimate of the price of a particular property at a specified time without or with less human intervention post-initiation (RICS 2021). They are well applied in the real estate field for diverse purposes, such as mass appraisal, mortgage portfolio management, taxation, cost/benefit analyses for potential public expenditure, and so forth, by different stakeholders, such as valuers, mortgage lenders, regulators, and city planners.

In the *AVM* literature, there are many statistical models fitted to market prices of properties. Some characteristics could distinguish these models, such as flexibility, number of "parameters", and dependency on the amount of data. Imagine a highly structured linear model versus a fully nonparametric model, the latter is fully flexible but needs a large set of data. Meanwhile, The linear model might be too restrictive. The optimal model might be in the middle: semi-parametric. Anglin and Gencay (1996), Parmeter et al. (2007), and Haupt et al. (2010) examine the performance of semi-parametric, non-parametric and fully parametric models respectively on the same data set. They found different "best" models. Anglin and Gencay (1996) find that the semi-parametric model outperforms the parametric model significantly. Then, Parmeter et al. (2007) claims that the non-parametric model is better when the categorical variables are added to the estimation. On the opposite, the opinion of Haupt et al. (2010) is that a previously proposed parametric specification cannot be rejected. In real estate data, spatial information is usually available and hard to measure. Bourassa et al. (2010), Liao and Wang (2012), and Liu (2012) consider the spatial effect on property prices via some complicated parametric models, and examine the potential improvements in *AVM*s. With the development of machine learning and artificial intelligence techniques, they become popular to estimate transaction prices of properties. Kagie and Van Wezel (2007), Kok et al. (2017), Mayer et al. (2019), and Schulz and Wersing (2021) apply a batch of machine learning models, such as Random Forest, Gradient Boosting Machine, Neural Network and so forth. They find that the machine learning models benefit the accuracy of price predictions. Kagie and Van Wezel (2007) apply the boosting trees to estimate property prices in six cities of the Netherlands. They handle the missing value issue by replacing them with the mean of the non-missing values of the variable, and then, compare the model performance with a linear model, the benchmark. Kok et al. (2017) compare the performance of several machine learning approaches on the full data with only complete cases, which means they remove around 30% of full observations that have missing values. A huge data loss may limit their results. Mayer et al. (2019) and Schulz and Wersing (2021) study the performance of more machine learning approaches in a dynamic setting (rolling windows). This mimics the updating process and maintenance of AVMs in industrial practice. A more recent paper (Sing et al. 2022) develops an artificial intelligence-based automated valuation model (AI-AVM) using the boosting tree ensemble technique to predict housing prices in the private and public housing sectors of Singapore. They also found that the boosting model is the best predictive model that produces the most robust and accurate predictions for housing prices compared to the decision tree and other regression models.

In the property valuation industry, AVMs are also applied with leading machine learning and artificial intelligence techniques in some big valuation service providers, for instance, Zillow, CoreLogic, Zoopla, and REA Group. Zillow uses a sophisticated neural network-based model that incorporates millions of data to calculate a “Zestimate” of a residential property, the nationwide median error rate for the “Zestimate” is lower than 7.5%. Besides accurate estimates of property prices, machine learning techniques allow AVMs to have high data tolerance. It allows different data types in the estimation, thus, it will be easier to do estimation with categorical variables and spatial information. Moreover, the quality of collected data may not always be guaranteed, the practice problems of missing values are common because of, for instance, human error. This issue becomes worse when the collected data are usually multi-sourced and multi-formatted. Some machine learning techniques could solve missing values when estimating. In advance, these companies greedily expect that *AVM*s could satisfy the demands of their commercial clients, such as the confidence of the predicted prices for mortgage lenders (commercial banks). In practice, loss functions merit more attention as diverse types of loss functions could satisfy different purposes of valuations. The default (the squared error loss function) could not match the developing demands in the industry. More importantly, *AVM* users hope to benefit from the full potential that technology promises for valuations, potentially making them quicker, less prone to human error, and sensitive to a far wider range of evidence than is currently the case (RICS 2017). This is worth an investigation in the future.

2.3 Residential property price indices

The price of residential property needs to be accurately estimated for *AVM* users, who may be potential property buyers. Also, monitoring the development of residential property prices is considered important, especially in times of economic turbulence Eurostat (2013). For instance, central banks use aggregated price trends as a leading macroeconomic indicator of economic growth. Rising house prices are often associated with periods of economic growth while falling house prices often correspond with a recession of the economy. However, the measurement for residential property price development varies per country, and even within a country, there are sometimes two or more competing methods in use. For example, in Australia, the Australian Bureau of Statistics uses the stratification method to measure the house price dynamics. Meanwhile, the hedonic regression method and repeat sale method are applied in the real estate industry, such as REA Group and CoreLogic AU. Generally, an index is built to document the price dynamics. In the literature on residential property price index (*RPPI*), it is intended to measure pure price changes, real estate prices need to be adjusted for quality change. Individual residential properties have their unique qualities. It is necessary to somehow control for any variations in the amounts of the price-determining characteristics of the properties to compile a constant quality *RPPI* or quality-adjusted *RPPI*. Four main methods have been suggested in the literature to control for changes in the amounts of property characteristics: stratification or mix-adjustment methods, repeat sales methods, hedonic regression methods, and appraisal-based methods.

The mix-adjustment method is a straightforward and computationally simple algorithm to adjust for changes in the quality mix of samples of residential properties in different periods. It is to divide or stratify residential property transactions according to some price-determining characteristics. For example, Wood (2005) groups house price observations into sets or "cells" of observations on houses with similar locations and physical attributes. The groups or the stratification of residential properties are homogeneous. The average or the median selling price within each group or stratification can be used as a constant quality price for that type of property, this price is a proxy. A batch of these prices in each stratification is aggregated up with weights to give a "mix adjusted" price or an overall index by regular index number theory. Such a mix adjustment method is also known as the stratification method. Prasad and Richards (2008) propose a novel stratification method and test the method using the Australian data. They grouped suburbs according to the long-term average price level of dwellings in those regions, rather than just clustering smaller statistical areas into a larger one. The residential property price indexes for eight capital cities published by the Australian Bureau of Statistics (ABS) is a practical example of such a method. For this method, the composition of the residential property sample is quite important. A change in the composition of the residential property sample will alter the number of observations in each stratification. However, if the stratification is defined sufficiently precisely and reasonably, so that all elements of the stratification have similar prices and price trends, then such changes will not systematically affect the mix-adjusted price.

The repeat sales method addresses the quality mix problem by comparing properties that have sold more than once over the sample period. This method only requires that the transaction prices and the identities of the transactions and the residential properties are available. The concept of this method is first introduced by (Bailey et al. 1963). Restricting the comparison to properties that have sold repeatedly ensures that the price relatives compare like with like, assuming that the quality of the properties remains unchanged. The standard repeat sales method is based on a regression model using the repeat sales data only, those repeat sales in all periods are pooled. An advantage of repeat sales methods is that they only require information about the sales, such as the identity of the sold property, transaction price, and date. The rest factors affecting property prices are held constant during the period of two consecutive sales. Additionally, standard repeat sales regressions are easy to run, and the price indices are easy to construct. The Standard & Poor's CoreLogic Case-Shiller Home Price Indices are the leading measures of residential real estate prices for the United States, as the best-known example that applies the repeat sales method.

However, there are several potential flaws in the repeat sales method. One drawback could be that heteroskedasticity is found when the time interval between two consecutive sales is not identical. Case and Shiller (1987, 1989) argued that changes in prices include components whose variance increases with the interval of sales so that the assumption of a constant variance of the errors for ordinary least squares (OLS) is violated. They proposed a weighted least squares approach to correct this issue. The weights are derived from the regression that the squared OLS residuals are regressed on an intercept and the time interval between sales. However, some studies found that this adjustment may not be necessary, such as Leishman and Watkins (2002) using Scottish data and Jansen et al. (2008) using Dutch data. These studies concluded that the standard repeat sales method using OLS was not inferior. Also, some methods are similar to the repeat sales method, such as Deng et al. (2012). They used a matching procedure to construct samples of private residential sales in Singapore. An advantage of the matching procedure is that it is easier to characterize changes in the full distribution of quality-adjusted sales prices, rather than just the means.

Another potential drawback of the repeat sales method is the issue of "revisions": when new periods are added to the sample, and the model is re-estimated, the previously estimated price indices will change. The issue comes from the fact that "new" usable transactions have their previous transaction in a period for which an index has already been estimated. The new transactions are provided, therefore it is unlikely that the estimates for the previous periods remain the same with the additional information. Clapp and Giaccotto (1999) and Clapham et al. (2006) are motivated to discuss this issue in their empirical studies. Sample selection bias happens in the repeat sales method which could be another drawback. This method uses data that only includes the properties that have been sold at least twice during the investigation period. If the investigation period is short, the properties that are sold frequently may differ from those that are not, then, there might be reasons for being hot to hand around quickly. When the investigation period is extended and the coefficients are re-estimated, the bias might decrease as the number of observed repeat sales increases.

Over time, a dwelling could undergo renovations or be restructured. To address this issue of housing feature changes, some improvements are developed. For example, some intend to edit the repeat-sale model (e.g. Shiller 1993, Clapp and Giaccotto 1998, Goetzmann and Spiegel 1997), the others intend to advocate hybrid models, that exploit all sales data by combining repeat sales and hedonic regressions and address not only the quality change problem but also sample selection bias and inefficiency problems, (such as Case and Quigley 1991, Quigley 1995, Hill et al. 1997).

The hedonic regression method is an index method that utilizes information on the relevant property characteristics to estimate the quality-adjusted price index using regression techniques, inspired by the hedonic theory (Rosen 1974). There are two main different ways to estimate hedonic price indices. One is the time dummy approach has been prominent in the real estate literature. This approach models the price of a property (or the logarithm of the prices) as a function of its physical characteristics and a set of time dummy variables. As the time dummies could measure the temporal effect quarterly or yearly, the data for all periods are pooled. When the index number of the first period is set as the reference period, the rest index numbers will be the coefficients of time dummies (or exponential of the coefficients if using the logarithm of dependent variable) (for example, Crone and Voith 1992, Gatzlaff and Ling 1994, De Haan 2004, Diewert et al. 2007, Agarwal et al. 2021). The time dummy method is easy to apply and interpret but makes strong assumptions: implicit prices (beta coefficients) of the housing characteristics are constant over time. The process of index construction is easy, as only the time-dummy coefficients vary over time.

A problem with the time dummy method is the issue of the revision like the repeat sales method. If the new sample data are added, the coefficients of the characteristics may slightly change, including the coefficients of time dummies. Consequently, the time dummy method violates time fixity (Hill 2004), the previously computed price index numbers have to be updated by those newly computed. Shimizu, Nishimura and Watanabe (2010) overcame this problem of the time dummy method by using moving windows. A hedonic model with time dummies is estimated in each window. Then, by chaining the last period-to-period (such as quarter-to-quarter) changes in each window, a non-revised time series is obtained. The index numbers could be calculated when the reference period is determined.

The second main approach is to estimate separate regressions in the cross-sections of each period and to assess the price index for a property or a set of properties with given housing features, such as all properties that were sold in one quarter. In this case, the index value must be computed as an average of all predictions in one cross-section. This approach allows the implicit prices of characteristics to vary over time. Thus, it is more flexible than the time dummy variable approach and does not need to concern revision issues.

There are two distinguished variants: the *characteristics prices approach* and the *hedonic imputations approach*. The characteristics price approach is to predict prices in a reference period and another period for a "standardized" property with fixed quantities of characteristics. The computed price relatives of the "standardized" property would be the index numbers when the index number is set to 1 in the reference period. If The different "standardized" property is determined, the different index numbers will be given. Thus, the "standardized" property needs to be the most representative, however, the characteristics of the most representative property may not be constant in the long term. The terminology of this approach differs between authors. For example, Crone and Voith (1992) and Knight et al. (1995) refer to this approach as the "hedonic method" directly, while Gatzlaff and Ling (1994) call it the "strictly cross-sectional method".

The imputations approach is to impute the "missing prices" for the properties in all periods. The price relative between the reference period and the other periods will be the index number when the reference index number is set to 1. The imputation approach could be traced back to Hill and Melser (2008) or before in the housing field, and well discussed more generally, such as Triplett (2006). Based on the usage of imputed prices, there are two varieties, single imputation index, and double imputation index. Also, there are two types of index calculations, Laspeyres type (arithmetic and geometric) and Paasche type (arithmetic and geometric) respectively. In addition, Törnqvist type or (Fisher type) is the geometric mean of the Laspeyres type index and Paasche type index. The imputation method is conceptually more complicated than the characteristics prices approach (Hill 2013b).

Assessment-based method Many countries tax real estate property and are likely to have an official property valuation office that provides periodic appraisals of all taxable real estate properties. Assessment-based methods combine selling prices with appraisals to compute price relatives (sale price appraisal ratios) and control for quality mix changes. The Sale Price Appraisal Ratio (SPAR) method is based on the matched model methodology. In contrast to the repeat sales method, it relies on all (single and repeat) sales data, and there is no revision of previously estimated indices. Of course, the method can only be applied in countries where reliable assessed values of the properties are available.

Overall, diverse index methods have their advantages and disadvantages. The different methods are solutions to a problem: what data is observed? If there is no housing feature observed, the repeat sale method could be the only choice, even if there is a data loss. If the housing features are observed, the quality-controlled repeat sale method and hedonic regression method could be equally used. The selection of indexing methods may highly rely on the data observed.

2.4 Land valuation and development

Land development is a decision-making problem like Hamlet's famous soliloquy from Shakespeare's tragedy, "To be or not to be; that is the question". Landowners have the right to develop their land if they wish, but they are not obliged to do so[1]. Land development means that they construct a building on their land, thereby converting the land into a building that can be rented out or sold. However, the future is uncertain. It should be explained whether the development is rational. A real option could theoretically explain this question. The option value can be assessed under certainty (see examples in Donald 1970, Arnott and Lewis 1979, Capozza and Helsley 1989). With uncertainty, the option value increases because the risk premium arises (see examples in Titman 1985, Williams 1991).

Land development could be modeled as a call option, that gives landowners or landowners a development opportunity for their lands. It is a right, not an obligation. When a project of land development is proposed by the landowner, a decision could be made at any time, whether a building should be constructed on the vacant land or it should be preserved. For a simple example, the present house price is P, the current construction cost is C, and the land value is L. If $L > P - C$, the landowners should wait. Otherwise, there is a loss to develop the vacant lands. If $L \leq P - C$, then the vacant lands should be developed immediately[2].

The theory of real options is based on the theory of financial options, but some assumptions that seem to be appropriate in the context of financial assets seem less palatable in the context of real options. A real option is an economically valuable right to let the option holder make a decision which often concerns business projects or investment opportunities. This option is

[1] Obviously, legal requirements, such as building permission, zoning regulations, and the building code, place restrictions on the development process, but do not affect the principle right of development

[2] The detailed derivation of real option theory is explicated in Section 7.2

referred to as "real" because the involved asset of the business project or investment opportunity is typically tangible, such as lands and buildings, instead of a financial instrument. Pindyck (1991) and Dixit and Pindyck (1994) give introductions of real option into the literature. Some basic models of irreversible investment are reviewed to illustrate the option-like characteristics of investment opportunities. Meanwhile, the role of risks and opportunity costs are discussed in the irreversible investment. Land development is one kind of irreversible investment. The construction cost is the investment cost, the house price is the investment output. Meanwhile, the development process is irreversible in a relatively short term. Capozza and Helsley (1989) theoretically discuss the price of land under the case of certainty when the landowners have perfect foresight, which means the future is assumed as determined. Thus, the potential risks are not mentioned. Titman (1985) study the case of uncertainty for the price of land and explain the land valuing (as option pricing) using a binomial tree model. The analysis conducted in the paper demonstrates that the diversity of potential building sizes provides an option to the landowner. The option becomes more valuable when uncertainty about future prices increases, and this leads to a decrease in the chance of land development in the current. This relationship may have important macro-implications. If the government initiates a monetary policy to stimulate the development of vacant land, the policy may actually lead to a decrease in the development if there is uncertainty about its duration or its effect. McDonald and Siegel (1986) improve the discrete-time model to a continuous-time model using the Wiener process (also known as one-dimensional Brownian Motion). The values of investment output and investment cost are assumed as two continuous stochastic processes. The decision is determined by the dynamics of the gap between the values of investment output and investment cost. Applying this model, Williams (1991) computes the optimal exercise policies of the real option, such as optimal prices of the developed and undeveloped properties, analytically and numerically for the option to develop or to abandon in real estate. Capozza and Li (1994) discuss vacant land conversion, which could be treated as a capital-investment decision involving the choice of both the timing of investment and the intensity or level of capital investment. They derive the conditions for this optimization. When the conditions are matched, the conversion (or the capital investment) should take place. They then conduct comparative statistics and analyze the case when real estate is taxed. Additionally, under almost the same settings, Capozza and Li (2002) models the development decision when net rents or cash flows are evolving randomly and geometrically. Moreover, they derive the conditions under which positive responses of investment to interest rate increases can occur. Overall, the real option theory gives theoretical guidance on when to wait and why waiting can lead to better decisions. The theoretical applications discussed so far motivate us to examine the qualitative and quantitative effects of the variables, that could affect the optimization of development in empirical analysis.

Empirical applications start by assuming that the processes of the underlying value drivers are stochastic or deterministic and solve for the option value using these drivers. The equations derived under theoretical models are used and are calibrated to empirical land prices, building prices, rents, etc. Quigg (1993) is the first to examine the empirical predictive power of a real option model using the transactions of 3,200 developed properties and 2,700 undeveloped land parcels in Seattle from the second half of 1976 to the end of 1979. The real option model used allows stochastic processes for property prices and development costs, which correspond to the model explicated in McDonald and Siegel (1986) and Williams (1991). The transactions of developed properties are used to establish hedonic regressions in each year and zoning category. The fitted regressions are for predicting market prices of buildings that could be constructed on undeveloped lands. Then, the real option value and the intrinsic value are computed for each undeveloped parcel, where zero timing flexibility is assumed when calculating intrinsic value. The average option premium is estimated at around 6% and ranges from 1% to 30% varying by building type. Quigg (1993) also run a regression of land price on the intrinsic value and option premium with a constant. The regression shows that the constant is not far from zero, and the coefficient for intrinsic value is close to 1. Also, the coefficient for the option premium is positive and statistically significant. Thus, Quigg (1993) concludes that the option valuation model has some explanatory power for prices.

Holland et al. (2000) examines if the supply (or new investments) of commercial real estate in the US is described better by real options or 'neoclassical' investment theory. New investment is measured by the square feet of new construction starts. Effectively, they examine the role of idiosyncratic risk and systematic risk on investment. Meanwhile, other factors are explicitly controlled, such as interest rate, construction cost, and expected growth rate of asset cash flows. They find that the change in total uncertainty exerts a negative effect on investment. Chu and Sing (2021) examines the effects from the demand side. They allowed the demand shock and the cost functions to be dependent on the intensity of real estate development. It was found that demand uncertainty delays development activities. Somerville (2001) considers land development as a compound option, which could be divided into multistage, permit-to-start, and start-to-completion. Regressions are conducted using Canadian data. The house price volatility is the most important variable in the regressions, it shows that anticipated future volatility should delay the starts. As Somerville (2001) discusses, the $GARCH$ measure of house price volatility might be riddled by measurement error issues, instead, a vector autoregression (VAR) might be a more appropriate modeling approach. Besides the house price volatility, the interest rate has a macro-implication on land development. Capozza and Li (2001) investigate the effect of interest rates on new investment. They find that interest rates can have a positive influence on building permits if the growth rate and/or the volatility of the growth rate is high, following the theoretical analysis of in Capozza and Li (2002).

Cunningham (2006) tests whether greater price uncertainty should delay the timing of development and raise the land prices, the dataset used includes real property transactions with parcel characteristics for the Seattle area. Firstly, the constant-quality house price index is constructed using house transaction data. Then, A seasonal autoregressive model is fitted to each of the price series and used later for house price forecasts. These time series are used to estimate the time-varying house price volatility. The proportional hazard model is estimated to test whether the house price volatility affects the timing of development; the linear model is estimated to investigate the effects of house price volatility on the value of vacant land. The results show that an increase of one standard deviation reduces the development likelihood by around 11% and raises vacant land prices by 1.6%. After conducting several robustness checks, Cunningham (2006) concludes that landowners consider their real options, even in competitive and economically important sectors, like new home construction. A similar study was conducted by Bulan et al. (2009) using condominium data in Vancouver from 1979 to 1998. Bulan et al. (2009) also find that an increase in risks leads landowners to delay new constructions. An increase of one standard deviation in volatility leads to a 9% decline in land prices, equivalent to a 13% decline in the probability of development. More recently, Fan et al. (2022) empirically tests the volatility effects on land development options using Singapore's government land sales data. Similar results were found that development land option premiums increase by 5% on average with one standard deviation increase in conditional volatility. However, the high volatility effects on development options are negative in an extremely high volatility state, and risk-averse developers exercise development options earlier than that predicted by the risk-neutral model. This reveals that the real estate developers' behavior changes in different states of market volatility. Moreover, Bulan et al. (2009) involves the effect of competition into consideration for the development timing. The degree of competition could affect development timing through the interaction with the volatility of return. More similar studies are investigated recently, such as Wang et al. (2016) and Zhang et al. (2021).

2.5 Research gaps in literature

Transaction-level data is necessary when conducting an analysis or investigation on the previously introduced three topics, automated valuation model (*AVM*), residential property price indexing (*RPPI*), and land valuation and development.

From the review of *AVM* literature, it is known that machine learning techniques do improve the accuracy of price predictions for residential properties, meanwhile, there are diverse choices of machine learning techniques(Mayer et al. 2019, Schulz and Wersing 2021, for example). However, there is a hidden research gap, which is missing value issues. The impact of this issue could be significant that the *AVM* cannot be implemented when the missing rate is high. (Krause and Lipscomb 2016) discuss the data issues, including missing value issues, in the real estate field. They find that the missing values are common in the real estate data due to diverse reasons. While it might be possible to correct individual observations manually for missing values, this becomes extremely time-consuming and impossible once a data set is large. Thus, it is necessary to find a solid solution for missing values in the real estate data, especially transaction-level data, when conducting *AVM* using machine learning techniques.

While machine learning techniques are applied for the valuation of residential properties using transaction-level data, the extension of valuation is worthwhile to discuss. The residential property price index is highly related. Most classic indexing approaches rely on the "workhorse" model, the linear model. The machine learning applications in the *RPPI* field are relatively less discussed. It is already known that machine learning techniques could benefit the accuracy of predictions. However, it is unsure whether those accurate price predictions would benefit the residential property price index. Meanwhile, the interpretation of machine learning techniques is necessary when combined with the time dummy approach of indexing. It is necessary to investigate how well the machine learning techniques could explain the time effects.

In academic research, transaction-level data are also used to establish an analysis for land development, including land valuation and timing of development. The theory suggests that a land development project is a real option, the value is affected by the "investment" output and the "investment" cost. If the land is developed immediately, the intrinsic value of the option is the difference between the current sale price of the building and the total construction cost spent (Quigg 1993, eq. 7). In the literature, the effect of interest rate is discussed, as it discounts the future value of the investment and affects the investment cost. On the "investment" output side, the effect of house price volatility is well studied in several previous literature (Wang et al. 2016, Zhang et al. 2021, see examples). However, it is also worthwhile to discuss the effect from the cost side. Especially, the *ABS* suggests the construction cost may vary during the building process. The building process usually takes around one or two years. When building a new home, the final cost of construction may differ from the initial expectations at the building approval stage and/or the start of construction Australian Bureau of Statistics (2020a). Thus, including the construct cost and its volatility in the land development analysis could comprehensively study the determinants of the vacant land value and optimal timing of development.

2.6 Concluding remarks

The empirical studies for the analysis using transaction-level data on three different aspects have been reviewed in this chapter. Motivated by the development of machine learning, three studies are formed for finding some potential solutions for the current issues and combining machine learning with real estate practice, based on the literature review of *AVM*, *RPPI*, and land valuation and development.

Firstly, the *AVM* literature suggests that machine learning techniques could improve the accuracy of price prediction surprisingly using the available housing characteristics. The predictive accuracy is ranked high compared with some non-parametric and classic parametric models, if not the highest, among the highest tier. Thus, machine learning techniques could benefit *AVM*s when they are established and deployed. However, missing values are a usual data issue in residential real estate. They do obstruct to conduct *AVM*s. Thus, the first topic focuses on finding an answer to "What to do with observations that suffer from missing values when estimating prices of residential properties using an automated valuation model?".

For the second, several commonly applied indexing methods are reviewed. In different countries or areas, the preferred indexing approach is various. The findings from the literature show that the selection of indexing approaches highly relies on the type of data observed. With transaction-level data, there are several unsure questions, "Is the automated valuation model with machine learning techniques always better than the linear model (the workhorse mentioned above) when compiling the price index?", "Whether more accurate price predictions lead to a better price index?", and "How well do the machine learning techniques explain the time effect on valuation?", in the indexing field for residential properties.

The last topic analyses land development which also needs transaction-level data support. The analysis for land development could investigate the factors that affect the optimal timing of development and contribute to land pricing. The previous empirical studies find the effects from the investment output side, such as house price volatility, and some effects from “policies”, such as interest rate and money policy. However, the investment cost side is also valuable, and there are uncertainties during the building process. Thus, it is worthwhile to study “What are the effects of the factors, that are from the investment cost side, on the valuation of vacant land and optimal timing of development?”. These gaps of three topics summarized and identified from the literature review will be studied in this thesis.

Chapter 3

Data

3.1 Introduction

As described in Chapter 1, transaction-level data are used to conduct quantitative analysis in the real estate field. These data are often collected by surveyors or government agencies, such as *Landgate* in Perth, Western Australia, and some of these data are then assembled by data aggregators, for instance *CoreLogic*. These data usually contain the observations that record the details about the transactions and the transacted properties, as well as complementary spatial data such as coordinates and other points of interest. However, the coverage and the quality of this information highly rely on the data source and the method of data collection.

In this chapter, the discussion of the data begins with a basic summary of the local housing market and the data sources. Then, the thorough introduction of data is described in three aspects, data identification, data coverage and quality, and data preparation. Firstly, in Section 3.1.1, it briefly introduces the basic information about the local housing market. Section 3.2 answers the question, "What is transaction-level data?". It also summarizes the sources of all used data sets, and different sourced data sets could cooperate for specific purposes. Then, in Section 3.3, the data identification is studied, which helps to distinguish the individual transactions and diverse types of properties. The data coverage shows the availability of variables that could be included in the price estimation. Additionally, the data quality is discussed, and the missing value issue is detected in the data set. Section 3.4 describes the data preparation for empirical analysis. The data preparation involves data selection, data washing, and data error treatments. Section 3.5 sums up and concludes the chapter.

3.1.1 Study area and local market

Western Australia is the largest state of Australia, occupying the western 33 percent of the land area of Australia excluding external territories with around 2.7 million inhabitants (June 2021, rank 4th in Australia). Around 80% of the population in the whole state lives in Perth, the capital city of Western Australia. Figure 3.1 depicts the map of local government areas (LGA) in the Perth metropolitan area.

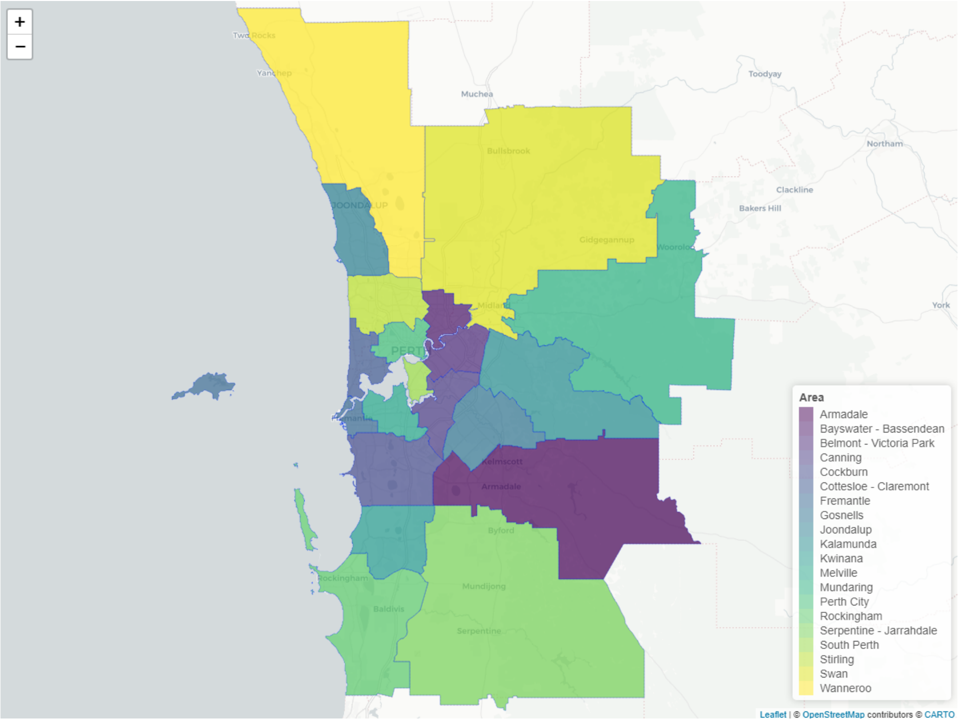

Figure 3.1: The map of *LGA*s in Perth, WA. The color scale is only to discriminate each LGA.

Perth as a study area is selected due to the following reasons. Firstly, a large volume of property transactions are processed in the Perth metropolitan area, around 50,000 properties are transacted each year. Meanwhile, it is allowed to view the details about the transactions and the transacted properties. The full history of a parcel, from vacant land to building(s), could be observed by chaining the sequence of transactions on the same land. The second reason is the density of the population. Due to the low density of population and low mobility in the regional area, some rural housing markets are dispersive and more independent of each other. The amount of local transactions is relatively small, and the available information on sold properties is relatively incomplete. The data maintenance is not as well organized as what has been done in Perth. Even, some transaction details documented in Perth are not complete either. As mentioned above, the majority of the population lives in the Perth metropolitan area. Thus, the local housing market could be more mature and more representative than the rural markets. Additionally, for a practical consideration, familiarity with the city may be helpful for the study. Meanwhile, it could be convenient to get support from the data provider. All of these make the Perth metropolitan area suitable as the study area.

The local housing market in Perth has experienced some ups and downs in the past 20 years. The first dramatically increasing happened in the first decade of the 21st century, which follows the global trend. The local housing market is super hot, property prices are increasing rapidly. Between 2010 and 2015, it was a price adjustment period. There are some price fluctuations, but in general, the price is relatively stable. After 2015, Perth's housing market has seen a decline in prices due to a decrease in demand and an oversupply of properties. The number of sales listings in 2015 increased by around 30%, compared to the number in 2014. And the median of selling days is over a month. During the Covid pandemic, the market has started to go up again with a surge in demand for properties. Currently, it presents a decline sign due to rapidly rising interest rates.

In general, Perth is known for having a more affordable housing market compared to other major Australian cities such as Sydney and Melbourne. According to the recent data published by the Australian Bureau of Statistics (ABS), the mean dwelling price in Perth is around AUD 600,000, which is much lower than the average, AUD 920,000 in the whole of Australia. In more popular and affluent suburbs, prices can be much higher, while in more affordable areas, prices can be cheaper. The rental market in Perth is also relatively affordable. The median rent for a house is around AUD 550 per week, while the median rent for a unit is around AUD 480 per week. There is an around 50% increase, compared with the rents in 2018. However, the rents are still cheaper than the rents in Sydney when they are over AUD 700. In the local housing market, it is offered a range of property options for buyers or investors. For residential property sales each week, 70% of the sales are houses, and 25% are unit sales. The rest is vacant land sales.

Additionally, 70% of Western Australian households owned their own home, including 43% owned with a mortgage and 27% owned without. The median mortgage outstanding in Western Australia was AUD 280,000 in the financial year 2019 – 2020, a decrease from AUD 324,000 in 2017 – 2018. This mortgage pressure on a household is acceptable when the median total personal income was AUD 55,000 in 2020. Overall, the Perth housing market is dynamic and constantly and healthily evolving.

3.2 Transaction-level data

Transaction-level data are commonly used in residential real estate analysis, they are often collected and aggregated by some real estate firms (e.g. *CoreLogic*) or some government agencies (e.g. *Landgate* in Western Australia). In these data, the observations contain the details including basic information about the transactions, such as the date and the price when the transaction is made, and the housing characteristics of transacted properties, such as the number of rooms and land size, as well as complementary spatial and location data such as census statistics for the land area.

The nature of transaction-level data is longitudinal and spatial. It does not like the panel data. Thus, all properties in the data are not consecutively observed in each period. The observations in one period only represent the sold properties, that are a part of the whole. For example, a property is rarely sold more than once in one year. Transaction-level data always come with spatial information. Handling spatial data properly requires the use of a Geographic Information System and/or advanced database systems. Additionally, the spatial information is usually available in a separate file, for example, cadastral files. This may require matching and combining the available information in the transaction file. These two natures make managing transaction-level data difficult. Besides nature, it is the biggest headache that transaction-level data are often dirty. It requires a big effort in data preparation. The common "dirt" concern is the issue of standardization, data errors, missing data, non-arms length transactions, mislabeled data, and outliers. Among them, the issue of missing data is relatively difficult to solve, and this issue is commonly found especially when the data set is large.

Transaction-level data are not flawless but are necessary for residential real estate analyses. Because detailed information about individual properties can not be obtained from other sources. Meanwhile, they are critical to understanding the property value and the investment potential. For accurate valuation, transaction-level data provides information about the physical characteristics of a property, such as its size, its facilities, and location, which are important factors in determining its value. Accurate valuation is essential. Broadly, for market analysis, transaction-level data is used to identify trends and situations in the local housing market. The dynamics of market prices could be more accurately measured and be more representative. Meanwhile, the price history for different types of properties generated from transaction-level data can provide some insights into the current market conditions. Further, for investment analysis, transaction-level data could be used to evaluate the potential returns and/or risks of the investment for a real estate activity. For example, landowners could use this information to determine whether it is beneficial to build a house on the vacant land at the current.

In transaction-level data, some clues about building activities during the “life cycle” of a parcel could be identified, such as construction, renovation, demolition, and rebuilding. Such that, the structures may not be constant after the dwellings were built. The changes in structures are normally observed in the changes in the housing characteristics between two consecutive transactions. For example, the house owner builds a swimming pool in the backyard after he bought the house. When the owner sells the house, the number of swimming pools changes from 0 to 1. However, some structure changes are not observed. For instance, the design changes, but the number of facilities does not change. Hence, it is assumed the structures of dwellings do not change if the housing characteristics don't change between two consecutive transactions of the properties.

3.2.1 Data source

There may be diverse sources that transaction-level data are available. In Australia, the Valuer's general office or the land information authority in each state holds transaction-level databases of their jurisdiction. All of these data are government-owned. Access to these transaction-level data could vary widely from completely free and open to very expensive and difficult to procure. In addition, there are some industrial real estate firms (such as *CoreLogic*, *Australian Property Monitors*, and *PropTrack*) that collect government-issued raw transaction-level data. After standardizing and formatting, they re-sell the transaction-level data for a fee depending on the amount of data purchased. Meanwhile, these firms also provide some free open-source data sets gathered in some national research infrastructures, such as the Australian Urban Research Infrastructure Network (*AURIN*). However, there is limited information released or published. If the data source is reliable, luckily, the details of properties are mostly complete and well-formatted.

This section mainly introduces the sources of the data sets, the information the data sets include, and the purposes for observing these data sets. The data sets used originate from the *Landgate* and the *Australian Urban Research Infrastructure Network* (*AURIN*).

The *landgate* data

The *Landgate* is the main source of the data sets, that are operated by the Western Australian Land Information Authority, formerly the Department of Land Information, the Department of Land Administration, and the Department of Lands and Surveys. It is the statutory authority responsible for property and land information, that maintains the official registers of land ownership and survey information in Western Australia. The data obtained from the *Landgate* have two data sets, one contains the details of property transactions, which are collected from the authorized documents for each property transaction. The other includes the geographical information of properties observed from surveying. These two data sets are the base of research topics in this thesis, which are to compare the prediction performance of *AVM* implementation, to compile residential property indices, and to analyze the prices of vacant lands and the optimal timing of development. Additionally, this data set was well-used in previous academic studies, such as Goh et al. (2012) and Leishman et al. (2013).

The *AURIN* data

The complementary data of transaction information are from the *AURIN*. *AURIN* facilitates the development, deployment, and long-term support of advanced data, analytical methods, simulation models, and visualization capability for the adoption of high-impact research within government and industry across Australia. It allows direct access to high-quality, spatially enabled, national-scale data assets covering economics, population, health, housing, transport, climate impacts, social policy, critical infrastructure, planning, and land use. These resources originated from the partners of *AURIN*, such as Australia Property Monitors (*APM*), a part of Domain Group. The *AURIN* data observed include some basic information about the local market. The data set has two time-series data, the minimum and maximum transaction prices in the Perth local housing market. They are used for achieving specific purposes, such as identifying arm-length transactions.

3.3 Data description

Data identification is the first stage in the data description. It usually determines what kind of observations should be estimated in the models for various research aims. It is followed by an overview of data coverage, and then, the quality of the data will be discussed.

3.3.1 Data identification

The initial consideration of data description is to identify the data that are necessary to conduct residential real estate analysis. Firstly, the property type is identified using the classification variable. Three groups of properties are divided, vacant lands, residential dwellings, and non-residential buildings. The dividing rules for the three groups of properties refer to the definitions of vacant lands, residential dwellings, and non-residential buildings, which are clarified in the Functional Classification of Building Structure 1999 (FCB, revision 2011) and Census Dictionary (2016)[1]. The summary of property classes and the three groups are shown in the Appendix.

Then, each transaction needs to be distinguished. Typically, they record different transaction prices, structural characteristics, and the spatial or geographical characteristics of the transacted properties. As this information originates from two *Landgate* data sets, the link is necessarily required. The polygon identity number (**PIN**) is the link between the transaction data and the geographic cadastral data. Besides the **PIN**s, the **address**, the **parcel ID**, the **land ID**, the **LGA** (local government area), the **transaction date** and the **application number of the transaction** are helpful to recognize duplicate records and to understand the sale history of one parcel. Thus, longitudinal data could be constructed, each cross-section observes different transactions of different properties, meanwhile a timeline goes across the cross-sections. Each property could be sold several times during the research period, which means the transactions of the property could be observed in different cross-sections. The identity variables are also necessary in the data preparation stage. The details of the preparation process are summarized in Section 3.4.

[1] These two official documents are available on the *Australian Bureau of Statistics* website.

3.3.2 Data coverage and quality

The *landgate* data

The *Landgate* data have details of transacted properties and cover all transactions in the Perth metropolitan area. They could be a fair representation of transaction activities including arm-length transactions and non-arm's length transactions. Some privacy information is not available, such as the details about buyers and sellers. The rest information is obtained based on the agreement and license between *Curtin University* and the *Landgate.* The *Landgate* data can provide most of the required information in two separate data sets. These two data sets are combined by matching the **PIN** of each observation. Then, the *landgate* data include the information of transacted properties, such as the transaction prices and the structural characteristics of properties (for instance the number of bedrooms or bathrooms and wall and roof material, etc.), and the geographic cadastral information, that includes the latitudes and longitudes as a measure of location.

A total of 1,641,657 transaction records were received from January 1901 to December 2020. These include all properties transacted in the Perth area. This data set also contains some renovated and digitized historical records from old paper reports. Most of these historical records are the transactions processed before 1988. A detailed description of the variables included in the *Landgate* data is given below:

- **Address:** This contains the street number, the street name, the street suffix, and the suburb name.

- **Parcel ID:** The identity number of the parcel. The number is unique.
- **Land ID:** The identity number of the land, is paired with parcel ID. The properties built on the shared ownership land may have the same land ID, but their parcel IDs are different. The number is unique.
- **PIN:** It is the polygon identity number, which is used to link the transaction details with the geographical information of the transacted property. Some PINs are not available for some reasons. The main one is that the land survey is not complete. The PIN will be updated when the survey report is finished and uploaded to the *Landgate*. The number is unique.
- **Land area:** The total area of the land.
- **LA DESC (LGA):** The local authority in which the address falls. It is also called a local government area (LGA) which is a slightly different geographical system from the geographical system used in the *Australian Bureau of Statistics* (*ABS*).
- **Land use and property use:** The land use code and the property code. The *R* means 'Residential', the *I* stands for 'Industrial', the *C* indicates 'Commercial', the *F* means 'Farming', the *M* indicates 'Miscellaneous', and the *V* is 'Vacant'.
- **Zoning:** The local authority zoning code.
- **Property class:** The classification of the property.
- **Buyers and sellers:** The name of buyers and sellers, but for privacy purposes, this information is not available. Thus, the arm-length transactions can not be recognized by the buyers and sellers.

- **Sale price:** The transaction price of the sold property.

- **Date:** The transaction date.

- **Multiple sale:** The variable has the value '1' which means the sale includes more than one lot. '0' mean the others.

- **Application number:** The sale transfer document number for the sale.

- **Wall:** The wall type description for the property.

- **Roof:** The description of the roof type.

- **House area:** The area of the house. Generally, it represents the living Area. This would exclude areas such as the garage or carport under the main roof (UMR), balcony, storeroom if the only access is from outside the house, alfresco dining under the main roof unless it has substantial walls, etc.

- **Year build:** The year the residence was built.

- **Beds:** The number of bedrooms.

- **En suite:** The number of en suite bathrooms.

- **Baths:** The number of bathrooms.

- **Dining:** The number of dining rooms.

- **Carports and garages:** The total number of attached carports or garages to the residence, detached carports or garages from the residence, and carports or garages under the main roof. This variable is calculated from six variables, **CARPR**, **CARPRT DET**, **CARPRT UMR**, **GARAGE ATT**, **GARAGE DET**, and **GARAGE UMR** separately.

- **Kitchen:** The number of kitchens.
- **Family:** The number of family rooms.
- **Game:** The number of games rooms.
- **Meals:** The number of meals rooms.
- **Lounge:** The number of lounge rooms.
- **Study:** The number of study rooms.
- **Number of units:** This usually shows the count of the number of residences where “larger than one” exists.
- **Tencrt:** The number of tennis courts.
- **Pool:** The description of the pool type. The “B/G” means the pool is below ground, and the “A/G” indicates the pool is above the ground.
- **Coordinates:** The coordinates include the accurate longitudes and latitudes of the property. If the **PIN** is not available for the property, the coordinates are set to “NA”.

The details of transactions stored in the *Landgate* data are mainly from the surveyors’ or solicitors’ reports or other sources. Most records have good quality, the information about properties and transactions is complete, and no erroneous entries. However, some records are suffering from missing values and erroneous zeros. For example, there are blanks in some categorical variables and *NA*s in the coordinates, probably due to late updating issues. In Addition, some erroneous zeros are in some structural characteristics that can not be zero as long as the property exists, such as land area, floor size, and the number of bedrooms or bathrooms. These zeroes could mean the values are unknown rather than the non-presence of the characteristics. These issues are the main shortcomings of the *Landgate* data. They are also quite common in transaction-level real estate data sets (Krause and Lipscomb 2016). Compared with the transaction-level data used in other research[2], the overall quality of the *Landgate* data is competent. The missing rates of housing characteristics are mild, the highest rate is around 35% in the **Floor area**. Except for missing values, The issue of erroneous zeros is amply settled in Section 3.4.

The *AURIN* data

The *AURIN* data are used as the complement for the *Landgate* data. The data set includes a time series for the highest and lowest sale prices of properties in the Perth housing market. The time series are built for houses and units[3] in each *SA4* statistical area. The description of the statistical area geographical system and the relationship between *SA4* statistical areas and *LGA*s are outlined in the appendix. The *SA4*s are defined as the **southeast**, the **southwest**, the **northeast**, the **northwest** and the **central** by the *West-*

[2] Kagie and Van Wezel (2007) use the Dutch housing data that have the same data issues also. The missing rates of features are from 0.2% to 70%. In the Sydney (Australia) housing data set, about 47% of full data have one or more missing values in the housing characteristics (Hill and Scholz 2018, Web appendix). Graz (Austria) residential transaction data contain 11,250 incomplete cases, around 40% of total cases (Steurer et al. 2021). In Zillow Prize data available in Kaggle, around 0.38% of observations don't have information about the number of bathrooms and bedrooms. 9.25% of lot size is unknown and 72.82% of air conditioner details are missing.

[3] The houses and units are distinguished by the property classification. Such as apartments, home units, and flats belong to the group of units.

ern Australia's real estate institute (*REIWA*) and the *Australian Bureau of Statistics* (*ABS*). The *AURIN* data set contains the maximum and minimum transaction prices for the period from 1994 to 2021. Each data point represents an aggregation of recent twelve months' worth of data and is reported monthly. Such that, the report each December describes the lowest price and the highest price in the whole year. These data cover all the $SA4$ areas in Perth. A detailed description of the variables included in the *AURIN* data is given below:

- **Year and month:** This contains the time of the data points. That indicates the period of data aggregation.
- **Property categorization:** The category of the sold properties. This data set just has two categories, "house" and "unit".
- $SA4$ **code and** $SA4$ **name:** The codes and name of each $SA4$ area in Perth.
- **Maximum sold price:** The highest sold price in the local $SA4$ submarket spanning the recent year.
- **Minimum sold price:** The lowest sold price in the local $SA4$ submarket spanning the recent year.

Figure 3.2 depicts the minimum and maximum prices for houses and units in the **central** area of Perth metropolitan from 1994 to 2020. The graphs for the other four *SA4*s are shown in the Appendix. The difference in minimum prices between houses and units is negligible, both of them have an upward trend. For the maximum prices, the climbing trend is still clear. The gap between houses and units is much larger, which means houses could be much more expensive than units.

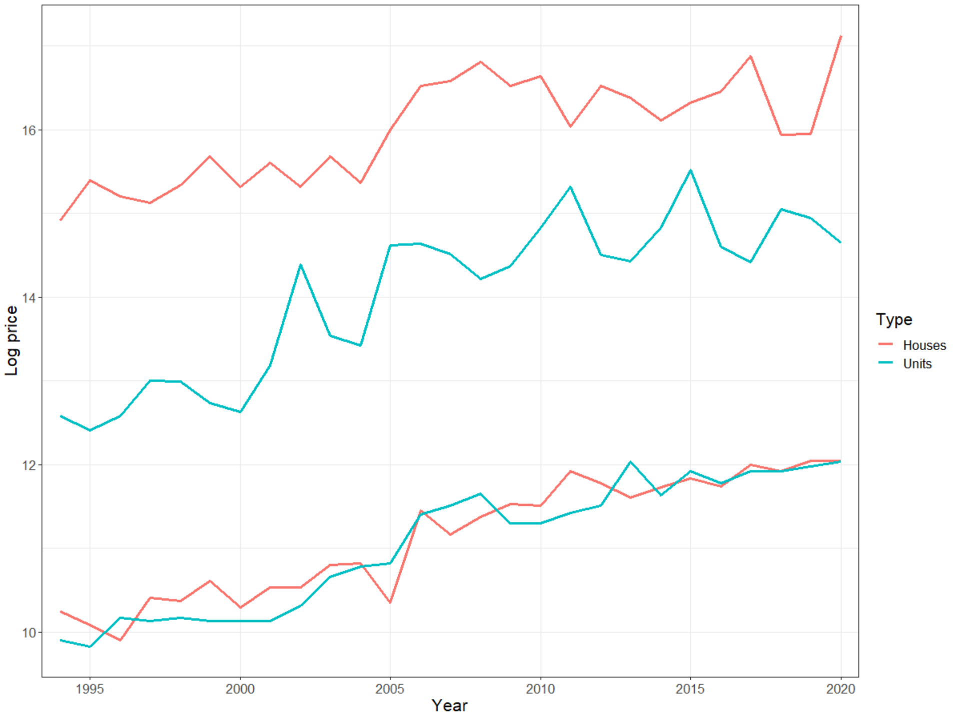

Figure 3.2: The minium and maximum of transaction prices in the central area (*SA4*) from 1994 to 2020. To depict the minimums and maximums in one graph, the transaction prices are under a logarithm scale.

This data set could support the *Landgate* data for identifying the market transactions. In the *Landgate* data, there are some non-market (non-arms length) transactions whose prices are much lower than usual. These could be the transactions between family members because of divorce or heritage. Usually, they could be identified by their stakeholders. However, due to privacy, buyers and sellers are not visible in the data set. Alternatively, the minimum and maximum transaction prices in each *SA4* for different types of properties are used as the lower and upper boundaries. In each $SA4$ area, if the transaction prices fall out of the min-max intervals, the transactions are recognized as non-market (non-arms length) transactions. The data quality is well maintained by *AURIN* and its partners. There is no missing value in the time series.

3.4 Data preparation

The data are necessarily prepared before modeling, irrelevant observations should not be included. This section describes the process of data preparation, which involves three different stages, the data washing stage, the data errors treating stage, and the variable transformation stage. The final data set, however, is not summarized after data preparation. Because, for diverse research topics and purposes, the requirements of the data set are slightly different, even if all of them are generated from the same origin. The detailed data adjustments and the description of the data set used in each topic are described in Chapter 5, Chapter 6 and Chapter 7 respectively for their research purposes.

3.4.1 Data washing

Data washing starts from the identification and detection process of irrelevant transactions. Firstly, the non-residential buildings are not relevant to any topics in the thesis. This type of property is identified by the property classification, introduced in Section 3.3.1, and should be removed from the

Landgate data.

Secondly, the transactions, that involve multiple properties or units[4] (also called "bundle sale" or "package sale"), are removed as their transaction prices could not reflect the real price of each unit or property, meanwhile, the structural characteristics may not be recorded accurately.

Thirdly, duplicate records are detected by the identity information of properties and transactions. If two or more records have the same property IDs (parcel ID, PIN, and land ID) and the same transaction information (transaction price, date, and application number), they will be identified as duplicate records. The records that have more available information are preserved in the data, but their copies are abandoned.

Then, the records are cleaned up if the properties are restructured or refurnished. These observations are detected when the "year build" is larger than the year of sale. Additionally, the records of off-plan properties should be excluded from the *Landgate* data. Because they refer to the properties which are available to purchase before it has been constructed. Their prices are very volatile in short periods, and their building features are not inherently available. These sale records are identified when the records don't have any valuable building information (only zeros or blanks in variables). Overall, the removal of these records will not cause bias. The full process of data washing

[4] There are variables called "multiple sale" and "number of units" to locate these transactions. "Multiple sale" indicates several properties are sold together in one transaction. "Number of units" means the number of units included in the transaction, such as the two-storey building having 10 flats.

is shown in Table 3.1. After washing, all irrelevant observations are filtered out.

Table 3.1: The details of the data preparation process. shows the number of observations deleted and the reasons why the observations are removed. The overlaps happen because some observations could match multiple washing rules. The non-residential properties only contain the commercial properties. Vacant lands are included in the dataset. The data-washing process does not filter the non-arm-length transactions, they are cleaned in each chapter.

		Number of observations	Filtered observations
Landgate data		1,646,157	
	Non-residential		-56,044
	Bundle sales		-52,746
	Duplicate records		-42,136
	Structure upgrade		-1,123
	Off-plan property sale		-62,993

The non-arm length transactions should be removed also, they are identified by using the lower and upper boundaries of market values provided in the *AURIN* data. The removals are described separately in the data adjustments' sections of Chapter 5, Chapter 6 and Chapter 7, those are along with the description of the study periods and the property types. The summaries of the data subsets are described respectively in each topic.

3.4.2 Data errors treatments

Erroneous zero values[5] in variables are found in the *Landgate* data due to the problem of data collecting and data updating probably. The principal step is to identify whether the information is missing, or if it simply represents the lack of positive values. This issue is corrected according to the concept of gross living area, a suggestion from Krause and Lipscomb (2016). Building features are divided into three groups, essential characteristics, living characteristics, and add-on facilities. Essential characteristics are the most fundamental characteristics of a building including bedroom, bathroom, lounge, kitchen, wall, roof, building area, land area, and the year of build. If missing values and zeros appear in these features, they can be uniformly assumed that the value is not available at the transaction moment. The living characteristics include a dining room, family room, game room, meal room, study room, and carport. They are not as necessary as the essential characteristics. Zeros in these characteristics could indicate that the building doesn't have these features. Missing values mean that the information about the living characteristics is unknown. Then, add-on facilities are the characteristics that have the least significance to residential dwellings for basic living purposes, such as swimming pools and tennis courts. Zeros or missing values represent no presence of the add-on facilities as opposed to missing or unknown values. Such that, all the erroneous zeroes are solved, and the missing values are dealt with by applying some advanced strategies that will be introduced and discussed in Chapter 4.

One special case is the coordinates. Missing values or zeros don't mean the

[5] For example, zeros appear in the number of bedrooms, the number of bathrooms, floor area, and year of build.

coordinates are inexistent, they indicate the geographical information of the properties are not currently updated. Missing values or erroneous zeros shown in the coordinates are filled by geocoding the addresses of properties. Coordinates are generated by *OpenStreetMap API* and *Google Map API* for ensuring accuracy. Then, in the *Landgate* data, the coordinates of all properties are available.

3.4.3 Variable transformation

Some categorical building features are transformed into dummies. If the wall of one property is made of bricks, "brick wall" is 1. Contrariwise, the others except blanks are set to zero. When blanks appear in the wall material variable, it is decided that they are missing values. Because the wall is one of the essential characteristics. The same action is taken on the roof material variable. However, the swimming pool is an add-on facility. Krause and Lipscomb (2016) suggest that a missing or null value may signify no pool belonging to the property as opposed to a missing or unknown value. Different from what is done to the wall and roof, "B/G" and "A/G" are equal to 1 which means the property has a swimming pool. Blanks are set to zero directly. The time dummies, such as quarter dummies, and location dummies, such as LGA dummies, are generated for the specific model specifications.

3.5 Concluding remarks

This chapter describes the data that have been collected, gathered, verified, and prepared in the study area (Perth, Western Australia) for the empirical studies. The main data set is collected from the *Landgate*, it includes information about property identification, transaction identification, structural characteristics, and location of sold properties. The complementary data set originates from the *AURIN*, it contains two time-series data for the maximum sold prices and minimum transaction prices in one year. This chapter prepares a "playground" that the empirical analysis could "play" on.

Initially, two data sets are described in detail, starting from the data identification. The identity variable in the data set could locate transactions that are identified as residential dwelling sales, vacant land sales, and commercial building sales for different research purposes. Then, the data coverage and the quality are discussed. The data set covers detailed characteristics about properties, such as structural characteristics and location, and the information about the transactions, such as transaction time and amount of the transaction. The highlighted section is data quality. Overall, the quality of the data set is adequate, however, it has flaws. The practical issues, such as missing values and erroneous zeroes, are the obstacles that should be necessarily overcome.

In data preparation, three types of characteristics are defined for the missing values and erroneous zeros, essential characteristics, living characteristics, and add-on facilities. They are ranked by their necessity for a building, the different treatments will be applied to them. Through the treatments, the erroneous zeroes issue is solved. The missing value issue needs more advanced methods to deal with, this will be discussed in Chapter 4. The necessary transformation for categorical variables is conducted, such as wall material and swimming pool.

Finally, in this chapter, the prepared data set contains all the transactions of vacant lands and residential properties in Perth from 1988 to 2020. Except for "vacant land", the dominant type of property is "house", which takes more than half of the local market share. It is followed by "group house", "home unit" and "flat". The prepared data set could be subdivided for different research aims, the subsets are applied in the Chapter 5, Chapter 6, and Chapter 7. The selection rules for a specific subset are mainly applied to the temporal variable and the property types. The more detailed and specific adjustments to the prepared data are described in each chapter.

Chapter 4

Methodology

4.1 Introduction

There are two essential questions in transaction-level data analysis, *which estimation method is appropriate for analyzing the data*, and *which technical tool should be applied when a solution of missing value issue is necessary.* This chapter studies the estimation methods that could be applied for *AVM*, and those are thoroughly discussed. The regression model provides a function that describes the relationship between one or more independent variables and a dependent variable. The model forms generally have three types, parametric models, semi-parametric models, and non-parametric models. The tree-based machine learning model, which is non-parametric, is selected and described in Section 4.2. Meanwhile, some commonly used loss functions are discussed below, they are applied for optimizing the price estimation. Interpretation and evaluation are two necessary steps after the estimation. However, for machine learning techniques, the interpretation is not an easy step. A method for calculating the average marginal effect is discussed in Section 4.4. Evaluation of different *AVM* implementations answers the first question, the evaluation criteria are introduced in Section 4.5.

The solutions for the problems associated with missing values are provided in the second half of this chapter. The missing value issue has been discussed in Chapter 3. Currently, the gold standard for solving this is to impute the missing characteristics using the observed information. Along with the improvement of computing techniques, some methods allow missing values to participate in the estimation process. This may squeeze out more valuable hidden information from the missing values rather than imputing them or abandoning them. The technical tools and how they are applied in the regression models are described in Section 4.6.

4.2 Tree-based machine learning model

Consider a data set $\mathcal{D} = \{\mathbf{x}_i; y_i\}_{i=1}^{N}$, where y_i denotes the i^{th} observation of the response variable $\mathbf{Y}$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{iM})$ contains the i^{th} observation of the M predictors and define $X_m = (x_{1m}, \ldots, x_{Nm})'$ as the m^{th} predictor that contains N values. Thus, $\mathbf{X} = (\mathbf{x}'_1, \ldots, \mathbf{x}'_N)' = (X_1, \ldots, X_M)$ is the $N \times M$ data matrix containing the N observations of all the M predictors. Some elements in $\mathbf{X}$ can be missing. It is assumed that there is some relationship between $\mathbf{Y}$ and $\mathbf{X}$, which can be written in the very general specification

$$\mathbf{Y} = f(\mathbf{X}) + \varepsilon \tag{4.1}$$

Here f is some fixed but unknown function of $X_1, \cdots, X_M$, which represents the systematic information that $\mathbf{X}$ provides about $\mathbf{Y}$, and ε is a random error term, which is independent of $\mathbf{X}$ and has mean zero.

4.2.1 Decision tree

Growing a decision tree is a recursive partitioning process. This process could be roughly summarized by two steps (James et al. 2017a):

1. The predictor space, the set of possible values for M predictors, is divided

into J distinct and non-overlapping regions, $R_1, \ldots, R_J$.

2. The predictions of observations are calculated in their region, R_j.

The partitioning process starts from selecting the predictor (X_m) and the cut point (c), such that, the data space is divided into two regions, $R_{Left}(m, c) = \{\mathbf{X}|x_{im} < c\}_{i=1}^{N}$ and $R_{Right}(m, c) = \{\mathbf{X}|x_{im} \geq c\}_{i=1}^{N}$. In this step, all the M predictors are considered and each observed value of the selected predictor is a potential cut-point. All observations are assigned to the left or the right child node from the parent node. The tree stops growing when the stopping criterion is met. The nodes that do not have child nodes are named the terminal nodes. A five-region two-predictor schematic example of the decision tree is shown in Figure 4.1. A more detailed description of the decision tree can be found in Breiman et al. (1984) and Hastie et al. (2009).

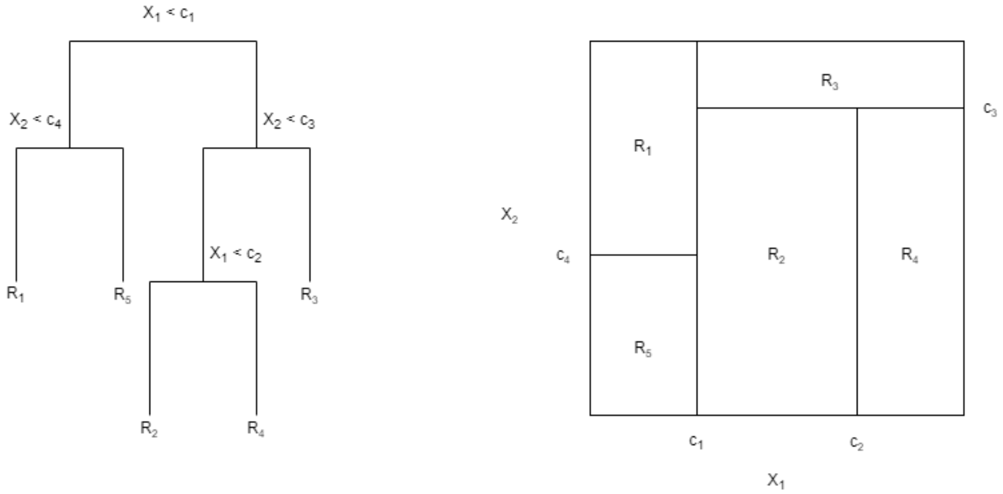

Figure 4.1: The example of the decision tree. The left: the tree corresponding to the partitions in the right panel. The right: an example of recursive binary partitioning on a two-dimensional feature space.

Single decision trees are highly interpretable. The entire model can be completely depicted by a binary tree as shown in Figure 4.1 that is easily visualized. Trees more closely mirror the human decision-making process. However, the predictive performance may not be as good as some regression approaches.

4.2.2 Gradient boosting machine

Gradient boosting machine (*GBM*) (Friedman 2001, 2002) is one of well well-known tree-based machine learning methods. As discussed in Chapter 2, the performance of *GBM* is competitive in empirical analysis. The fundamental of *GBM* is the decision tree[1] (Breiman et al. 1984), its detailed algorithm is described below.

GBM is a model for improving the accuracy of predictions resulting from a decision tree. *GBM* grows trees sequentially: each new tree is grown using information from previously grown trees, and all the trees are in the same "family line" (James et al. 2017a). The preliminary of *GBM* is to decide the hyperparameters[2], J terminal nodes (or the depth (or size) of each tree d), the number of trees (or iterations) B, and the learning rate λ. It starts with an initial guess of response $F_0(\mathbf{X})$,

$$F_0(\mathbf{X}) = \gamma_0. \tag{4.2}$$

[1] In this article, a decision tree is for solving regression problems. The description introduces the characteristics of a decision tree when it builds a regression model. For the classification problem, the description is slightly different.

[2] These parameters could be called the tuning parameters also. They must be tuned to achieve the best performance. The depth (or size) of each tree d is commonly used in implementation packages rather than J terminal nodes.

In each iteration b, the current "pseudo"-residuals are calculated using the previous information.

$$\tilde{y}_{ib} = -\Big[\frac{\partial L(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\Big]_{F(\mathbf{X})=F_{b-1}(\mathbf{X})}, i = 1, \ldots, N. \tag{4.3}$$

where $L(\cdot)$ is the loss function, which is defined as a function of the estimated values ($\hat{\mathbf{Y}} = F(\mathbf{x}_i)$) and the actual values ($\mathbf{Y}$). Then, a J terminal nodes' tree is grown using the current "pseudo"-residuals $\tilde{y}_{ib}$ and all observations $\mathbf{x}_i$.

$$\{R_{jb}\}_1^J = J\ terminal\ nodes'\ tree(\{\mathbf{x}_i; \tilde{y}_{ib}\}_1^N). \tag{4.4}$$

The outputs of the tree in the iteration b are

$$\gamma_{jb} = \arg\min_{\gamma} \sum_{\mathbf{x}_i \in R_{jb}} L(y_i, F_{b-1}(\mathbf{x}_i) + \gamma). \tag{4.5}$$

Finally, the predictions are updated using the outputs.

$$F_b(\mathbf{X}) = F_{b-1}(\mathbf{X}) + \lambda \cdot \gamma_{jb} \cdot \mathbf{1}(\mathbf{x}_i \in R_{jb}). \tag{4.6}$$

After repeating B times, the final output predicted values are

$$\hat{\mathbf{Y}} = F_B(\mathbf{X}) = F_0(\mathbf{X}) + \sum_{b=1}^{B} \lambda \gamma_b. \tag{4.7}$$

Moreover, Friedman (2002) creates a stochastic variant in that the trees are trained on a subset of randomly drawn observations without replacement from the full training set at each iteration. *GBM* aggregates a sequence of decision trees, this could substantially improve the predictive performance of trees (James et al. 2017b). However, it also makes the process more complicated than decision trees, which can not be easily visualized. Such that it has to be with the help of some interpretation tools for explaining the effect of the independent variable (X_m) on $\mathbf{Y}$.

4.3 Loss functions

Estimation is the process of finding an approximation of the actual value for some purposes. The loss function is used for selecting an optimal approximation. In the optimization, it seeks to minimize a loss function that maps the decisions of estimate selection to their associated costs[3]. In mortgage portfolio management, for instance, the fund managers may be more concerned about the overpriced properties than the undervalued properties in their cases, due to hedging the total risk of the portfolio. The overpriced should suffer more losses. Such that, the decision makers' preference and their economic intuition should be represented by the suitable function for the loss optimization (Frisch 1981). Depending on the context of the problems, there are many diverse loss functions. Some of the more popular choices are the quadratic loss function (A.K.A. the squared error loss function), and the quantile loss function, which has a special case – the absolute error loss function[4].

4.3.1 Quadratic loss function (Squared error loss)

The quadratic loss function is a common measure of how accurate a model is. As the name suggests, the word "quadratic" means that the highest term in the function is a square, and the difference between the predicted values

[3] It is a convention that the theory is formulated in terms of trying to minimize the losses rather than trying to maximize the gains, but they make no difference.

[4] These two loss functions are mostly used for solving regression questions. For classification questions, the losses could be described in the other way, such as using misclassified rates. In this research, the classification questions are not covered.

and the actual values are squared, shown in Eq. 4.8,

$$L = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{4.8}$$

where y_i is the ith observation of $\mathbf{Y}$, $\hat{y}_i$ is its estimate. Such that, the loss is always positive, and it is only concerned with the magnitude of the difference irrespective of their direction (or sign). However, due to squaring, predicted values, which are significantly away from their actual values, are penalized heavily in comparison to the less deviated predictions. Such that, the predictions could not be robust and accurate when there are outliers. In addition, the quadratic loss function is continuous and differential, these mathematical properties make it easier to calculate gradients. The application of this loss function in regression problems will lead to forecasts of the mean when the loss function is minimized (Hyndman and Athanasopoulos 2018).

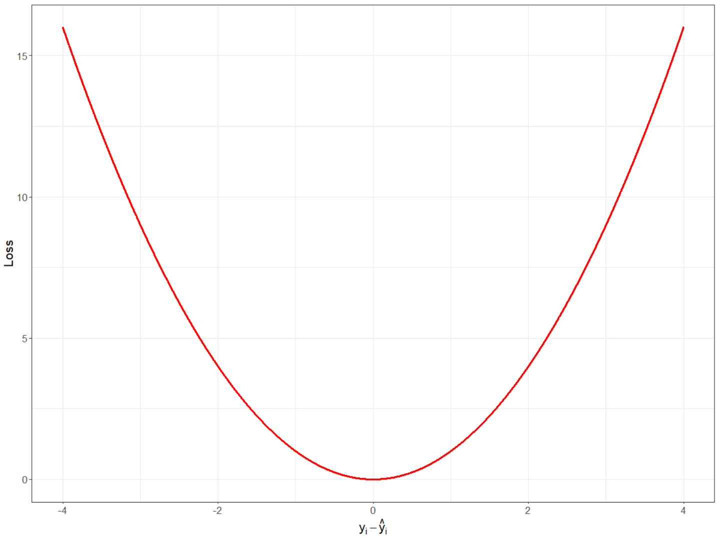

Figure 4.2: The schematic explanation of the quadratic loss function. The x-axis is the difference between the actual values and the predicted values. The range of these differences is limited in the interval $[-4, 4]$. The y-axis shows the losses of each pair of values.

4.3.2 Quantile loss function

As the name suggests, the quantile loss function is employed to predict the quantile of the response variable (y_i). Inherently, the forecasts are along with their uncertainty. As shown in Eq. 4.9, the q is the target quantile of the response. The distances between actual values and predicted values are computed and weighted by $1 - q$ and q, which depends on the sign of the difference between the actual and predicted values.

$$L(q) = \sum_{i:y_i<\hat{y}_i} (1-q)|y_i - \hat{y}_i| + \sum_{i:y_i \geq \hat{y}_i} q|y_i - \hat{y}_i| \tag{4.9}$$

In most practical forecast problems, the uncertainty of predicted values is keen to be introduced. Because a forecast with its uncertainty is quite helpful for decision-making in risk management or portfolio management, rather than a simple predicted value of conditional mean provided by the minimized quadratic loss function. The prediction interval could be computed using the standard error of the prediction. However, it is based on an assumption that residuals ($y_i - \hat{y}_i$) should have constant variance across values of independent variables. If this assumption is violated, the quantile loss function could be an alternative for solving the problem. It provides the predicted quantiles of the response (predicted value with its uncertainty), and these predicted quantiles could construct sensible prediction intervals even for residuals with non-constant variance or non-normal distribution.

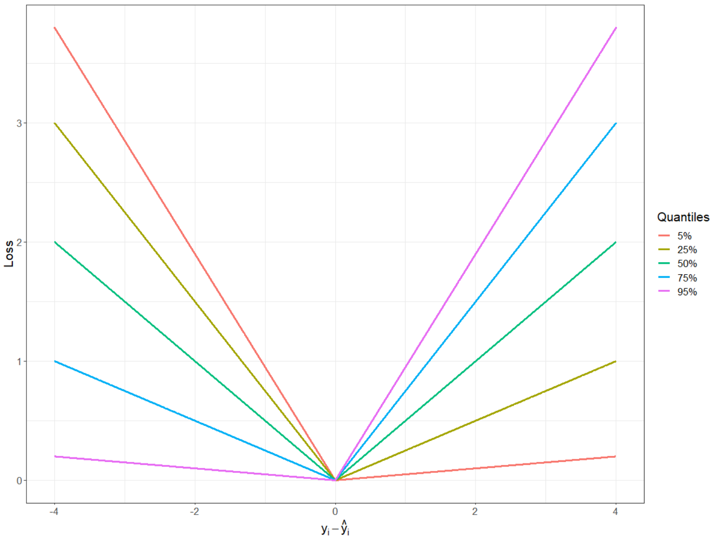

Figure 4.3: The schematic explanation of the quantile loss function. The x-axis is the difference between the actual values and the predicted values. The range of these differences is limited in the interval $[-4, 4]$. The y-axis shows the losses of each pair of values. Except for the loss function for 50% quantile, the rest loss functions for quantiles are asymmetric, such as the functions for 5% quantile, 25% quantile, 75% quantile and 95% quantile in the graph.

Special case: absolute error loss

A special case of quantile loss function is to calculate the median (0.5 quantile). When $q = 0.5$ is substituted into Eq. 4.9, the outcome is similar to the absolute errors, shown in Eq. 4.10, and they are mathematically equivalent in terms of optimization. The absolute error loss is directly measured by the absolute difference between predictions and actual values. A forecasting method that minimizes this loss will lead to forecasts of the median(Hyndman and Athanasopoulos 2018).

$$\begin{aligned} L(0.5) &= \sum_{i:y_i<\hat{y}_i} 0.5|y_i - \hat{y}_i| + \sum_{i:y_i\geq\hat{y}_i} 0.5|y_i - \hat{y}_i| \\ &= 0.5\sum_{i=1}^{N} |y_i - \hat{y}_i| \qquad (4.10) \\ &\propto \sum_{i=1}^{N} |y_i - \hat{y}_i| \end{aligned}$$

The absolute error loss measures the magnitude of the difference without considering the direction (or sign) which is the same as the quadratic loss function. However, it is not always differential, such that, it needs more complicated tools to compute the gradients. In addition, absolute error loss gives the equal weight to every residual $(y_i - \hat{y}_i)$. Predicted values, which are significantly away from their actual values, are penalized equally in comparison to the less deviated predictions. Thus, absolute error loss is more robust to outliers since it does not make use of squares. Besides the previous two loss functions, there are some other loss functions, such as Huber loss function (Huber 1964), that are not applied in this thesis.

4.4 Model interpretation tools

The interpretable models could explain the effect of predictors on the response directly, such as the linear model. However, not all models are inherently interpretable, such as the non-parametric models. One set of enhancements for developing the ability to interpret such models or parts of models is termed model-agnostic interpretation methods, such as partial dependence plot (PDP), marginal plots, accumulated local effect (ALE), global surrogate

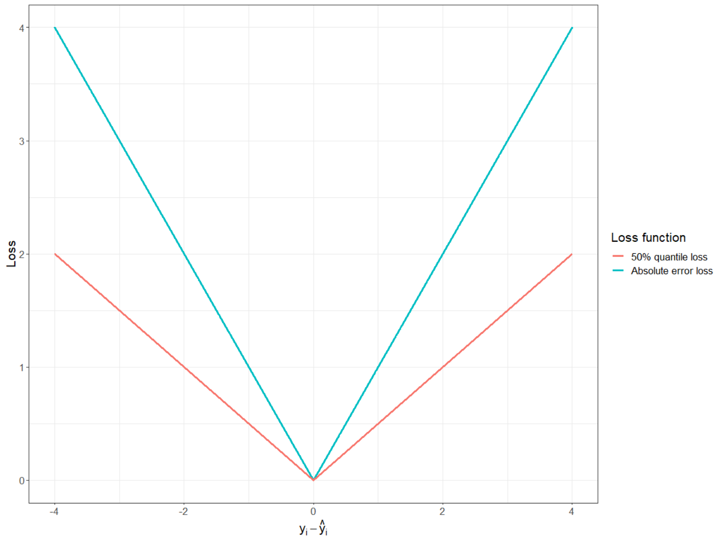

Figure 4.4: The schematic comparison between 50% **quantile loss function and the absolute error loss function.** The x-axis is the difference between the actual values and the predicted values. The range of these differences is limited in the interval $[-4, 4]$. The y-axis shows the losses of each pair of values. They are equivalent when the total loss is assessed.

and permutation feature importance. These methods have three desirable aspects (Ribeiro et al. 2016):

- Model flexibility allows model developers to freely use any models they like without worrying about model interpretation.
- Explanation flexibility does not limit methods to a certain form, such as graphical form or numerical form when they are applied to explain models.
- Representation flexibility allows methods to be able to use different feature representations as the model being explained.

The commonly applied model interpretation tools are *PDP* and *ALE*. Friedman (2001) apply partial dependence plot for interpreting *GBM*, that provides a comprehensive summary of the dependence of the response on the joint values of the input variables. Unfortunately, such visualization is limited to low-dimensional arguments. Different from *PDP*, accumulated local effect calculates the marginal effects between two conditions or values of X_m, and it could work with higher dimensional arguments. In this thesis, the *ALE* is selected to interpret the implementations using *GBM*, because it is a faster and unbiased alternative to a partial dependence plot, and it avoids the extrapolation problem in *PDP* (Zhu and Apley 2018).

4.4.1 Accumulated local effect

The accumulated local effect describes how predictors influence the prediction of a non-interpretable model on average. The *ALE* of the predictor X_m is defined as

$$ALE(x_m) = \int_{z_{0,m}}^{x_m} E[\frac{\partial f(X_1, \cdots, X_M)}{\partial X_m} | X_m = z_m] dz_m - c \tag{4.11}$$

where $z_{0,m}$ is an approximate lower bound of X_m. The constant c is chosen such that $ALE(x_m)$ has a mean of zero to the marginal distribution of X_m. An ALE of the effect of x_m (a particular value of X_m) is a plot of an estimate of $ALE(x_m)$ versus x_m, and it shows the effect dependence of $f(\cdot)$ on x_m. The estimate of the ALE is

$$\widehat{ALE}(x_m) = \sum_{k=1}^{k(x_m)} \frac{1}{n_{mk}} \sum_{x_m \in A_m(k)} [f(z_{k,m}, \mathbf{x}_{i,\backslash m}) - f(z_{k-1,m}, \mathbf{x}_{i,\backslash m})] - \hat{c}. \tag{4.12}$$

This equation replaces the integral in Eq. 4.11 with a summation and the derivative with a finite difference. $\mathbf{x}_{i,\setminus m}$ is $\mathbf{x}_i$ with all predictors except the mth ($\setminus m$). $A_m(k)$ $(= \{(z_{k-1,m}, z_{k,m}], k = 1, \cdots, k(x_m), \cdots, K\})$ is one of the K intervals that the sample range of X_m is sufficiently fine partitioned, $z_{k,m}$s are the values of interval boundaries. n_{mk} is the number of observations that fall into the kth interval $A_m(k)$. $k(x_m)$ denotes the index of the interval into which x_m falls, $x_m \in (z_{k(x_m)-1,m}, z_{k(x_m),m}]$. The details of ALE are shown in Apley and Zhu (2020).

ALE helps to extract the average marginal effects of variables from the machine learning techniques. However, this solution is not perfect compared to some parametric models. The limitation of this method is the statistical inference. After the average marginal effects are calculated, the values lack statistical information, such as standard error. Even though the bootstrapping method could be helpful, it still is hard to clarify whether the marginal effects are statistically significant or not.

4.5 Evaluation criteria

For evaluating the accuracy of models or the precision of indices introduced above, the difference between the price estimates and their true values should be calculated. The models estimated for price predictions use the logarithm response. This could help to alleviate efficiency losses due to heteroskedasticity during estimation. However, these predictions are for real properties, the industrial practitioners may be more interested in predictions of market values, not in predictions of log market values. Such that, the log prices are then back-transformed to the normal scaled prices. They are reverted with the smearing estimator of Duan (1983).

$$\hat{P}_i = exp(\hat{p}_i)\Big(\frac{1}{N}\sum_{i=1}^{N} exp(\varepsilon_i)\Big). \tag{4.13}$$

Therein, $\hat{p}_i$ is the log-scaled price predictions, $\hat{\varepsilon}_i$ is the estimated error term in the models, such as GAM and GBM described above. N is the number of observations in the model. There are other methods of re-transformation, but Schulz et al. (2014) find that Eq. 4.13 shows benefits.

The assessments for models and indices use percentage errors of the normal scaled prices, that give the equal error evaluation scale to the cheap properties and the expensive ones. This process follows the model evaluation in Schulz and Wersing (2021). The percentage error (e_i) is calculated as the following equation:

$$e_i = \frac{P_i - \hat{P}_i}{\hat{P}_i} \times 100,\ \ i = 1, \cdots, N. \tag{4.14}$$

where P_i is the normal scale prices. Then, the percentage errors are summarized by different error metrics. Diverse error metrics could provide the different ranks of models' performance, the metrics selection follows the recommendations in Steurer et al. (2021). The selected evaluation criteria are summarized in Table 4.1.

PER(a) gives the fraction of percentage errors that fall out of the interval $[-a, a]$. For instance, given $a = 10$, we assume $PER(10) = 50$. This informs that there are half of the valuations those errors are larger than 10 percent of their valuations. These evaluation criteria are used throughout the thesis.

Table 4.1: The evaluation criteria. shows the metrics used to assess the performance of models based on prediction percentage errors (e_i). The percentage error range calculates the fraction of percentage errors that fall outside the interval $[-a, a]$. $\mathbf{1}(\cdot)$ is the indicator function that takes the value one if the argument is satisfied and value zero otherwise.

Metrics Name	Formula
Mean Percentage Error	$MeanPE = \frac{1}{N_{test}} \sum_{k=1}^{N_{test}} [e_k]$
Median Percentage Error	$MedianPE = median[e_i]$
Mean Absolute Percentage Error	$MAPE = \frac{1}{N_{test}} \sum_{k=1}^{N_{test}} \lvert e_k \rvert$
Root Mean Squared Percentage Error	$RMSPE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} [e_i]^2}$
Percentage Error Range	$PER(a) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(\lvert e_i \rvert \geq a)$

4.6 The solutions for missing values

Data with missing values are ubiquitous and inevitable in many applications, due to system errors or bad process of data collection, for example. Before dealing with missing values, it is suggested to distinguish the missing data mechanism, which concerns the relationship between missingness and the values of variables in the data. Rubin (1976) formalizes the concept of the missingness mechanism, through the simple device of treating the missing data indicators as random variables and assigning them a distribution. There are three mechanisms, missing completely at random ($MCAR$), missing at random (MAR), and not missing at random ($NMAR$). Most of the theory and practices build upon the MAR assumption. For instance, Dempster et al. (1977) propose a variant of the expectation-maximization algorithm which provides unbiased estimators even in the presence of missing data, under MAR assumption. This assumption is also critical to most of the missing data imputation literature, because it guarantees the distributions of the same missing data indicator are identical. When the assumption does not hold, many algorithms may break down.

Except for dropping all missing values, in prediction tasks, practitioners often first impute missing values and then train a model on the imputed data set(s). This approach is called "impute-then-regress". Multiple imputation is the "gold standard" of this approach for statistical inference with missing data. However, this approach requires imputing the potential missing value in the testing data sets. The other approach is to deal with missing values and regressing tasks simultaneously, instead of sequentially, which could be called "regress with missing". The missing value node strategy described below is one of this approach.

4.6.1 Gold standard – multiple imputation

Multiple imputations refer to the procedure of replacing each missing value with a vector of imputed values that are calculated in each of the simple single imputations. The single imputation could be described as follows:

$$\Phi_m(\mathbf{X},\ MI) = \begin{cases} X_m & if\ MI_m = 0, \\ \phi_m(X_{obs}) & if\ MI_m = 1. \end{cases} \tag{4.15}$$

The $\Phi_m(\cdot)$ is a function of $\mathbf{X}$ and the missingness indicators (MI) for M predictors. If MI is equal to 0, it means the predictor does not contain missing values, the predictor could be directly used in the regression model. Otherwise, the missing values should be imputed by imputation function $\phi_m(\cdot)$ using all the observed non-missing predictors. Then, the imputed data set ($\tilde{\mathcal{D}}$), the output of function $\Phi_m(\mathbf{X},\ MI)$, could be treated as a complete data set.

Multiple imputation simply repeats this process k times and independently generates k completed data sets. Multiple imputation is first proposed in Rubin (1978), and a comprehensive treatment is given in Rubin (1987).

Single imputation has the practical advantage of allowing standard completed data methods of analysis to be used. However, a single value for a missing value could not validly reflect sampling variability or display uncertainty about the correct model. Compared with the single imputation, multiple imputation generates a vector of independent imputed values. Thus, the pitfall of a single imputation could be avoided, and the sampling variability or uncertainty would not be obliterated. The payment of this improvement is that multiple imputations take more work to compute the imputations and analyze the results.

The full analysis on data set $\mathcal{D} = \{\mathbf{x}_i; y_i\}_{i=1}^{N}$ could follow a two-step procedure named as *Impute-then-regress procedure*. The incomplete data set $\mathcal{D}$ is first imputed k times using a function $(\Phi_m^{(k)})$, k imputed data sets $(\tilde{\mathcal{D}}^{(k)})$ are generated. The regression model is applied on the k completed data sets $(\tilde{\mathcal{D}}^{(k)})$ separately, a family of models is the outputs, $f^{(1)}(\mathbf{X}^{(1)}), \cdots, f^{(k)}(\mathbf{X}^{(k)})$. The final prediction of the test sample, if train and test sets are split, is the average of predictions computed using each of $f^{(1)}(\mathbf{X}^{(1)}), \cdots, f^{(k)}(\mathbf{X}^{(k)})$. Performing multiple imputations of the test set and averaging the predictions leads to a consistent estimator(Josse et al. 2020). In addition, *Impute-then-regress procedure* is Bayes consistent for all missing data mechanisms and almost all imputation functions(Morvan et al. 2021).

4.6.2 Missing value node strategy

The missing value node strategy is designed for tree-based models, which allows them to handle missing values in the fitting procedure. Dealing with missing values and fitting a model could be processed simultaneously, "regress with missing". Missing value node strategy treats missing values (NA) as a new category of data. Thus, trees could group and assign the observations, that have missing values in the primary selected predictor of one split, to a new node, rather than to the left node or the right node. The non-missing values are partitioned as a normal decision tree. The observations in the missing value nodes could be partitioned again as long as the estimation would be improved[5]. As missing values are assigned to a new special node, there is no chance of misclassifying observations, meanwhile, the observations with missing information are studied under a different path. A schematic example is shown in Figure 4.5. If there are missing values in the selected predictor (X_1) of the top split, the observations with missing values are assigned to $R_{na,1} = \{\mathbf{X}|x_{i1} = NA\}_{i=1}^{N}$. The rest is distributed to the left if $x_{1i} < c_1$, and to the right if $x_{1i} \geq c_1$ $(i = 1, \ldots, N)$. If X_2 is unobserved in some observations, they are assigned to $R_{na,2}$ or $R_{na,3}$

[5] Missing value node is generated in each split as a security mechanism, even the primary selected predictor of one split has no missing value. This is for prediction purposes, in case missing values appear only in test data.

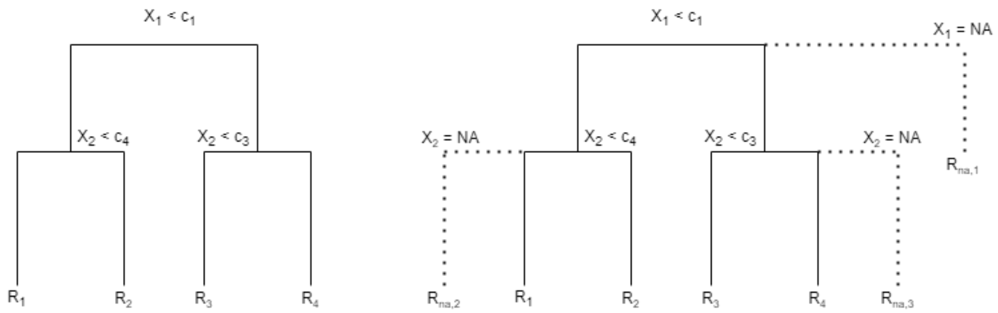

Figure 4.5: The example of "missing value node" approach. The left: a normal tree without missing nodes. The right: a tree with missing nodes $R_{na,1}$, $R_{na,2}$ and $R_{na,3}$ (rules: $\{X|X_1 = NA\}$ and $\{X|X_2 = NA\}$.).

4.6.3 Other strategies for missing values

Similar to the missing value node strategy, Breiman et al. (1984) design a type of variable, called "surrogate variable", which could be used to deal with missing data during estimation. The surrogate variable is to search for the most appropriate one in the pool of variables as a proxy when the selected predictor has missing values. A schematic example[6] is shown in Figure 4.6. If there are missing values in the selected predictor (X_2) of the left split, the surrogate variable (X_3) could help to decide whether the observations with a missing x_{i2} are assigned to the left. If x_{i2} is not missing, $R_1 = \{\mathbf{X}|x_{i1} < c_1 \textit{ and } x_{i2} < c_4\}_{i=1}^{N}$. Otherwise, $R_1 = \{\mathbf{X}|x_{i1} < c_1 \textit{ and } x_{i3} < c\}_{i=1}^{N}$. Compared to the missing value node strategy, the surrogate variable could misclassify the observations to a child node. Because it is not easy to search for a perfect surrogate variable in the observed predictors' pool as a proxy.

[6] The details about surrogate variable could be found in Breiman et al. (1984) and Hastie et al. (2009).

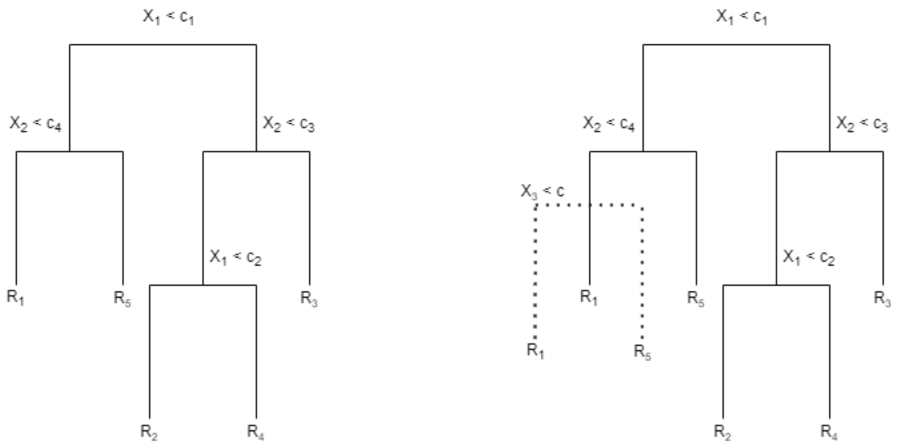

Figure 4.6: The example of surrogate split. The left: a normal tree without surrogate variables. The right: a tree with surrogate variable X_3 in case X_2 is missing in some observations.

Another strategy allows the imputation and the regression model to be learned simultaneously, named an "on-the-fly imputation". Tang and Ishwaran (2017) and Bertsimas et al. (2021) achieve this approach in random forest and linear models respectively. This approach is not that significantly different from "impute-then-regress". In addition, this approach currently could not store the records about the imputation during the estimation process. Thus, the output results are difficult to trace back, and this strategy is not that friendly when making predictions for testing data. Compared to "on-the-fly imputation", mean imputation (Wilks 1932) has a simple process, and is easy to understand. Mean imputation is to replace missing values in one variable by the mean of the observed values of the same variable. This approach is computationally efficient and straightforwardly applicable. However, this approach could lead to severely biased estimates even if the data are *MCAR* (Jamshidian and Bentler 1999). In addition, mean imputation could underestimate the standard error of variables. Because variables would be invariant if the missing values in them are replaced by the sample means.

4.7 Concluding remarks

This chapter outlines the applied tools for the analysis of residential real estate on two aspects, the appropriate method for estimations and the missing value issue. It begins with the description of the gradient boosting machine, which is the representative model for conducting *AVM*, due to its competent performance. Then, the loss functions used to optimize the estimations have been introduced, which could be designed for the specific preference of estimation users or economic intuition. For general practical or academic purposes, two types of loss functions (quadratic loss function and quantile loss function) are commonly used. Therein, the absolute error loss function is a special case of quantile loss function when the quantile is 0.5. The interpretation and evaluation of models are also discussed. The relationship between the independent variables and the response could be analyzed by the accumulated local effect for non-parametric models, which is an efficient and unbiased alternative to the partial dependence plot. Evaluation criteria select the commonly used error metrics, recommended in Steurer et al. (2021).

The rest of this chapter recites two paths for dealing with missing values, "impute-then-regress" and "regress with missing", except dropping them. The former imputes all missing values before doing regression. The "gold standard" is multiple imputations, that could generate multiple different and independent imputed data sets. The latter handles missing values and does regression simultaneously instead of sequentially. The missing value node strategy is one of its representatives for the tree-based machine learning models. Comprehensively, the other methods are discussed, but not applied. Because the methods used in this thesis are better or equivalent to the other strategies for missing values.

The different tools are applied under different combinations for diverse research purposes in three topics. In Chapter 5, *GBM* is applied with two loss functions and three different strategies for missing values. However, the tools for interpretation are not used as the main purpose is to investigate the accuracy of predictions in the comparison of various implementations. In Chapter 6 and Chapter 7, the best implementation from Chapter 5 is applied to construct the residential property price index and implement the analysis for real option. *ALE* is applied to explain the effects of input variables from *GBM*. The details are described in the section on research design respectively for each topic.

Chapter 5

Price prediction with missing values

5.1 Introduction

Automated valuation models (AVM) use one or more mathematical techniques to provide an estimate of the price of a particular property at a specified time without or with less human intervention post-initiation (RICS 2021). They are well applied in the valuation field of real estate analysis for diverse purposes, such as mass appraisal, mortgage portfolio management, taxation, cost/benefit analyses for potential public expenditure, and so forth, by different stakeholders, such as valuers, mortgage lenders, regulators, city planners, etc.

In this chapter, an empirical study of AVMs is conducted, that can provide accurate predictions while missing values occur in the *data set*, for the housing market of Perth metropolitan area, Western Australia. Regarding the estimation modeling, the gradient boosting machine (GBM) is applied. Because, it outperforms the other machine learning models in most of the previous empirical studies (Kok et al. 2017, Mayer et al. 2019, Schulz and Wersing 2021). Additionally, GBM is flexible, it could apply different types of loss functions for regression problems (Friedman 2001). Meanwhile, it allows categorical and numerical variables as inputs, and it could automatically and simultaneously deal with missing values during estimation.

Missing values in the transaction-level data seriously obstruct conducting a regression model. Eliminating the incomplete observations, containing missing value(s) in one or more variables, is by far the most common method of dealing with missing values. This is the approach adopted in many studies, for example, Schulz et al. (2014), Steurer et al. (2021) and Kolbe et al. (2021). The advantage of this extreme approach is simplicity since a standard analysis can be applied without further modifications (Little and Rubin 2002). However, it not only often leads to a sizeable loss of observations available for estimation, it may potentially have a selection bias. The non-missing part of the dropped observations could be also valuable in estimation. Simply ignoring them is unwise and may reduce the practicability of a model. Because the predictions in the forecasting stage can not be accurately made or not be even provided once the property characteristics are missing. The second widely used approach is imputation, which is the process of replacing missing data with substituted values. The substituted values are imputed using known information, such as the mean of other non-missing cases or the predicted value based on other observed variables. To reduce the noise due to imputation, Rubin (1987) develop a method for averaging the outcomes across multiple imputed values. The imputation approach requires a further step of calculation or regression for "filling" the missing values in the data preparation before estimation. This may make the procedure more complicated and hard-to-operating, compared with dropping the incomplete observations. Recently, combined with machine learning techniques, the methods used for imputation are more "accurate" and allow for imputation on basically any type of data (see examples Stekhoven and Bühlmann 2012, Tang and Ishwaran 2017, Morvan et al. 2021). But, the procedure is still complex. The approach preferred is to preserve all relevant observations and handle missing values before estimation or when the estimation is on the fly, such as Knight et al. (1998), Kagie and Van Wezel (2007) and Hinrichs et al. (2021).

*GBM*s are trained under these three schemes. One scheme applies the complete case strategy for missing values that *GBM*s only learn from the complete cases, which is adopted in many previous studies. The second scheme is to train *GBM*s on the original data, and the missing value node strategy is deployed for unknown values. The last one is to train *GBM*s on the data that the missing values are filled by multiple imputation strategy. The linear hedonic models are chosen as the benchmarks for evaluating the *GBM*s' performance. Two types of loss functions are used in both *GBM*s and linear hedonic models. The squared error loss function (LS) is the most commonly used in the previous literature. The absolute deviation loss function (LAD) may provide more robust estimates than LS when there are some outlier values. This ensures that the effect of other interference is eliminated and controlled. Models are fitted with a rolling window procedure. Out-of-sample predictions are to assess the predictive performance of models. This analysis could answer one of the main research questions, "What to do with observations that suffer from missing values when estimating prices of residential properties using an automated valuation model?", and two sub-questions, "whether *GBM* is a competent model that shows the similar results as the previous studies?" and "which missing value strategy or model implementation is recommended?".

Our main findings are as follows. First, the *GBM* with missing node strategy is recommended to solve the issue of missing values automatically. *GBM*s outperform the linear hedonic models, no matter which loss function and missing value strategy are applied. For the *GBM*s trained on the original data, around 42% (15%) of the valuations deviate by more than 10% (20%) from the sale prices. The *GBM* implementations using only complete cases show worse performance in the *GBM*'s family, but they outperform the benchmarks. This is in line with the results in Kagie and Van Wezel (2007) and Mayer et al.

(2019). Second, the results are more accurate when the absolute deviation loss function is applied in models. *GBM* applying *LAD* is more recommended than applying *LS*. Third, missing values may not have to be arbitrarily or experientially removed before doing estimation, when a method is available to robustly deal with them. *GBM* using *LAD* is a competent choice, which could provide accurate results and prevent potential sizeable data loss while there are missing values in data.

5.2 Data for *AVM* implementations

For the *AVM* aspect, the *data set* used is a subset of the *landgate* data which have been cleaned from 2015Q1 to 2020Q4. In total, there are 3,638 non-arm-length transactions dropped. Each observation documents information about transaction information, parcel details, and dwelling details, such as sale price, date, location, property types[1] and the number of housing features. From 2015 to 2020, there were 21,231 (12% approx.) newly established properties sold, 174,137 transactions (observations) in total. Missing values are preserved in the *data set*.

The summary of variables is presented in Table 5.1. It summarizes the statistics for 174,137 observations. The price is deflated by the *RPPI* from the *ABS*, the reference period is the financial year 2011–2012. In the table, some characteristics show low levels of missingness, the rates are higher than 0.09%

[1] The most property type is houses (66%), followed by group houses (13%) and the rest types.

but lower than 1.6%. The missing rate of floor area, however, is around 35%, which is much heavier than the others. Compared with other data sets used in real estate literature[2], the missing rates of the *data set* is mild.

Table 5.1: The summary statistics for the sold residential properties in the data set, Jan 2015 - Dec 2020.

Variable	NA num	Mean	S.D.	Variable	NA num	Mean	S.D.
Price ($,000)	0	573.221	416.015	Dining	0	0.685	0.468
Land size (m^2)	0	927.504	5,054.839	Family	0	0.594	0.496
Floor area (m^2)	61,114	155.240	75.039	Game	0	0.203	0.407
Age (year)	1,243	27.504	21.937	Meal	0	0.308	0.464
Bedrooms	1,139	3.221	0.854	Study	0	0.197	0.407
Bathrooms	290	1.645	0.599	Car ports	0	1.499	0.770
Lounge	1,139	1.007	0.086	Tennis court	0	0.001	0.028
Kitchen	1,139	1.010	0.105	Pool	0	0.169	0.375
Tile-roof	2,707	0.798	0.401	latitude	0	-31.970	0.167
Brick-wall	169	0.935	0.247	longitude	0	115.856	0.096

5.3 Analysis design

To provide up-to-date price predictions, models are periodically maintained by adding new market transactions and removing old transactions in real estate industrial practice. Thus, the rolling windows strategy is applied to study the predictive performance of *AVM*s. The training data sets contain observations in a two-year (eight-quarter) length period. The first training period is from the

[2] The comparison with the other data sets used in the literature was discussed in Chapter 3.

first quarter of 2015 to the fourth quarter of 2016. The window is repeatedly shifted by one-quarter and re-apply the models. The testing data sets are constructed by observations transacted in the following quarter of training periods. In total, we generate 16 different windows on the *data set*, the fraction of missingness in each window is presented in Table 5.2.

Table 5.2: The proportion of observations in each window with missing characteristics. All numbers are in percentage.

Missingness (%)	Window 1	Window 2	Window 3	Window 4
Training	38.56	38.39	38.27	38.08
Testing	38.03	38.17	37.55	38.56
Missingness (%)	Window 5	Window 6	Window 7	Window 8
Training	38.13	37.94	37.75	37.68
Testing	36.47	37.55	37.01	37.14
Missingness (%)	Window 9	Window 10	Window 11	Window 12
Training	37.58	37.40	37.20	37.0
Testing	36.64	36.48	35.25	36.5
Missingness (%)	Window 13	Window 14	Window 15	Window 16
Training	36.66	36.24	36.04	35.71
Testing	33.28	36.05	34.86	33.78

Model specifications are similar to those of standard hedonic price models. The response is the logarithm of transaction price[3] and the predictors are the characteristics that could explain variance in the price.

$$y_i = p_i = log(P_i) = f(\mathbf{x}_i) + \epsilon_i. \tag{5.1}$$

[3] The log-scaled outputs are transformed back to natural scaled sale prices. Fitting the models to prices in natural scale, or to prices in logs with back-transformation, that is the question. Schulz and Wersing (2021) find the latter has advantages.

where $f(\cdot)$ is an unknown function, and we use two different models to place some structure on this function. The whole process is implemented using the statistical software R, version 4.1.2 (R Core Team 2021).

5.3.1 Gradient boosting machine

In *GBM*s, the predictors include structural characteristics, location features (latitude, longitude, and local government area (*LGA*s)), a temporal feature (continuous quarter number), and property types. Two groups of *GBM*s are trained on the modified training sets (complete cases strategy, C, and multiple imputation strategy, MI). One is to omit the observations with missing values, the other is to fill in the missing values using the available information. The third group uses the original training sets. All of them are grown using either *LS* ($GBM(LS)$, $GBM(LS, MI)$ and $GBM(LS,\ C)$) or *LAD* (GBM (LAD), GBM (LAD, MI) and GBM $(LAD,\ C)$) separately. The hyperparameters are tuned by grid searching and ten-fold cross-validation on the training sets. We use root mean squared error and mean absolute error to measure the performance in the tuning stage. The optimal hyperparameters are selected by comprehensively considering two metrics[4].

[4] If two sets of hyperparameters have the same ranking, we choose the most parsimonious and conservative one. The idea is similar to the one standard error rule that is applied in Hastie et al. (2009).

5.3.2 Linear regressions

Referring to the benchmark linear hedonic model in Leishman et al. (2013)[5], distance to CBD is added to explain the location effect. In addition, a subset selection is applied to the structural characteristics, and ten-fold cross-validation is used to decide the optimal form of linear regressions[6]. The lounge, family room, and tennis court are excluded due to their low relevance. Therefore, linear models use structural characteristics excluding lounge, family room and tennis court, location features (distance from CBD[7] and LGAs dummies), a temporal feature (quarter dummies) and dummies of property types[8]. For a fair competition with GBMs, both of LS ($Linear$ (LS, MI)) and LAD ($Linear$ (LAD, MI)) are employed.

When multiple imputation strategy is applied, missing values are imputed by missForest[9], which shows the best performance of missing value imputation in Tang and Ishwaran (2017). For linear models, due to they can not inherently and internally deal with missing values, the multiple imputation strategy is applied by default. Otherwise, 35% of testing data can not be predicted by linear models due to missing floor area. All implementations are summarized in Table 5.3, and all of them can run automatically.

[5] Because this chapter studies the local housing market in Perth also. Moreover, we have the same data source ("*Landgate*").

[6] The form shows the lowest root mean squared error and mean absolute error.

[7] The distance between the location of the property and Perth CBD (city town hall) is calculated using coordinates.

[8] Since linear regression is unable to handle categorical variables, we use LGAs dummies, quarter dummies, property type dummies instead.

[9] In total, the *data set* is imputed by five times. Thus, five imputed data sets are generated.

Table 5.3: The description of models.

Model	Model description
GBM (LS)	unmodified training sets, GBM and LS loss function.
GBM (LAD)	unmodified training sets, GBM and LAD loss function.
GBM (LS, MI)	multiple imputed training sets, GBM and LS loss function.
GBM (LAD, MI)	multiple imputed training sets, GBM and LAD loss function.
GBM (LS, C)	modified training sets with complete cases only, GBM and LS loss function.
GBM (LAD, C)	modified training sets with complete cases only, GBM and LAD loss function.
Linear (LS, MI)	multiple imputed training sets, linear function and LS loss function.
Linear (LAD, MI)	multiple imputed training sets, linear function and LAD loss function.

5.4 Empirical results

The quarterly rolling windows are implemented on the *data set*. Within *GBM*s' family, the testing data sets used are slightly different. Those applied missing node strategy and complete case strategy use the unmodified testing data sets which contain missing values[10]. The others applied imputation strategy use the imputed testing data sets by default[11]. For linear hedonic models, the missing values in the test sample are also imputed to ensure that the transaction prices can be forecasted. The overall results of the different implementations are shown respectively. Figure 5.1 depicts the percentage error range. In Figure 5.2, the error metrics through the 16 rolling windows are presented, and in Table 5.4, the overall performance of all implementations is

[10] The security mechanism of missing value node method ensures that every observation in the testing sample has a prediction even if there is no missing value in the training sets.

[11] The testing data sets are also imputed using the same imputation model for the training sets when multiple imputation strategy is applied.

summarized. Some additional results are shown in the appendix.

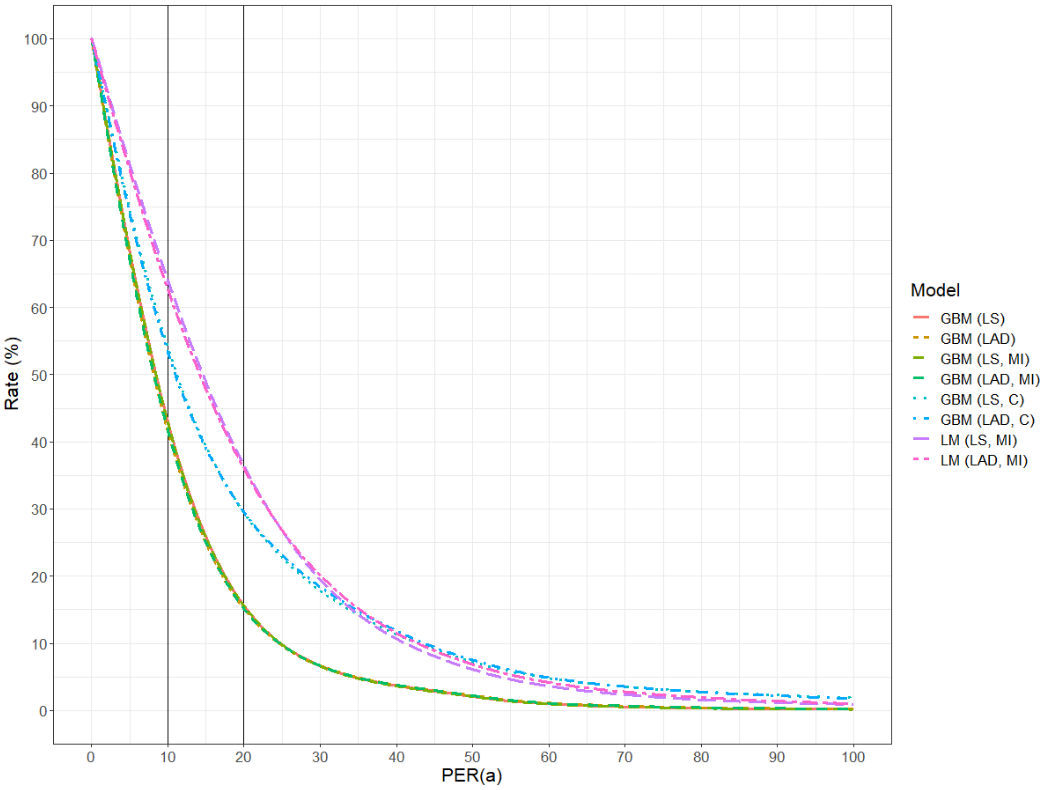

Figure 5.1: The percentage error range for each model. Two vertical lines indicate the PER(10) and PER(20). The percentage errors are truncated at 100.

The *PER* shows the proportion of percentage prediction errors that lie outside a specified range. The "L" shape of *PER* indicates that every observation in the test sample is perfectly predicted. Three aspects are worth mentioning in the overall scope. Initially, in Figure 5.1, every model depicts a convex line, which means the transaction prices of observations in the test sample are acceptably predicted. Then, most models tend to underestimate the prices of properties for the whole period (from 2017Q1 to 2020Q4), see the positive entries for the *MeanPE* and *MedianPE* in Table 5.4. However, the biases in 16 rolling windows are not always positive, as depicted in graphs A and

B of Figure 5.2. The trend of biases may be explained by the local housing market trend. Starting from 2015Q1, the Perth local residential property market experienced a gentle long-term decline by about -0.7% per quarter until 2019Q3. Then, the market prospers again by an average 1.8% increase per quarter from 2019Q4[12]. If the local market is in recession, the predicted prices tend to be overestimated or less underestimated. On the contrary, the models tend to underestimate or slightly overestimate market prices of properties in the prosperity period. Additionally, Table 5.4 shows that *MeanPE* is always larger than *MedianPE* in all six models, which means that all the residual distributions have positive skewness.

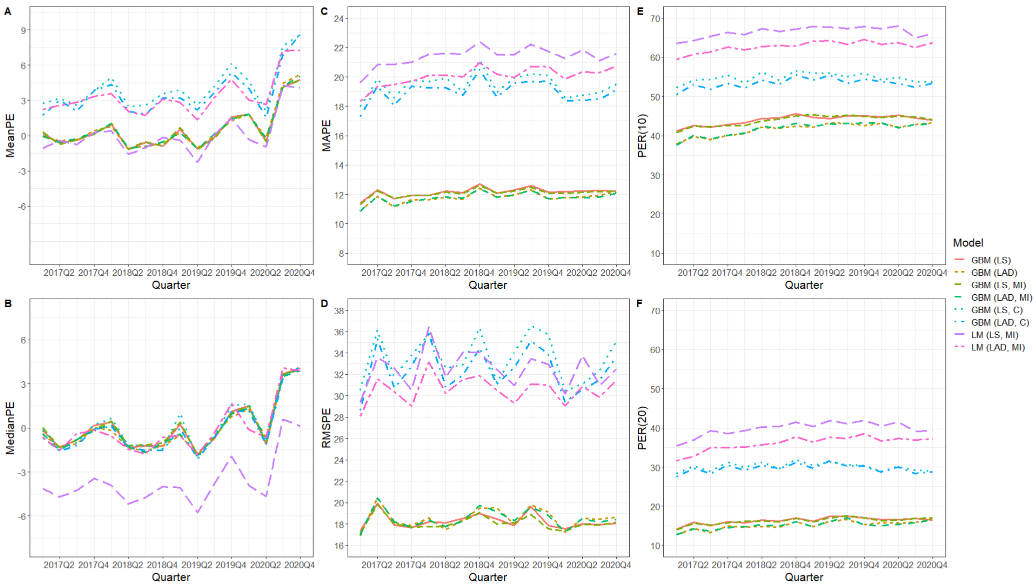

Figure 5.2: The errors for each test quarter in 2017Q1-2020Q4. depicts *MeanPE* (**A**), *MedianPE* (**B**), *MAPE* (**C**), *RMSPE* (**D**), *PER(10)* (**E**), *PER(20)* (**F**) of each model through 16 rolling windows. The mean of percentage errors is added to the comparison to comprehensively present the distribution of percentage errors.

[12] This information refers to the quarterly residential property market reports and residential property price index (*RPPI*) issued by the Australian Bureau of Statistics (*ABS*).

5.4.1 *GBM* versus *LM*

Through results shown in figures and tables, the *GBM*s outperform the linear hedonic models with the same settings in all sixteen windows (*GBM*s (*LS* & *LAD*, *MI*) versus the linear models (*LS* & *LAD*, *MI*)), this result is in line with Mayer et al. (2019), Schulz and Wersing (2021). Averagely, the *GBM*s reduce mean absolute percentage error by around 8%, around 40% improvement to the linear hedonic models. They also present around 12% less for *RMSPE*, and around 40% more accurate to the benchmarks. In Figure 5.1, the convex *PER* lines of *GBM*s are more close to the "L" shape. For *PER(10)* (*PER(20)*), there are about an average of 42% (15%) of observations that the absolute prediction errors are more than 10% (20%) of their valuations in the *GBM*s, there are about 63% (36%) for the same metrics in the linear models. These improvements could be possibly caused by the modeling algorithm (the $f(\cdot)$). Compared with a linear model, *GBM* could measure more complex relationships between predictors and the response and the interactions among predictors. Thereby, the linear model has less explanation power than *GBM*. Additionally, the partitioning process could distribute extreme values to some nodes by isolating them from the normal observations. These extreme values might be well treated within their nodes, instead of estimated with all observations. This may avoid huge accuracy drops in some extreme cases especially when compared with linear models. Thus, the predictions provided from the *GBM*s tend to show more reliable results than the linear hedonic models under the same settings. Surprisingly, *GBM*s (*LS* & *LAD*, *C*) present higher *RMSPE*s than the benchmarks. This difference may be affected by the application of missing value strategies.

Table 5.4: The predictive performance of models. summarizes the evaluation metrics for each model. Metrics are calculated using all test samples. Log-scaled price predictions are transformed back to a natural scale. All numbers are in percentage.

Model	MeanPE	MedianPE	MAPE	RMSPE	PER(10)	PER(20)
All test samples ($N = 113,104$)						
GBM (*LS*)	0.77	0.31	11.86	17.82	42.99	15.57
GBM (*LAD*)	0.84	0.18	11.77	18.52	41.60	15.14
GBM (*LS*, *MI*)	0.77	0.33	11.81	17.68	42.85	15.50
GBM (*LAD*, *MI*)	0.83	0.19	11.76	18.46	41.70	15.11
GBM (*LS*, *C*)	4.26	0.41	19.05	33.01	54.01	29.43
GBM (*LAD*, *C*)	3.84	0.12	18.95	32.44	53.28	29.52
Linear (*LS*, *MI*)	0.30	-3.17	19.84	30.19	64.03	36.42
Linear (*LAD*, *MI*)	3.58	0.14	20.04	30.58	62.69	35.98

5.4.2 Missing value strategies

Three missing value strategies are comprehensively compared within the *GBM*'s family. Overall, the *GBM*s using missing value node strategy are the recommended choice regardless of loss functions. The complete case strategy has the simplest process – the observations with missing information are discarded carelessly. However, the drawback is significant, data loss in the training sets. In addition, the low accuracy of predictions could be expected[13], when missing values occur in testing sets. In Table 5.4, the *GBM*s (*LS*, *C* and *LAD*, *C*) show roughly 7% more *MAPE*, 14% more *RMSPE* and 14% more *PER*s than the *GBM*s (*LS* and *LAD*). This reflects that the accuracy of the predictions can not be guaranteed when the trained *GBM*s encounter

[13] The worse situation is that the prediction can't be provided, for example when complete case strategy is applied on a linear model.

the missing values for the first time. The multiple imputation strategy uses complete cases and observed characteristics (omitting the target variable) to impute missing information multiple times in the training and testing sets. Compared with the *GBM*s (*LS* and *LAD*), the performance is at the same level. The differences in *MAPE*, *RMSPE*, and *PER*s are negligible, less than around 0.1%. However, this strategy requires imputation steps before and after modeling, which makes the procedure much more complex than the other two. The above two have their shortcomings, one has a simple process but low accuracy, and the other is accurate enough but complex. The *GBM*s applying missing value node strategy is more recommended. When predicting for test samples with missing values, the trained *GBM*s are deliberate and the accuracy of predictions is guaranteed. Because they have thoroughly learned missing values in the training sets. Meanwhile, imputation for test samples is not required, which makes the whole valuation process more concise.

5.4.3 Squared error loss verse absolute deviation loss

In Table 5.4, the *GBM*s using the absolute deviation loss function always have smaller *MedianPE*, and mostly the *MeanPE* is lower when the least square loss function is applied. Similarly, when *LAD* is applied, the *MAPE* is smaller; the *RMSPE* is lower when using *LS*. Even, the difference of *MAPE* and *RMSPE* among the same group of *GBM*s, that using the same training sets, is not that significant. Such that, *PER*s are preferred. When the original training data sets are applied, *GBM* using *LAD* reduces *PER(10)* by 1.4%, *PER(20)* by 0.5%. This indicates that more observations are accurately predicted when *GBM* applies *LAD*. Similarly, when the modified training datasets are used, there are more observations that their prediction errors are less than ten percent of their valuations while *LAD* is applied. In the benchmarks, the linear hedonic model using *LAD* also shows 1.3% less *PER(10)* and 0.4% less *PER(20)*. However, the linear hedonic model (*LAD*) has a slightly higher *MAPE* and *RMSPE* than the linear hedonic model (*LS*). Because there are some extremely inaccurate predictions provided by the linear hedonic model applying *LAD*. In summary, *LAD* is a slightly better choice than *LS* in the valuation for residential properties. Because it is less sensitive to extreme cases and can prevent the influence from them. This makes estimates for normal observations more accurate and robust.

In addition, *LAD* is a special case of quantile loss function. The predictions of models using *LAD* come with a 50% confidence which models using *LS* don't have. To *AVM* users who care more about risks, such as *REIT* investors, the models using *LAD* might be more recommended. Because the price predictions accurately estimate the median prices rather than the mean price. It is better to control the risk especially when the price distribution is positively skewed. This extension of loss functions could benefit some special purposes in the practice of real estate valuation, such as risk management for mortgage portfolios.

In summary, when data include missing values, *GBM* using *LAD* outperforms the other implementations in predictive accuracy. The absolute deviation loss function is an ideal alternative to the commonly used squared error loss function when robust predictions are necessary. The most important is that the missing value node strategy is more recommended when applying tree-based machine learning techniques for valuation in residential real estate.

5.4.4 *GBM*'s Interpretation

To interpret the derived estimation is always useful, which involves gaining an understanding of those particular input variables that are most influential in explaining the variation of the dependent variable. For *GBM*, in Friedman (2001), the relative importance of input variables is defined as the relative influences (I_m), of the individual input variables (X_m), on the variation of the dependent variable ($\mathbf{Y}$) over the joint input variable distribution ($\mathbf{X}$), which is formulated as

$$I_m = \left(E_{\mathbf{X}} \left[\frac{\partial \mathbf{Y}}{\partial X_m} \right]^2 \cdot var_{\mathbf{X}}[X_m] \right)^{\frac{1}{2}}. \tag{5.2}$$

A higher relative importance means a higher contribution of the input variable to explain the variation of the dependent variable. Table 5.5 summarizes the score and the rank of relative importance for all input variables via sixteen rolling windows.

In Table 5.5, two input variables are the most influential, *LGA* and *floor area*, compared with the others. One emphasizes the approximate location of the residential property, the other shows the building size. This indicates that the location and the size of one residential property could significantly affect its market value. Meanwhile, it also reflects that the buyers would like to pay more money for a residential property that has a better location and/or larger size. The following three variables, *latitude*, *longitude* and *land size*, make this more solid, those are ranked from the third to the fifth. Except for location and size, the age and the type of residential property are also important factors, ranked sixth and seventh. The age could explain the effect of depreciation. The older property is indicated to have a lower price, compared with the new property with the same or similar structure, because of the depreciation. The *property class* identifies the price gap between different property types. Compared with the location and size, these two variables are less influential to property prices. Even they could identify the price differences between "new" and "old" properties and between diverse property types.

Table 5.5: The summary of relative importance for all the input variables via sixteen windows presents the relative importance score for each input variable and its ranks. The higher score means a more influential contribution to the variation of the price for residential properties. The values in the brackets mean the times when the input variable wins the rank in the sixteen rolling windows.

Variable	Score	Rank	Variable	Score	Rank
Land size	9.155	5 (16)	Family	0.222	17 (6)
Floor area	21.668	2 (16)	Game	0.401	15 (16)
Age (year)	3.580	6 (16)	Meal	0.236	16 (5)
LGA	29.991	1 (16)	Dinning	0.142	19 (13)
Bedrooms	1.016	12 (8)	Study	0.775	13 (10)
Bathrooms	2.262	8 (16)	Car ports	1.175	10 (10)
Lounge	0.032	20 (13)	Tennis court	0.005	22 (16)
Kitchen	0.020	21 (13)	Pool	1.848	9 (16)
Tile-roof	0.655	14 (12)	Latitude	12.041	3 (16)
Brick-wall	0.218	18 (4)	Longitude	11.243	4 (16)
Property Class	2.753	7 (16)	Temporal Variable	1.097	11 (2)

Surprisingly, in Table 5.5, the housing structural factors and temporal variables are not that crucial. All housing structural variables are ranked below the eighth. It means they may have a small effect on property prices. The property prices may not be significantly changed due to there being one extra bedroom or bathroom. Unexpectedly, the temporal variable ranks low, at the eleventh. There may be a potential reason. The growth of the housing market is not strong during the research period, which means the quarterly growth rates are not significantly different from zero. Thus, the two-year length may not be enough for each rolling window to recognize the temporal effects. This reflects that, when the housing market is not that hot, there is no huge price difference in purchasing a residential property early or late.

5.4.5 Robustness

Valuers are not only interested in the accuracy of property price estimates but are also concerned about the uncertainty of valuation model performance. Because the future is uncertain, models could face variant properties that may be well predicted or may not. In industrial practice, valuers collect all information currently available to evaluate residential properties. This practical approach is simulated by only re-sampling the whole testing sample with replacement from 2017Q1 to 2020Q4. This simulation could provide different parallel "futures" but the same "past" to mimic the future uncertainty. For each replication in each window, the training sample is the same as it is in the quarterly rolling window procedure. After training, models forecast market prices for different bootstrapped new testing samples in the following quarter. 20,000 replications are conducted. The simulation procedure with rolling windows is depicted in Figure 5.3.

The distributions and boundaries of performance metrics are shown in Fig-

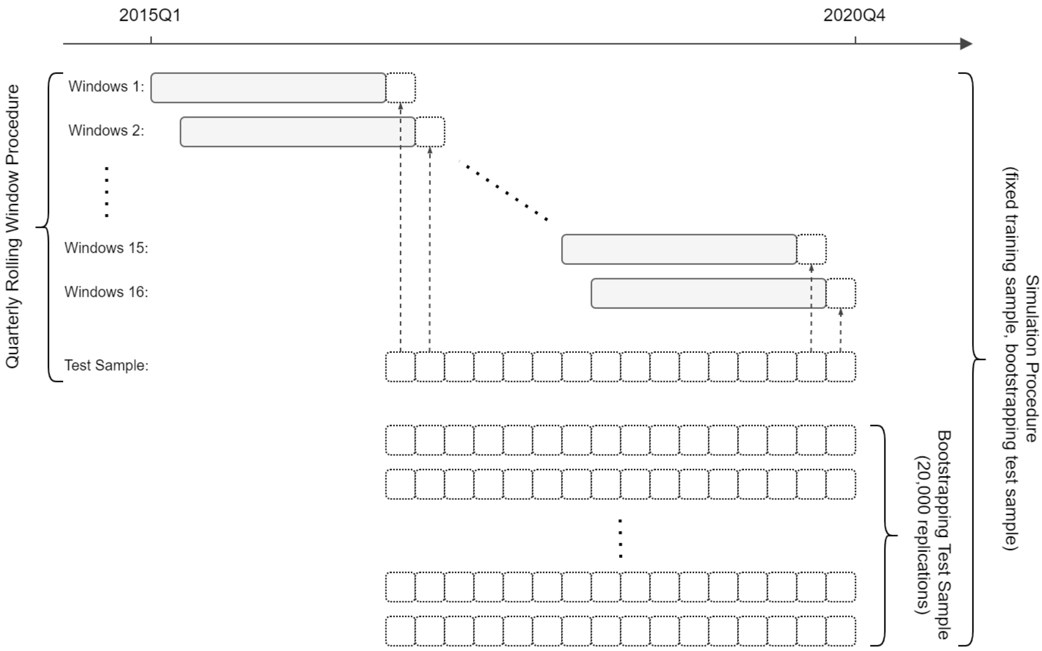
2015Q1
2020Q4
Quarterly Rolling Window Procedure
Windows 1:
Windows 2:
Windows 15:
Windows 16:
Test Sample:
Simulation Procedure
(fixed training sample, bootstrapping test sample)
Bootstrapping Test Sample
(20,000 replications)

Figure 5.3: The schematic explanations of test sample simulation procedure with rolling windows. The 16 training samples in the simulation procedure are the same as they are in the rolling window procedure. The models are done on the training sets. Then, they predict the transaction prices for observations in the bootstrapped testing samples.

ure 5.4 and Table 5.6 respectively. In Figure 5.4, the four error distributions of *GBM* (*LAD*) can't be easily distinguished from the error distributions of *GBM* (*LS*), *GBM* (*LS*, *MI*), and *GBM* (*LAD*, *MI*), there are heavily overlaps, such as in the *MAPE* distribution. The *RMSPE* distribution of *GBM* (*LAD*) is fatter and more flat-topped, this indicates that *RMSPE*s of the *GBM* using *LAD* are unstable and sensitive when future situations are uncertain. The four *GBM* implementations, *GBM* (*LS*), *GBM* (*LAD*), *GBM* (*LS*, *MI*), and *GBM* (*LAD*, *MI*), are the first tier, the rest implementations show a large gap of accuracy with them. The second tier is the *GBM*s using the complete case strategy. Most error distributions of *GBM* (*LAD*, *C*) are positioned on the left of the error distributions of *GBM* (*LS*, *C*). Even if they have larger

*RMSPE*s than the benchmarks, they are better than the benchmarks in the other error metrics.

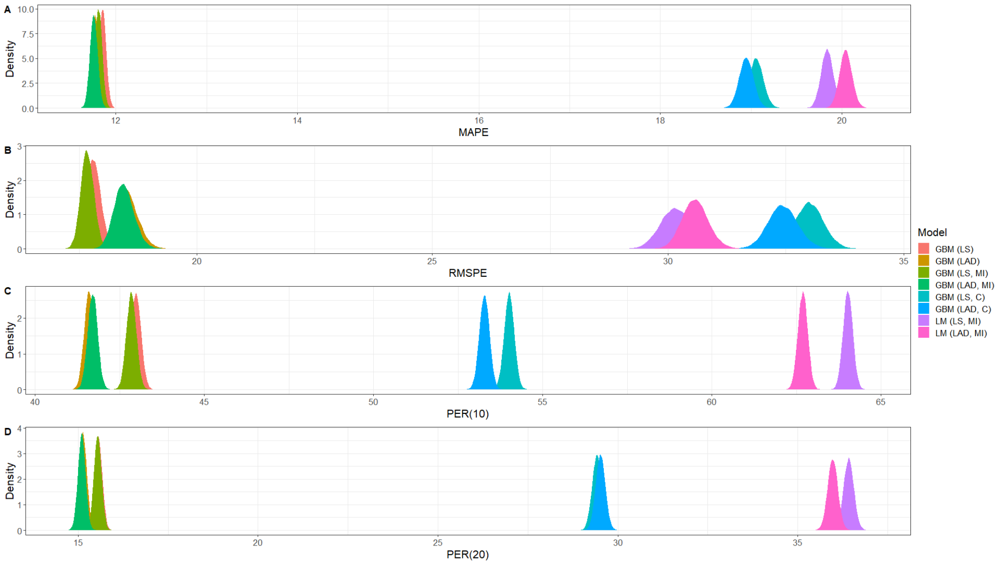

Figure 5.4: The error distributions for each model in the simulation. depicts the distributions of *MAPE* (**A**), *RMSPE* (**B**), *PER(10)* (**C**), *PER(20)* (**D**). Test samples are bootstrapped in each iteration. In total, there are 20,000 replications done in the simulation. The gap between *GBM*s and linear hedonic models is eliminated. The left graphs are for *GBM*s (*LS* and *LAD*), the graphs in the middle are for *GBM*s (*LS*, *C* and *LAD*, *C*), linear models (*LS* and *LAD*) are on the right.

Table 5.6: The intervals of evaluation criteria for models. summarizes the maximum and minimum evaluation criteria for each model in all 20,000 bootstrapped test samples. Log-scaled price predictions are transformed back to a natural scale. All numbers are in percentage.

Model	MAPE	RMSPE	PER(10)	PER(20)
All 20,000 bootstrapped test samples				
GBM (*LS*)	[11.69, 12.03]	[17.30, 18.59]	[42.44, 43.62]	[15.08, 16.07]
GBM (*LAD*)	[11.59, 11.94]	[17.82, 19.46]	[40.90, 42.24]	[14.71, 15.54]
GBM (*LS*, *MI*)	[11.65, 11.99]	[17.22, 18.32]	[42.30, 43.47]	[15.15, 15.99]
GBM (*LAD*, *MI*)	[11.57, 11.94]	[17.77, 19.37]	[41.08, 42.32]	[14.65, 15.56]
GBM (*LS*, *C*)	[18.74, 19.36]	[31.83, 34.29]	[53.49, 54.57]	[28.88, 29.91]
GBM (*LAD*, *C*)	[18.63, 19.23]	[31.34, 33.80]	[52.77, 53.99]	[29.00, 30.06]
Linear (*LS*)	[14.38, 14.80]	[28.87, 31.53]	[63.51, 64.68]	[35.85, 37.01]
Linear (*LAD*)	[13.96, 14.40]	[29.52, 31.69]	[62.11, 63.21]	[35.43, 36.55]

Table 5.6 summarizes the intervals of error metrics for the 20,000 bootstrapped test samples. There are some overlaps in the *MAPE*, *RMSPE*, and *PER(20)* intervals between the pair of models using *LAD* and *LS*. The difference between the *PER(10)* intervals is more obvious. Among models using the same loss functions, the *GBM* trained on the original training data sets is the best, with the intervals always on the left. Overall, the outputs of this simulation do not change the results in the rolling window analysis. We could have high confidence in claiming that the *GBM* using *LAD* is still the recommended implementation, no matter what future situations it will experience.

5.5 Conclusions

In this chapter, an investigation is conducted for automated valuation models for residential properties in the metropolitan area of Perth, Western Australia, when data contain missing values. *GBM*s and linear hedonic models are studied and estimated in the sixteen rolling windows, the missing value issue they may experience is dealt with using different strategies, such as the missing value node strategy and the imputation strategy. Among *GBM*s, three different model training schemes are conducted respectively, the complete-case-only scheme, the imputed data scheme, and the original data scheme. Meanwhile, *LS* and *LAD* are applied to assess the influence of the loss function on the price estimation. This case study mainly produces three important insights.

First, the *GBM*s are competent models to provide better predictive accuracy than the linear hedonic models, no matter which loss function is applied, when missing values are included in data. The reason for this improvement could be the recursive partitioning algorithm of *GBM*. The recursive partitioning process allows *GBM* to investigate more complex interactions between variables, that the linear model can not. Meanwhile, *GBM* partitions the data space into different regions, the best estimates are found in the local regions, rather than in the global.

Second, missing values may not have to be arbitrarily or experientially removed, when a method is valid to robustly handle them. Missing values are highly common in residential real estate data, due to data collection, multiple data sources, and bad data maintenance. Retaining them in data makes

*AVM*s more comprehensively estimate the prices of properties, and prevent potential sizeable data loss. As a payment, it requires that the *AVM* implementation must be robust to automatically deal with missing values. If only complete cases are used in estimation, models may provide *NA* predictions or inaccurate predictions when missing characteristics occur in the coming cases. *GBM* applying the missing value node strategy provides an option. It not only provides robust and accurate predictions but also avoids the risk of *NA* predictions for the coming cases. Thus, the missing value node strategy allows *GBM* could deal with missing values simultaneously, automatically, and robustly in estimation. However, it is unsure whether there are better alternative missing value strategies, and this leads to further investigation.

Third, the loss function is a worthwhile concern to investigate. In this chapter, the absolute deviation loss function is applied, because of its robustness to extreme values. The models applying *LAD* do show more accurate results than the models applying *LS*. *LAD* gives the same weights to all residuals, which leads to less sensitivity to extreme cases. It ensures that more normal observations are well predicted. Such that, the least squared loss function may not always be the best choice. This reminds the *AVM* builders or users that, in the future, various choices of loss functions could make *AVM* adaptable to match diverse purposes of residential real estate analysis, such as applying quantile loss for the management of mortgage default risk.

Chapter 6

Residential property price indices

6.1 Introduction

Chapter 5 has demonstrated that *GBM* provides much more accurate price predictions for residential properties than the hedonic linear model while missing values issue occurs in the transaction-level data. However, it is not investigated whether there are other applications for residential real estate analysis using machine learning models. In Hill and Scholz (2018), missing values are the main data issue. Multiple versions of the general additive model have to be estimated, each containing a different mix of housing characteristics, the characteristics containing missing values are omitted. Motivated by their paper, *GBM* is applied for price indexing purposes, and the complied price could measure the market price trend over time. Meanwhile, it deals with missing value issues. This chapter will use a machine learning model to compile the residential property price index (*RPPI*) when missing values occur in the dataset.

Three research questions will be investigated in this chapter. The first is to study whether it is achievable to compile *RPPI* using *GBM* with the classic indexing approaches. The hedonic regression method is deployed, because it is recommended in the literature in Chapter 2. Both of hedonic imputation approach and the time dummy approach are studied. The hedonic imputation approach relies on the predictions provided by the estimated regressions in each cross-section. The research question, of whether the more accurate price predictions could impact the resulting price index, will be answered in this chapter. For the time dummy approach, *GBM* needs an interpretation tool for extracting the marginal effect of the temporal variable. The index quality may depend on how well the *GBM* could be explained with respect to time.

The second is to compare the diverse indexing implementations, which are compiled using different models and different indexing approaches. The index comparison is studied in two aspects. Initially, the accuracy of rough appraisal using the indices is compared. *RPPI* should represent the aggregative price trend through the investigation period. Thus, the more precious appraisals provide, the more representative the index is. Then, the contemporaneous correlation between indices is studied. This is to examine whether the compiled indices could measure the price growth of the residential property market precisely. If two indices are contemporaneously correlated, these two indices may visualize the approximate paths of market price dynamics. Then, the difference between these two indices is negligible. In the comparison, the official *RPPI* published by *ABS* is the benchmark. Additionally, the competent alternative index could be suggested, and it is necessary while the *ABS* official *RPPI* stopped publishing in 2022.

Thirdly, the revision of indices is discussed, especially for the indices using the time dummy approach. Normally, a price index representing a long-term market trend should be revised routinely, the revision is to correct errors in the index and to update the index values due new information becomes available. If most of the index values always significantly change in each regular revision, it needs to be considered carefully when this index is deployed for measuring a long-term tendency. This is also a robustness check for indices examined in this chapter. A solid *RPPI* should be accurate and timely with no revision or minimal revisions.

6.2 Hedonic regression method

The hedonic regression method is reviewed in Chapter 2, and has two variants, one is the hedonic imputation approach, and the other is the time dummy approach. These two are briefly described in this section.

6.2.1 Hedonic imputation approach

A residential property is a composite with a bundle of essential characteristics. This bundle may include attributes of both the structure and the location. There is no market for individual characteristics since they can not be sold separately. A function of the characteristics can be used to determine the values of the properties and estimate the marginal contributions of characteristics to the value. Suppose that transaction-level data are available for all periods, $t = 1, 2, \cdots, T$, the log-price of property i sold in period t, $p_{i,t}$, is given by a function of its characteristics, such as structures and location:

$$p_{i,t} = f(\mathbf{x}_{i,t}) + \varepsilon_{i,t}, \tag{6.1}$$

where $\mathbf{x}_{i,t}$ is the vector of all available characteristics, such as $Age_{i,t}$ and $Floor_{i,t}$ are the current age and floor space in period t. The term, $\varepsilon_{i,t}$, denotes the error for property i. The estimations are done in each cross-sectional data set. Such that, no time dummies is included in Eq. 6.1. $f(\cdot)$ controls the relationships between property characteristics and transaction prices, some model structures could apply on $f(\cdot)$ for different research objectives, for example, the best-known hedonic specification, log-linear model:

$$p_{i,t} = \beta_0^t + \sum_{m=1}^{M} \beta_m^t x_{im,t} + \varepsilon_{i,t}. \tag{6.2}$$

The hedonic imputation method relies on the predictions from a hedonic model, that predicts prices over periods, which can then be inserted into a standard price index formula. To obtain a hedonic imputation price relative to comparing period t and period $t+1$, the Laspeyres type focuses on the properties sold in the earlier period t, meanwhile, the Paasche type focuses on the properties sold in the later period $t+1$. Taking a geometric mean of the Laspeyres and Paasche types of price relatives is the Törnqvist type. The Törnqvist type has an advantage in that both periods are symmetrically treated (Hill and Melser 2008). For the index compilation, the predictions of prices ($\hat{P}_{i,t}$) need to be transformed back to the natural scale. The formulas presented below use the double imputation, which means that the prices are predicted in both periods.

The Laspeyres type:

$$\mathbf{PI}_{DIL}^{t,t+1} = \prod_{i=1}^{N_t} \Big[\Big(\frac{\hat{P}_{i,t+1}(\mathbf{x}_{i,t})}{\hat{P}_{i,t}(\mathbf{x}_{i,t})} \Big)^{\frac{1}{N_t}} \Big] \tag{6.3}$$

The Paasche type:

$$\mathbf{PI}_{DIP}^{t,t+1} = \prod_{i=1}^{N_{t+1}} \Big[\Big(\frac{\hat{P}_{i,t+1}(\mathbf{x}_{i,t+1})}{\hat{P}_{i,t}(\mathbf{x}_{i,t+1})} \Big)^{\frac{1}{N_{t+1}}} \Big] \tag{6.4}$$

The Törnqvist type:

$$\mathbf{PI}_{DIT}^{t,t+1} = \sqrt{\mathbf{PI}_{DIL}^{t,t+1} \times \mathbf{PI}_{DIP}^{t,t+1}} \tag{6.5}$$

where $i = 1, \cdots, N_t$ mean the dwellings sold in period t, and $i = 1, \cdots, N_{t+1}$ present the residential properties sold in period $t + 1$. These three formulas calculate the price relatives between two consecutive periods. The aggregate price index is then compiled by chaining the relatives from these calculations.

6.2.2 Time dummy approach

Eq. 6.2 allows the characteristics parameters and functions to change over time. This is in line with the idea that housing market conditions determine the marginal contributions of the characteristics: it is not expected that those contributions are constant when demand and supply conditions change (Pakes 2003). Yet, it seems most likely that market conditions change gradually. Clapp and Giaccotto (1992) suggest that the assumption can be made that the implicit prices or functions of characteristics are constant over time, perhaps only for the short term. The market trend over time is explained by time dummies in the whole period, then:

$$p_{i,t} = \beta_0 + \sum_{t=1}^{T} \alpha_t D_t + \sum_{m=1}^{M} \beta_m x_{im,t} + \varepsilon_{i,t}. \tag{6.6}$$

D_t denotes the time dummy variable, that is 1 when the property is sold in period t, 0 in the others, α_t is its coefficient. This approach uses the α_t from the period 1 to period T to compile the index.

6.3 Data for indexing

The *data set* used for indexing is a subset of the *landgate* data from 2004 to 2020, which have been cleaned. In total, there are 8,600 non-arm-length transactions dropped. Within the period, there are 583,762 observations available in total, and 219,320 sale-resale pairs. There are some missing values shown in the housing characteristics, described in Table 6.1.

Table 6.1: The summary statistics for the sold residential properties in the dataset, 2004 - 2020.

Variable	NA num	Mean	S.D.	Variable	NA num	Mean	S.D.
Price ($,000)	0	557.894	395.851	Dining	0	0.669	0.474
Land size (m^2)	22	907.831	5,178.766	Family	0	0.570	0.500
Floor area (m^2)	239,023	147.697	72.154	Game	0	0.217	0.416
Age (year)	4,300	25.209	20.492	Meal	0	0.300	0.461
Bedrooms	4,940	3.177	0.859	Study	0	0.167	0.380
Bathrooms	1,003	1.557	0.853	Car ports	0	1.403	0.778
Lounge	4,940	1.008	0.107	Tennis court	0	0.001	0.028
Kitchen	4,940	1.006	0.090	Pool	0	0.144	0.351
Tile-roof	11,924	0.854	0.353	latitude	0	-31.969	0.164
Brick-wall	379	0.928	0.259	longitude	0	115.858	0.095

6.3.1 Train-test split

Testing samples are created, that could conduct an out-of-sample evaluation for the comparison of the indices. Nagaraja et al. (2014) divide the data into training and testing sets. The testing set contains the final sale of the properties which have been sold three or more times in the sample period. However, this separation could lead to an unbalance of stock numbers, that fewer observations are remaining in the latter investigation period. In this chapter, the data set is randomly split into the training (80%) and testing (20%) subsets. For comparing the indices, the sales of the same property observed in both training and testing subsets are used. The sale price in the training set will be updated by the compiled *RPPI*, and then, the updated price is compared with the market price of the same property in the testing set. This splitting method benefits in checking the accuracy of the estimation models and comparing the indices. The aggregative results for index comparison include the backward and forward updating. For all indices, the test sample is not involved in the index construction procedure. Thus, the assessment is fair.

6.4 Model specifications for indexing

No matter the hedonic imputation approach or the time dummy approach, all of them commonly start from a hedonic model, the price of a residential property is regressed on a bundle of characteristics (Hill 2013a, Eurostat 2013, Diewert and Shimizu 2015). In this chapter, the hedonic imputation approach is chosen since it is more flexible than either the time dummy approach (Silver and Heravi 2007). The hedonic imputation method is introduced in Section 6.2.1, and the price growth between two adjacent periods is calculated using the Törnqvist type of formula as shown in Eq. 6.5, and then the growths are chained together for compiling the aggregate *RPPI*.

Recall Eq. 6.2, one of the best-known hedonic specifications (the log-linear model, the "workhorse"), it is modified as

$$p_{i,t} = \beta_0^t + \mathbf{x}_{i,t}\boldsymbol{\beta}_{\mathbf{x}}^t + \beta_A^t Age_{i,t} + \beta_F^t Floor_{i,t} + \boldsymbol{\beta}_L^t \mathbf{L}_{i,t} + \varepsilon_{i,t}, \tag{6.7}$$

where the βs are the parameters of the characteristics in the period t. Therein, β_A^t, β_F^t, $\boldsymbol{\beta}_L^t$ capture the constant marginal effects of property age, floor size, and land separately. $\mathbf{x}_{i,t}$ includes all the other structural characteristics. However, these continuous variables affecting property values could be in a non-linear manner, especially the land contribution, $\boldsymbol{\beta}_L^t \mathbf{L}_{it}$ (including coordinates and land size). The non-linearity could be measured by Box-Cox transformation of each variable, applied in Chau et al. (2005), Shimizu, Takatsuji, Ono and Nishimura (2010).

In advance, a generalized additive model (GAM), which is well studied in the $RPPI$ field, could apply non-parametric splines to measure this non-linearity of predictors under a semi-parametric format, such as a spline surface for coordinates (Hill and Scholz 2018). The spline components included in semi-parametric hedonic models are a potentially more flexible alternative to Box-Cox transformations, such as three-dimensional spline defined over floor size, garage space, and age of the property in Bao and Wan (2004) and one-dimensional splines defined on land area and age respectively in Diewert and Shimizu (2015). If the linearity assumption is violated to the continuous variables, by applying GAM, Eq. 6.7 could be evolved to:

$$p_{i,t} = \beta_0^t + \mathbf{x}_{i,t}\boldsymbol{\beta}_{\mathbf{x}}^t + f_A^t(Age_{i,t}) + f_F^t(Floor_{i,t}) + f_L^t(\mathbf{L}_{i,t}) + \varepsilon_{i,t}. \tag{6.8}$$

In each period t, the transaction price of a residential property is evaluated by the parametric part, which measures the effect of other structural characteristics excluding age and floor space, and the non-parametric spline functions, which measures the non-linear age effect (a one-dimensional spline), floor space effect (a one-dimensional spline) and land effect (a two-dimensional spline surface for coordinates and one-dimensional spline for land size) separately. The interpret-ability of Eq. 6.8 is worse than the linear model (Eq. 6.7), because the non-parametric spline part is harder to explain than the parametric part.

Gradient boosting machine (*GBM*) is a flexible tree-based method, that could directly learn from the provided data. The model structure is different from the previously described models. It depends on each boosting tree that could recursively partition the data space. Then, the responses (p_{it}) are summarized in each non-overlapped region. The logic of *GBM* is similar to the stratification indexing method. However, the stratification rules of *GBM* are not constant, they vary in each tree and are set based on data. This is different from the stratification indexing method where the stratification rules are decided by the experts. For the cross-sectional data analysis, in each period, the model specification could be described as:

$$p_{it} = f(\mathbf{x}_{it},\ Age_{it},\ Floor_{it},\ \mathbf{L}_{it},\ \varepsilon_{it}), \tag{6.9}$$

where $f(\cdot)$ is tree structured. The flaw of *GBM* is that the estimation is not that transparent and easy to interpret as Eq. 6.7, even worse than Eq. 6.8. Accumulated local effect (ALE, Section 4.4.1) is applied to extract the marginal effect of temporal variable in Eq. 6.9 for achieving the time dummy approach.

The implementations of the residential property price index which apply different estimation models and two hedonic indexing approaches are summarized as follows:

- HI_{LM} index applies the hedonic imputation approach using a linear model.
- HI_{GAM} index applies the hedonic imputation approach using a general additive model.
- HI_{GBM} index applies the hedonic imputation approach using a gradient boosting machine.
- TD_{LM} index applies the time dummy approach using a linear model.
- TD_{GBM} index applies the time dummy approach using a gradient boosting machine.

The missing values in the data are imputed as the same procedure in Chapter 5. For the time dummy approach, the model specifications are similar to the model specifications for the hedonic imputation approach. The difference is that the time dummy approach adds time dummies or a temporal variable into models, and the model is regressed on the pooled data[1]. The TD_{GBM} index is special. This implementation includes a temporal variable, which contains continuous integer values, into *GBM*. The time effects of quarters are extracted by ALE, which is described in Chapter 4, and the time effects are explained by accumulating the marginal effects when the temporal variable increases. Such that, the values of time effects calculated by ALE in GBM are equivalent to the coefficients of time dummies in the linear model. Then, these values are used for constructing a price index for residential properties (TD_{GBM} index).

[1] The combination of *GAM* and the time dummy approach is not implemented. Because the spline requires extremely large computing resources when the number of observations is high. Meanwhile, time dummies are in the parametric part of *GAM*, where the difference may not be that significant compared with the linear model using the time dummy approach.

The residential property price index published by the Australian Bureau of Statistics (*ABS*) is the benchmark in the indices' comparison, referred to as *ABS* index. The quarterly *RPPI* (*ABS* index) available on the Australian Bureau of Statistics (*ABS*) website uses the stratification indexing method, which is compiled and maintained by *ABS*.

6.5 Empirical results

In this section, the accuracy of candidate models is examined first. After pricking the most accurate one, "whether the higher accuracy of predictions could benefit the index construction" is answered in the investigation of indices compiled by the hedonic imputation approach, because this approach highly relies on the price predictions. Then, "how well do the machine learning techniques explain the time effects on the market prices of residential properties" is studied by examining the *RPPI*s constructed by the time dummy approach.

6.5.1 Model accuracy performance

Table 6.2 presents the overall errors for the candidate models, that follow the evaluation criteria described in Section 4.5. The test samples in the quarter t are forecasted using the models estimated for the cross-section of the quarter t. The *GBM* is the most precise model that exhibits the smallest errors in all metrics except *MedianPE*. There are approximately 44.3% (17.6%) of out-of-sample predictions those errors are larger than 10% (20%) of their valuations.

In addition, it presents around 4.5 better *MAPE* and 1.6 better *RMSPE* than the *GAM*. *GAM* wins the silver medal. It clearly shows better results than the linear hedonic model (*LM*), such as an average of around 11.8% less in *PER*s. But, comparing with *GBM*, it shows an average of around 2.7% more in *PER*s. The least accurate model is the linear one, which uses the simplest model structure. The performance gaps are relatively significant compared with the other two models.

Table 6.2: The Accuracy of the index construction models. summarizes the combined percentage errors for each model under a strict cross-section process, those are used to compile hedonic imputation index. The performance rankings follow the rule, the lower the error value the model has, the more accurate and better it is. All numbers are in percentage.

Model	MeanPE	MedianPE	MAPE	RMSPE	PER(10)	PER(20)
LM	3.61	-0.07	18.70	26.40	59.32	31.40
GAM	5.54	0.34	17.01	21.38	46.86	20.33
GBM	1.95	0.79	12.73	19.71	44.26	17.56

Figure 6.1 depicts the *MAPE* (plot A) and *RMSPE* (plot B) for the candidate models in each quarter through the investigation. In the plot A, the *MAPE* lines are stratified. The *MAPE* for *LM* is consistently larger than the *MAPE*s for *GAM* and *GBM*. The gap between the *MAPE* lines of *GAM* and *GBM* is relatively narrow. In most quarters, the *MAPE* for *GBM* is smaller than the *MAPE* for *GAM*. In plot B, all the *RMSPE* lines are volatile. Therein, the linear hedonic model almost depicts the highest *RMSPE* through the investigation period. In 2004Q2, 2011Q4, and 2014Q1, the *RMSPE*s for *GAM* are rocket-up. In the rest periods, the *RMSPE*s for both *GAM* and *GBM* are closed. This reflects that the *RMSPE* line of *GAM* is slightly more fluctuated and sensitive than the *GBM*'s *RMSPE* line. These two plots in

Figure 6.1 ensure the performance rank of models through the investigation period, which are shown in Table 6.2.

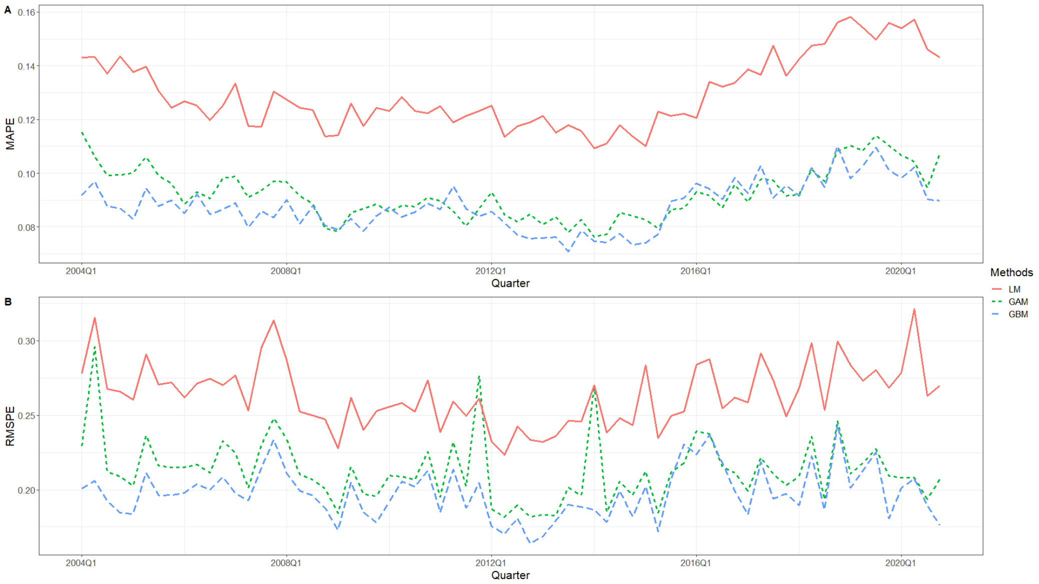

Figure 6.1: The Plot of quarterly *MAPE* and *RMSPE*. shows the *MAPE* and *RMSPE* of each candidate model in each quarter from 2004Q1 to 2020Q4.

In summary, *GBM* provides the most precise predictions, *GAM* takes the second place, and *LM* is the last. The predictions are necessary for the hedonic imputation approach. It has been demonstrated that the *GBM* provides the most accurate predictions from 2004 to 2020. Then, “whether the higher accuracy of predictions could benefit the index construction” is investigated in the following sections.

6.5.2 The comparison of indices

At first, the indices are compared graphically in two groups with respect to two approaches. The benchmark is the *ABS RPPI*. Then, the contemporaneous correlation is examined. Because the *RPPI*s measure the market dynamics equivalently, if the market growths are significantly correlated. In addition, the revision issue is discussed for the indices constructed by the time dummy approach.

Graphical comparison

In general, it can be seen from Figure 6.2 and Figure 6.3 that the price patterns are similar even if different indexing methods are used. Before around 2008, the local housing market in Perth had experienced a market rocket-up. The increasing rate is about 90%. Because the global economy was growing dramatically, especially in the real estate sector. Between 2008 and 2015, it went through ups and downs due to, such as the global financial crisis (2008-2009). After 2015, the market suffered a long-term gentle recession until around 2020. In these two periods, the prices are highly correlated with the local pillar industry, the mining industry. When the mining industry is hot, more inter-state or international immigration drives a higher demand for residential properties, which leads a higher prices. Otherwise, the price goes to the opposite. Apparently, in 2020, the price climbed again which may be caused by Covid-related effects. During the pandemic, the lockdowns cut off the supply of residential properties. Because all property constructions are not allowed, there is no new residential property joining the local market. Then, the price intends to increase due to the shortage of residential property supply.

The indices applying two different indexing approaches are compared respectively below. Figure 6.2 depicts the $RPPI$s using the hedonic imputation approach. The indices are constructed applying the candidate models – the linear hedonic model (LM), the generalized additive model (GAM), and the gradient boosting machine (GBM) – on the quarterly strict cross-sectional data sets. The price relatives between adjacent quarters are used to construct the index numbers applying Eq. 6.5. An index is compiled by chaining the price relatives. The first quarter of 2004 is the start and the reference number is set to 100. As a whole, all four indices depict a similar aggregative market trend spanning 17 years. However, after around 2008, the HI_{LM} index tends to underestimate the market trend. It diverges with the other three from 2008. Except for the HI_{LM} index, all indices almost overlap when the market is in prosperous periods (such as before 2008) and recession periods (such as after 2015). The HI_{GAM} index and HI_{GBM} index are shown higher index numbers than ABS index around 2015.

Figure 6.3 depicts the $RPPI$s using the time dummy approach. TD_{LM} index is compiled using the coefficients of time dummies in the log-linear model. The coefficients need to be transformed back to the natural scale. TD_{GBM} index requires calculating the marginal effects of the temporal variable via ALE (Eq. 4.11). The marginal effects also need a transformation. The reference period is the first quarter of 2004, and the reference index value is 100.

All three indices depicted in Figure 6.3 show a similar price trend through

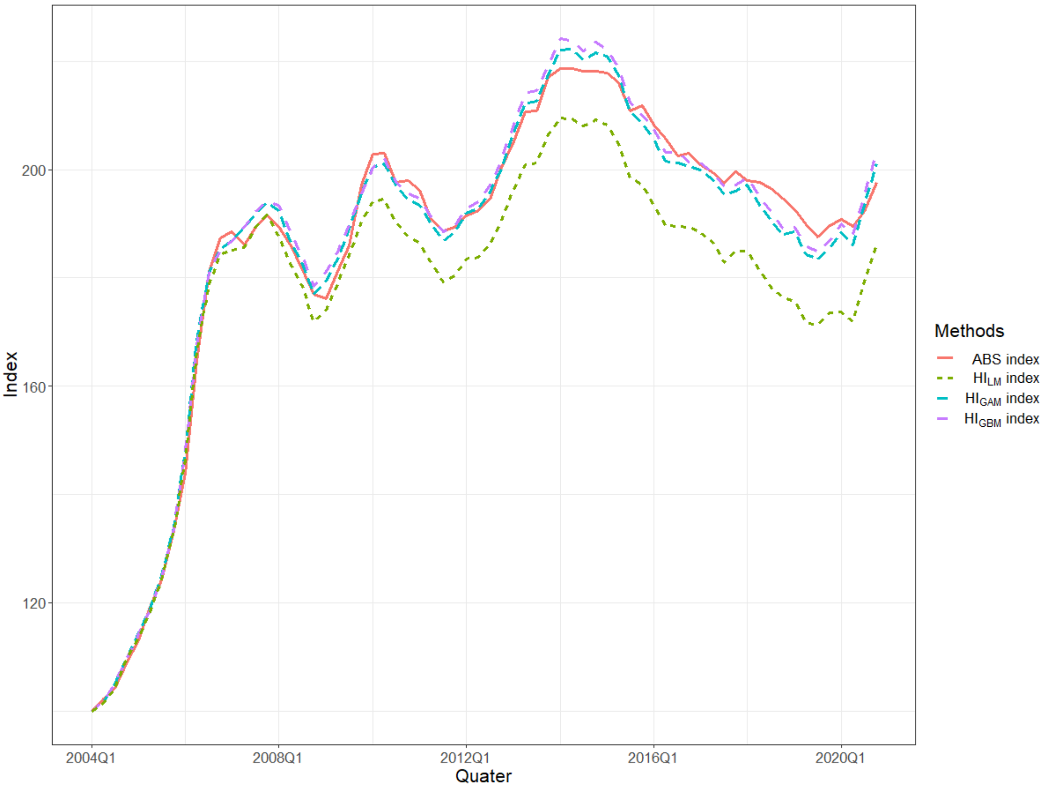

Figure 6.2: The Plot of quarterly indices using hedonic imputation approach. shows the candidate quarterly indices from 2004Q1 to 2020Q4. The reference period is the first quarter of 2004. For *ABS* index, it is updated under the new reference period setting, the original reference period is the fiscal year 2011-2012.

the period. In the prosperous period (before 2008), these three indices overlapped closely. However, after 2010, the difference is gradually emerging. Both the TD_{GBM} index and the TD_{LM} index are below the *ABS* index, but the TD_{GBM} index is closer to the benchmark. After 2015, the gaps between these three indices are growing until 2020.

From the visualizations of indices, it is difficult to suggest an indexing implementation outperforms the other indexing implementations using different indexing approaches or using various regression models. The comparison of

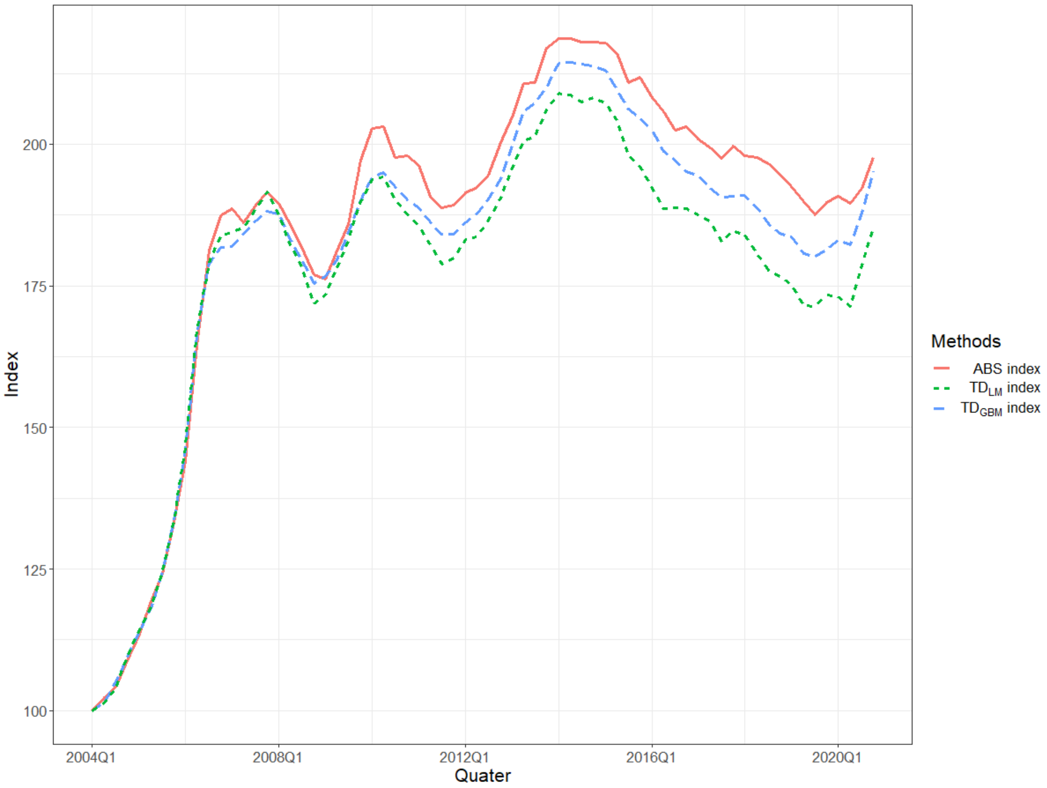

Figure 6.3: The Plot of quarterly indices using timing dummy approach. shows the candidate quarterly indices from 2004Q1 to 2020Q4. The reference period is the first quarter of 2004. For the *ABS* index, it is updated under the new reference period setting, the original reference period is the fiscal year 2011-2012.

these indices is still not complete. The visualization, however, intuitively reflects that the hedonic imputation approach and time dummy approach using GBM (HI_{GBM} index and TD_{GBM} index) can competently track the housing market dynamics, and are commensurate with the ABS index.

City-wide index accuracy

In the transaction-level data, the market prices of residential properties are not always available in each cross-section of time points. Thus, sometimes, *RPPI* is used to roughly estimate the market prices. For evaluating the accuracy of each index, this is mimicked. The repeat sales are paired, $P_i^{'}$ is the sale price of the property i observed in the training set, and P_i is the paired sale price of the same property observed in the testing set, which is assumed to be unknown. The price estimate of P_i is calculated:

$$\hat{P}_i = \frac{P_i^{'}}{\mathbf{PI}^{'}} \times \mathbf{PI},\ i = 1, \cdots, N_{test} \tag{6.10}$$

$\mathbf{PI}^{'}$ is the index number for the quarter that $P^{'}$ is observed. $\mathbf{PI}$ is the index number for the target quarter. The accuracy of the candidate indices is assessed. The percentage errors (e_i) are calculated as follows, the evaluation criteria are introduced in Section 4.5.

$$e_i = \frac{P_i - \hat{P}_i}{\hat{P}_i} \times 100,\ i = 1, \cdots, N_{test}. \tag{6.11}$$

The accuracy of city-wide *RPPI*s is evaluated using all available paired repeat sales in the training set and the testing set. The accuracy results are shown in Table 6.3.

Table 6.3: The evaluation metrics of city-wide quarterly indices. summarizes the evaluation metrics for each index. The performance rankings follow the rule, the lower error value the index has, the more accurate and better it is. All numbers are in percentage.

Index method	MeanPE	MedianPE	MedianAPE	RMSPE	PER(10)	PER(20)
ABS index	2.93	0.18	8.56	31.26	43.64	17.97
HI_{LM} index	6.70	3.31	8.88	32.62	45.26	19.51
HI_{GAM} index	3.55	0.59	14.55	31.31	43.45	17.77
HI_{GBM} index	3.09	0.21	8.48	31.13	43.39	17.70
TD_{LM} index	6.89	3.47	8.90	32.71	45.50	19.65
TD_{GBM} index	4.62	1.53	8.51	31.64	43.57	18.20

Compared with *ABS* index, *LM* indices, including the HI_{LM} index and the TD_{LM} index, present a worse performance. They show the largest *MedianPE*, which is over 3.3%, which means that more than half of price estimates tend to be overestimated. In addition, it also exhibits the largest values in the other error measures, such as *MeanPE* and *RMSPE*. Around 45.3% (19.5%) of price estimates those errors are larger than 10% (20%) of their predicted prices. Even though they are the least, the accuracy of the *LM* indices is acceptable. Meanwhile, the gap in the accuracy of compiled indices is not as significant as the gap in the accuracy of estimation models for predicting market prices.

Comprehensively considering all error measures, both of HI_{GBM} index and HI_{GAM} index deploying the hedonic imputation approach outperformed the benchmark. However, the improvement is not that significant. The *MedianPE* of HI_{GBM} index is positive, and smaller than the *MedianPE* of *LM* indices around 3%. That indicates more price estimates that are slightly overestimated than the number of price estimates that are underestimated. In addition, compared with the others, the HI_{GBM} index also shows lower error rates. Such as for *PER(10)* (*PER(20)*), HI_{GBM} index provides the smallest value, 43.49% (17.70%). Compared with the TD_{GBM} index, around 0.2% (0.5%) more appraisals using HI_{GBM} are more accurate. However, its *MeanPE* is slightly higher than the *MeanPE* of the benchmark, around 0.16. HI_{GAM} index could also be an alternative for the benchmark. It also has slightly small *PER*s, roughly an average of 0.2%. However, its *MeanPE*, *MedianPE*, *MAPE*, and *RMSPE* are all slightly larger than the metrics of the benchmark. Especially, the HI_{GAM} index shows the highest *MAPE* compared with the other competitors. This may indicate that the price estimates of the HI_{GAM} index have some extremely wrong appraisals. The TD_{GBM} index is also in the second tier with the HI_{GAM} index. However, the TD_{GBM} index is slightly deficient compared with the benchmark on most error metrics.

Comparison between the indices, *LM* indices are worse than the *GAM* and *GBM* indices. It indicates that the more accurate price predictions do benefit the accuracy of indices. However, the improvement in the accuracy of appraisals using indices is not as significant as the improvement in the accuracy of price predictions forecasted from models directly. The gap between the HI_{GAM} index and the HI_{GBM} index is relatively small, even *GBM* provides more accurate predictions. The difference between these two indices is graphically and numerically negligible, shown in Figure 6.2 and Table 6.3. In addition, calculating average marginal effects is a solution to interpret the time effects. It is proved that the machine learning technique could explain the time effect well, as TD_{GBM} is a competent alternative to the *ABS RPPI*.

Contemporaneous correlation between indices

RPPI is a key indicator that records the temporal price dynamics. Besides the accuracy of price estimates, the five indices compiled by the hedonic imputation approach and time dummy approach are compared with the benchmark in terms of their quarterly growth rates. The quarterly growth rates are calculated from each index, then, they are pairwise plotted and regressed for examining the contemporaneous correlation between indices (Shimizu, Nishimura and Watanabe 2010). The quarterly growth rates of one of the five indices say index Y, are regressed on the quarterly growth rates of another index, say index X, to obtain a simple linear relationship with a constant,

$$g_y = a + b \times g_X. \tag{6.12}$$

The hypothesis is that, if two indices equivalently measure the growth rates of the housing market, the coefficient, a, should be equal to 0, and the coefficient, b, should be 1. The coefficients of each regression and the p-value of each hypothesis test are summarized in Table 6.4.

In Table 6.4, all intercepts of growth rate regressions are not statistically and significantly different from zero, the slopes are all close to one when the significant level is 0.05. However, the hypothesis is needed to test whether indices are equivalent. The p-value of the hypothesis test for the TD_{GBM} index and the ABS index is rejected because it is the only one that is less than 0.05. It indicates that the growth rates of market prices measured by these two indices are diverse, and the path of the TD_{GBM} index might be different from the path of the ABS index through the investigation period.

Table 6.4: The summary coefficients for each regression and p-values for the hypothesis tests. These values are for investigating the contemporaneous correlation between the compiled indices and the benchmark. The asterisk indicates the default hypothesis test of regressions at 0.05 significant level.

Regression	Coef.a	Coef.b	p-value
HI_{LM} index vs ABS index	-0.0004	0.9554	0.3842
HI_{GAM} index vs ABS index	0.0010	0.9335	0.2984
HI_{GBM} index vs ABS index	0.0013	0.9180	0.1671
TD_{LM} index vs ABS index	-0.0005	0.9516	0.3373
TD_{GBM} index vs ABS index	0.0010	0.8802	0.0100

Figure 6.4 illustrates the pairwise comparisons among indices. The horizontal axis represents the quarterly growth rates of the benchmark index, the ABS index, while the vertical axis is the growth rates of the target index, for instance, the HI_{LM} index in the plot A. Most dots in the plots are close to the 45-degree line, implying that these indices are highly correlated with each other. The coefficients of correlation between the five indices and the benchmark are all above 0.93. Comprehensively considering Figure 6.4 and Table 6.4, the growth rates of five indices are statistically significantly correlated. Except for the TD_{GBM} index, the other four indices are equivalent when they measure the quarterly market growth.

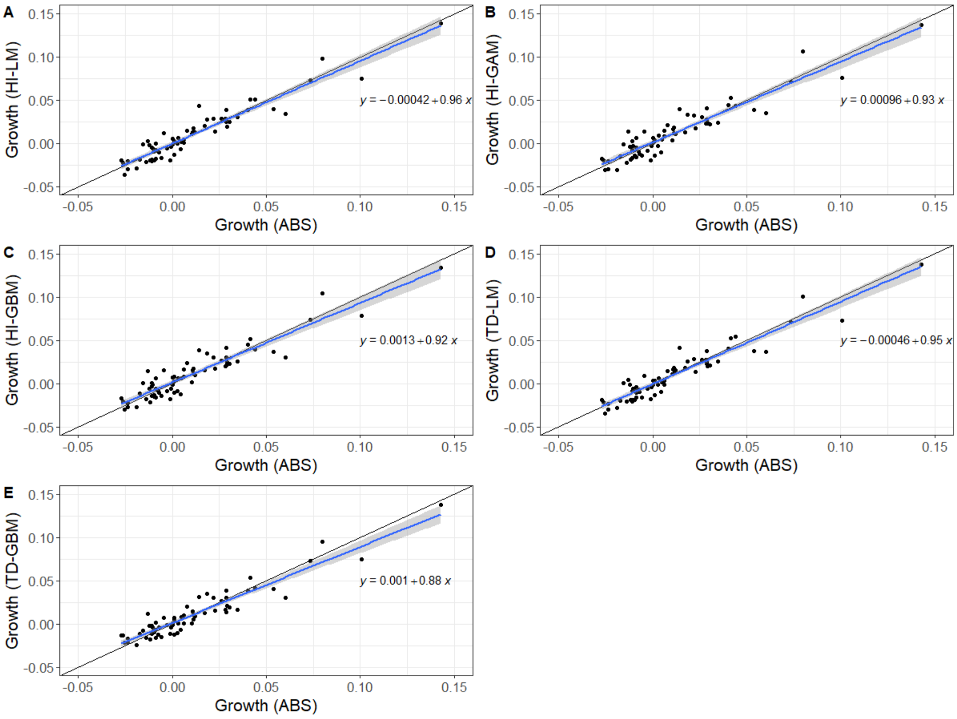

Figure 6.4: Comparison of the indices in terms of the quarterly growth rate. shows the pairwise comparisons among the four indices. In each plot, the black solid line is the 45-degree line ($y = x$); the blue line with ribbon is the smooth fitted line ($y = ax + b$) with its confidence intervals.

Index revision

Index revision evaluates the amount that previous index values revise when new available transactions are added. It is a necessarily considered criterion for *RPPI* which is continuously updated over time. Significant or systematic revisions of prior index values can be problematic, the *RPPI* could provide less in terms of social benefit (Deng and Quigley 2008). To indices used for rough appraisal purposes, systematic revisions, for example, sustained downward or upward adjustments over time, lead to biased estimates of property prices. To indices used as a macroeconomic indicator, significant revisions could result in an inaccurate inflation target, and mislead monetary policy. The reliable $RPPI$ should have fewer and insignificant revisions after its initial announcement.

The indices applying the hedonic imputation approach are built using cross-sectional data. Thus, they do not need to revise inherently except for systematic revisions. The new hedonic regression is estimated using the data in the new cross-section, and the index number is chained with the previous index values. On the contrary, the indices using the time dummy approach, such as the TD_{GBM} index, have to be revised when the new data join the pool. The revisions of indices applying the time dummy approach, TD_{LM} index and TD_{GBM} index, are examined.

The extending window is applied to mimic the index updating procedure for investigating the revision of each index (Clapham et al. 2006). The estimate of the price level in quarter q using information from the quarter 1 to the "current" quarter Q is defined as $\mathbf{PI}(q, 1, Q)$, where $Q \geq q$, $q = 1, 2, \cdots, 68$. Revision is the process as the estimated price levels evolve from the initial estimates $\mathbf{PI}(q, 1, Q')$ to the current estimates $\mathbf{PI}(q, 1, Q)$, starting from when the "first" index was compiled in quarter Q'. For instance, the initially announced index is built with 36 quarters (2004Q1–2012Q4) using each index construction method, $\mathbf{PI}(1, 1, 36), \cdots, \mathbf{PI}(36, 1, 36)$. Then, the pseudo-fresh data from the 37th quarter (2013Q1) are appended, and the index is recalculated, $\mathbf{PI}(1, 1, 37), \cdots$, $\mathbf{PI}(36, 1, 37), \mathbf{PI}(37, 1, 37)$. The revisions for quarters 1 to 36 are measured. The same is done for the 38th quarter (2013Q2, measuring revisions for quarters 1 to 37) and up through the 68th quarter (2020Q4). There is no revision number for the last quarter as it is only calculated once.

Then, the comparable revision paths that are $\mathbf{PI}(q,1,q),\cdots,\mathbf{PI}(q,1,67)$ are studied for the quarter q. The consecutive quarter-by-quarter percentage revision for quarter q at quarter Q is:

$$\Delta_q^{Q-1,Q} = 100 \times \Big(\frac{\mathbf{PI}(q,1,Q)}{\mathbf{PI}(q,1,Q-1)} - 1\Big) \tag{6.13}$$

The average percentage revision for quarter q from initial to current estimates is:

$$\Delta_q = \frac{1}{68-q}\sum_{i=q+1}^{68} 100 \times \Big(\frac{\mathbf{PI}(q,1,i)}{\mathbf{PI}(q,1,i-1)} - 1\Big) \tag{6.14}$$

For example, the revision for the 66th quarter is calculated. An index is constructed using the first 66 quarters length of data. The index number of the 66th quarter in the index is $\mathbf{PI}(66,1,66)$. Then, the new data for the 67th quarter is received, the entire index is recalculated and the new index number for the 66th quarter is revised to $\mathbf{PI}(66,1,67)$ due to changes in the models as a result of the additional observations. The difference($\Delta_{66}^{66,67}$) is the percentage difference between $\mathbf{PI}(66,1,66)$ and $\mathbf{PI}(66,1,67)$. When the data in the 68th quarter are available, the index is compiled again. The index number for the 66th quarter calculated using 67 quarters length of data is the latest one. The index number for the 66th quarter computed using 68 quarters length of data is the updated one. The new difference ($\Delta_{66}^{67,68}$) is computed. Such that, overall revision is measured as the mean of the revisions (Δ_q) for the quarter q, as shown in the previous equation.

The statistics in Table 6.5 focus on the panel of quarter-by-quarter index revisions ($\Delta_a^{Q-1,Q}$). The quarterly revision is generally small, the changes of all revised index values are lower than 1% of the previous index values. Thus, generally, the revised index will not significantly change the previously com-

Table 6.5: The index revision: quarter-by-quarter. summarizes the overall revisions, early revisions, and late revisions of indices. “Early” is defined as the newest eight revised index values in each revision window for each index; “late” as the remaining index values. Revision is measured as the percentage changes between the revised and non-updated index values. All numbers shown are in percentage.

Index method	All revisions				Early revisions				Late revisions			
	Min	Max	Mean	*S.D.*	Min	Max	Mean	*S.D.*	Min	Max	Mean	*S.D.*
TD_{LM} index	-0.0400	0.0463	0.0052	0.0091	-0.0368	0.0381	0.0092	0.0140	-0.0400	0.0465	0.0044	0.0075
TD_{GBM} index	-0.5755	0.7143	-0.0013	0.1585	-0.5046	0.5935	-0.0024	0.1685	-0.5755	0.7143	-0.0011	0.1548

piled index. Additionally, Table 6.5 summarizes the statistics of the “early” and “late” revisions respectively, the changes of the most recent eight revised index values, and the remaining. For the recent periods, the index numbers could experience higher revision changes. On the contrary, for the “older” quarters, the difference between previous index values and revised values is relatively small. The reason may be that the fresh coming-in data may have less influence on the “older” past but are more correlated with the index values of the recent periods.

In addition, for each index in Table 6.5, the revision shows diverse characters. Comparing with TD_{LM} index, TD_{GBM} index shows the negative means in Table 6.5. This presents that the TD_{GBM} index commonly has a downward revision. Meanwhile, the TD_{GBM} index shows larger standard deviations of percentage revisions than the TD_{LM} index. It could inform that the TD_{GBM} index may potentially have a more significant update. Thus, the TD_{GBM} index may suffer a more serious revision issue than the TD_{LM} index. However, the percentage difference is less than 1%, which is, somehow, not that notable between the revised index and the prior index. The standard deviations also reflect a wide variation between the revision paths, and a substantially wider revision spread could be depicted in Figure 6.5 for the TD_{GBM} index than

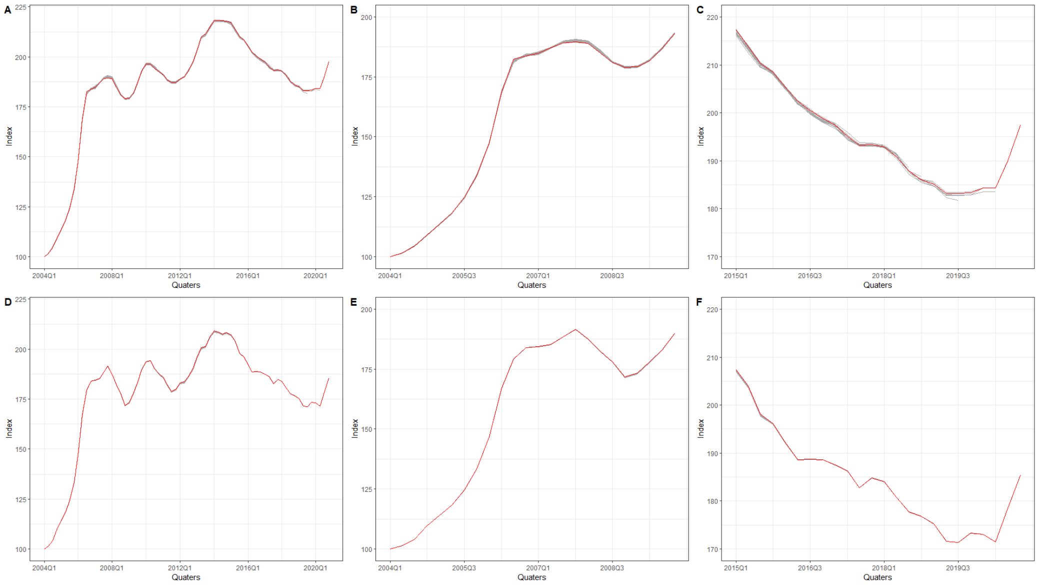

Figure 6.5: The revision paths for the TD_{GBM} index and the TD_{LM} index. shows the revision paths for two indices using the time dummy approach. The three plots of the upper panel are for the TD_{GBM} index. The red line is the index using all samples, the most recent revised index. The lower panel is for the TD_{LM} index. To depict the difference, plot B and plot E depict the period trunked from 2004Q1 to 2010Q4 (24 quarters, shown for the TD_{GBM} index and the TD_{LM} index respectively); plot C and plot F depict the later period trunked from 2014Q1 to 2020Q4 (24 quarters, shown for the TD_{GBM} index and the TD_{LM} index respectively). The grey lines are the previous indices from the previous period windows.

the TD_{LM} index. To the average revision for each quarter (Δ_q), the top four revised quarters are the 60th, 56th, 58th, and 64th quarters for the TD_{LM} index and the 66th, 63rd, 67th and 64th quarters for the TD_{GBM} index. In plot C of Figure 6.5, the area covered by the grey lines is wider than the area shown in plot F. It is proved that the index values for the recent periods may not accurately represent the trend of price dynamics, and potentially need a revision in the future periods. Overall, the revision of the TD_{GBM} index may generally have a downward tendency. The revision issue should be carefully considered when the time dummy approach is applied.

6.6 Conclusions

This chapter has conducted empirical analysis for five different residential property price indices, the implementations combine two hedonic indexing approaches, the hedonic imputation approach and the time dummy approach, and three representatives for different types of models, the parametric model (linear model), the semi-parametric model (generalized additive model), and the non-parametric model (gradient boosting machine). Several targets are to be achieved. The first is to conduct an empirical trial for machine learning techniques used in the indexing field. Next, the investigation is to examine whether the more accurate predictions could benefit the accuracy of indices, and how well the machine learning technique explains the time effects. Then, the contemporaneous correlation between indices and the revision issue of indices are discussed.

In the examination of five indices, the HI_{GBM} index shows the highest accuracy. Firstly, it provides the lowest error, when it is used for appraisal. Because it could accurately represent the price dynamics throughout the investigation period. Then, it does not need to be revised regularly. It allows the future index values to be appended to the current index by chaining the quarter-to-quarter price relatives. The other GBM implementation using the time dummy approach is not as good as the HI_{GBM} index. The price tendency is not that accurately measured. In addition, this index may suffer a serious revision issue. Overall, it is pursuable that the machine learning model could be applied for indexing purposes, applying the hedonic imputation approach and the time dummy approach.

The examination in this study also shows that more accurate predictions could not significantly benefit the accuracy of indices when the hedonic imputation approach is applied. The improvement in indexing is not as significant as the improvement in the accuracy of price predictions when applying diverse models. The results present that GBM provides the most accurate price predictions, GAM closely follows, and LM is the worst. This is the same as the results shown in Chapter 5. The accuracy is examined by using different error metrics based on out-of-sample forecasts. Intuitively, because of the accurate predictions, the HI_{GBM} index should show much lower errors to rough appraisals of residential properties. However, the improvement is not that notable. The accuracy of the HI_{GAM} index is similar to the accuracy of the HI_{GBM} index, and both of them are slightly better than the HI_{LM}. Thus, higher accuracy of price predictions may affect the accuracy of the index complied using the hedonic imputation approach. However, the improvement is limited. This finding is similar to the results in Hill and Scholz (2018). Additionally, comparing with the ABS index, both of HI_{GAM} index and the HI_{GBM} index are competent alternatives.

The contemporaneous correlation between indices is investigated, and quarterly growth rates are pairwise compared. All the correlation coefficients between indices are around or above 0.9. Four of five indices are equivalent when they measure the quarterly growth of market prices except for the TD_{GBM} index. This is reflected in that the dots of the paired quarterly growth rates are close to the 45-degree line. The F-test is employed to confirm the hypothesis that the slope of the smooth line is equal to 1 and the intercept is 0. At 0.05 significant level, only the test for the TD_{GBM} index is rejected. It means the TD_{GBM} index may present a different quarter-to-quarter growth path from theABS index. Between the other indices, the difference in quarterly growths is negligible, which confirms the results by Shimizu, Nishimura and Watanabe (2010).

For index revision, it is still an issue for indices compiled by the time dummy approach. The TD_{GBM} index is worse than the TD_{LM} index because it shows larger average revisions of index numbers. The TD_{GBM} index also has a downward revision tendency, and the revised index values may be different from the previous index value, around 1%. This proves GBM is empirically pursuable for indexing purposes, however, the drawbacks of time dummy approaches can not be overcome.

Chapter 7

Vacant land valuation and timing of development

7.1 Introduction

Before valuation and index compilation, the origin of residential dwellings starts from developing vacant lands, that are also relative to urban planning, such as metropolitan area expansion and old suburbs restructuring. Land development for a single parcel is a term project with three components, the input, the output, and the “effort” required. Commonly, the output of the project is a residential dwelling that is well studied in Chapter 5 and Chapter 6. Thus, this chapter intends to study the rest two components. The input is vacant land waiting for development, and the effort required is the construction process put on the vacant land. To study the first component, the valuation for vacant lands is conducted, which does not only evaluate the inherent characteristics of lands, size, and location but also includes some indices about the local housing market. Meanwhile, the “term” is assessed for investigating the factors that would inspire or delay the construction process.

Theoretical work in the past literature, for example McDonald and Siegel (1986), suggests that the construction project of residential properties is to exercise a real option of vacant land. Like a financial option, a real option allows one to pursue an investment at a future time point. If the future is uncertain and investment is durable and illiquid, the ability to make a different investment (or not to invest at all) in the future has an economic value, which is often referred to as a real option. By applying real option theory, land development could be considered an investment project. When landowners invest construction costs on their owned vacant land, the return on this investment is residential dwellings. Simply, by spending the construction cost (C), the landowner receives a dwelling with a value of P. The profit of this development project is therefore $P - C$. As both variables are random and growing at different rates, the landowner should wait until the profit is the largest, taking discounting into account. At this moment, the land value is represented as the option value and is affected by the contributors on the investment output side and the investment cost side.

Among the contributors, the investment cost side is paid more attention in this chapter. According to the *ABS* online article (Australian Bureau of Statistics 2020a), the final cost of construction can differ from initial expectations at the building approval stage and the start of construction when building a new dwelling. Almost half of newly constructed houses (43.8% of newly constructed houses) cost more to build than they were approved for, while 25% of newly built houses cost less in Western Australia. Only 31.2% of constructions are with no cost-changes. A similar situation also happens in the other property types, for example, townhouses. Thus, this chapter is motivated to consider the initial construction cost, its growth, and its uncertainty in the analysis, as they reflect what all the development projects could confront on the investment cost side.

For evaluating the option value (land value actually), the machine learning model, Gradient Boosting Machine (*GBM*), is employed. The model is well performed to evaluate the prices for residential properties in Chapter 5. It could be competent when assessing the vacant land value and optimal timing. However, this model is relatively hard to interpret, the effects of variables are hidden. Consequently, the accumulated local effect (*ALE*) introduced in Chapter 4 and applied in Chapter 6 could overcome this shortcoming by calculating the average marginal effects. Similar to financial options, the optimal timing of development is equivalent to finding the time when the profit $(P-C)$ reaches the maximum. At that moment, if the value of vacant land (L) is less or equal to the profit, the landowner decides to develop the vacant land, which means the vacant land is "dead". Otherwise, it is survived. Thus, survival analysis is appropriate to apply for analyzing the effects of the factors that could affect the timing of development. For the consistency of analysis tools for the option valuation and the timing optimization, *GBM* is also applied for survival analysis.

This chapter focuses on land development, which could be divided into two targets, land valuation and timing of development. The foundation of this analysis is real option theory that includes all three components of land development, the input (vacant lands), the effort required (construction cost), and the output (residential dwellings). To clarify land development, this chapter presents the theory of real options first, followed by an empirical analysis of land valuation and optimal timing. The next section (Section 7.2) introduces and explains the theory in detail. The data section (Section 7.3) summarizes the data sets selected and the variables generated. The analysis design and the empirical results about the land valuation and timing of development are described in Section 7.4 and Section 7.5 separately. The chapter concludes with a summary of the analysis.

7.2 The theory of real options

Different from financial options, which have financial assets as underlying, real options have real assets as underlying. Otherwise, the theories of these two types of options are very similar. Vacant land is such a real option where the landowner will receive a building if the land is developed. The development cost is the exercise price and the real option can only be exercised once[1]. The landowner decides to develop a building on the land corresponds to exercising the option.

Let's imagine that a landowner has a vacant land. The current real option to the landowner is to decide whether the vacant land should be developed now, at time t_0. If the landowner intends to develop the owned vacant parcel at time t_0, an amount of money (C_{t_0}) should be invested into the development project for construction, the construction cost[2], C_{t_0}. Consequently, the approximate market value[3] of the constructed property, P_{t_0}, reflects the total

[1] Obviously, an existing building could be demolished, thereby reviving the vacant land is another real option. This situation is ignored in this topic.

[2] An assumption needs to be made, that the structure to build is predetermined and optimized. Thus, the fluctuations in construction cost have not resulted from the structure changes. Making this assumption is appropriate in Australia. Because every building project needs approval from the local government before starting. In the application, the structure to build is required to be declared. Thus, it is assured when landowners decide to develop the vacant lands they own.

[3] Because the development project may take one or several years to construct a property, such as a house, and may take a while to sell the built property. Such that, the approximate market value is used as the value of the investment output. If the built property is for renting purposes, the approximate market value could represent the present value of the future rent payments.

package price of the land and the building on the land. Such that, the payoff of the development project at time t_0 is

$$V_{t_0} = P_{t_0} - C_{t_0}. \tag{7.1}$$

The dynamics of property price and construction cost are time-varied. At any future time t, if the landowner chooses to develop, the payoff of the project at time t is

$$V_t = P_t - C_t, \tag{7.2}$$

Thus, there is a series of V_t $(t \geq t_0)$. For comparing the potential payoffs of the project, assuming r is the discount rate, no matter when the landowner decides to develop the owned vacant land, the payoff of the project at time t_0, could be generally expressed as

$$V_{t_0} = (P_t - C_t)e^{-r(t-t_0)}, \tag{7.3}$$

where $t \geq t_0$. To rational landowners, they would compare the payoff with the intrinsic value of vacant land (L_{t_0}) at time t_0. Thus, to conduct the development project, it requires that the difference between the payoff of the project at time t_0 and the intrinsic value of vacant land should be non-negative. The decision rule is

$$\begin{cases} if\ V_{t_0} \geq L_{t_0}, & develope\ the\ land; \\ if\ V_{t_0} < L_{t_0}, & do\ not\ develope\ the\ land. \end{cases} \tag{7.4}$$

If $V_{t_0} \geq L_{t_0}$, the landowners would invest C_t to develop the land at a future time t $(t \geq t_0)$, then receive a whole property (building and land) which is worth P_t. Otherwise, the landowners would not develop the vacant land. The equilibrium of the development project requires that

$$\begin{aligned} L_{t_0} &= \max(V_{t_0}) \\ &= \max_t[(P_t - C_t)e^{-r(t-t_0)}] \end{aligned} \tag{7.5}$$

V_t represents the payoff of the development project at any future time t ($t \geq t_0$), r is a discount rate. There are two problems arisen, the land value and the development timing. To solve them, the keystone is the dynamic paths of price, P_t, and construction cost, C_t over time.

Assuming the price of residential property and the construction cost are stochastic. Thus, P_t and C_t include risk premium, because of uncertainties, and they are assumed following the geometric Brownian motion of the form,

$$\begin{aligned} \frac{dP_t}{P_t} &= g_p\, dt + \sigma_p\, dz_p, \\ \frac{dC_t}{C_t} &= g_c\, dt + \sigma_c\, dz_c. \end{aligned} \tag{7.6}$$

g_p and g_c are the growth rates, σ_p and σ_c are the standard deviations. Additionally, dz_p and dz_c are the increments of a standard Wiener process. The problem is formulated as a first-hitting-time problem. The argument establishes, that development should occur when the ratio, P_t/C_t, reaches the critical barrier, R^*. Then, Eq. 7.5 could be rewritten, and the expected land value is

$$\begin{aligned} L_{t_0} &= V_{t_0} \\ &= \max_t E_{t_0}[(P_t - C_t)e^{-r'(t-t_0)}] \\ &= E_{t_0}[(R^* - 1)C^* e^{-r'T^*}] \end{aligned} \tag{7.7}$$

E_{t_0} is the conditional expectation on the available information at time t_0, r' is the return rate with a risk premium to discount the future values[4]. T^* is the duration from t_0 to the first-hitting-time ($t^* \geq t_0$) when P^*/C^* ($=P_{t^*}/C_{t^*}$) first reaches the critical ratio, R^*, and $V_{t^*}e^{-r'(t^*-t_0)}$ is maximized.

[4] Because the future is uncertain. Thus, the discount rate should have a risk premium. The Capital Asset Pricing Model (*CAPM*) can be used to determine the risk-adjusted discount rate, $r' = r_f + \beta[E(r_m) - r_f]$, where $E(r_m)$ is the expected discount rate of the market.

Following the Itô's lemma, the solution of Eq. 7.7 at time t_0[5] is

- When $P_{t_0}/C_{t_0} \leq R^*$,

$$\begin{aligned} L_{t_0} &= (R^*-1)C_{t_0}\left(\frac{P_{t_0}/C_{t_0}}{R^*}\right)^{\alpha}, \\ E_{t_0}(T^*) &= ln\left[\frac{C_{t_0}}{P_{t_0}}R^*\right] \Big/ \left[g_p - g_c - (\frac{1}{2}\sigma_p^2 - \frac{1}{2}\sigma_c^2)\right], \\ R^* &= \frac{\alpha}{\alpha-1}, \\ \alpha &= \frac{1}{\sigma^2}\left[-\left(g_p - g_c - \frac{1}{2}\sigma^2\right) + \sqrt{(g_p - g_c - \frac{1}{2}\sigma^2)^2 + 2\sigma^2(r' - g_c)}\right]. \end{aligned} \tag{7.8}$$

- Otherwise, when $P_{t_0}/C_{t_0} > R^*$,

$$\begin{aligned} L_{t_0} &= P_{t_0} - C_{t_0} \\ T^* &= 0, \\ R^* &= \frac{\alpha}{\alpha-1}, \\ \alpha &= \frac{1}{\sigma^2}\left[-\left(g_p - g_c - \frac{1}{2}\sigma^2\right) + \sqrt{(g_p - g_c - \frac{1}{2}\sigma^2)^2 + 2\sigma^2(r' - g_c)}\right]. \end{aligned} \tag{7.9}$$

The analogy to the financial option, the development project under stochastic process is analogous to a perpetual American call option on a common stock. It gives the landowners the right to spend the construction cost (the exercise price of the option, C^*) and receive a property (a share of stock) that is worth P^* at the optimal time t^*. The option will not expire, as the optimal time t^* is a random variable. The value of the option (the land value) is derived above. In the stochastic process, the uncertainties play important roles. Risk

[5] The derivation of the solution is shown in the appendix in detail.

premiums are taken into account when the option is evaluated and the timing of development is decided. Eq. 7.8 are used to determine the covariant for the value of the option and the timing of development in the Section 7.4.

7.2.1 Simulations of the theory

Before simulation the theory, it is assumed that growth rates, g_p and g_c, are less than the discount rate, r', and the growth rate of price should be larger than the growth rate of construction cost, $r' > g_p > g_c$. These assumptions are based on the restrictions described in the real option theory, shown in the appendix. The construction cost (C_{t_0}) is set to $\$100,000$, the growth rates of price and construction cost are 0.02 and 0.015, and the discount rate is 0.05. The ratio of price and construction cost (P_{t_0}/C_{t_0}) is from 0.8 to 1.2. Thus, the red line in Figure 7.1 presents the simulation of land values without uncertainties, the deterministic process. When the process is stochastic, including uncertainties, the green line in Figure 7.1 depicts the land values to the initial ratio of price and construction cost, based on the solution shown in Eq. 7.8 and Eq. 7.9. At this moment, σ_p^2 and σ_c^2, are equal to 0.0002 and 0.0001 respectively, and the correlation between price and construction cost is set to 0.8.

Figure 7.1 depicts the different situations of a land development project at time t_0. If the initial ratio is less than the optimal ratio (the dash lines), at time t_0, the theoretical land value is larger than the option value (the payoff). The blue line is lower than the red line or the green line. Thus, the landowners should wait until the optimal ratio is reached. If the initial ratio is equal to or larger than the optimal value, the land value is the option value (the payoff, $P_{t_0} - C_{t_0}$), and the landowners will build immediately. The red or green line is merged with the blue line. Additionally, from Figure 7.1, when uncertainties join the simulation, the optimal ratio becomes larger. It means the output price needs to be higher when holding the construction cost the same if there are uncertainties. Meanwhile, there is a gap between the land values estimated with and without uncertainties. A risk premium is added to the land value if the future is uncertain, while the initial ratio is the same.

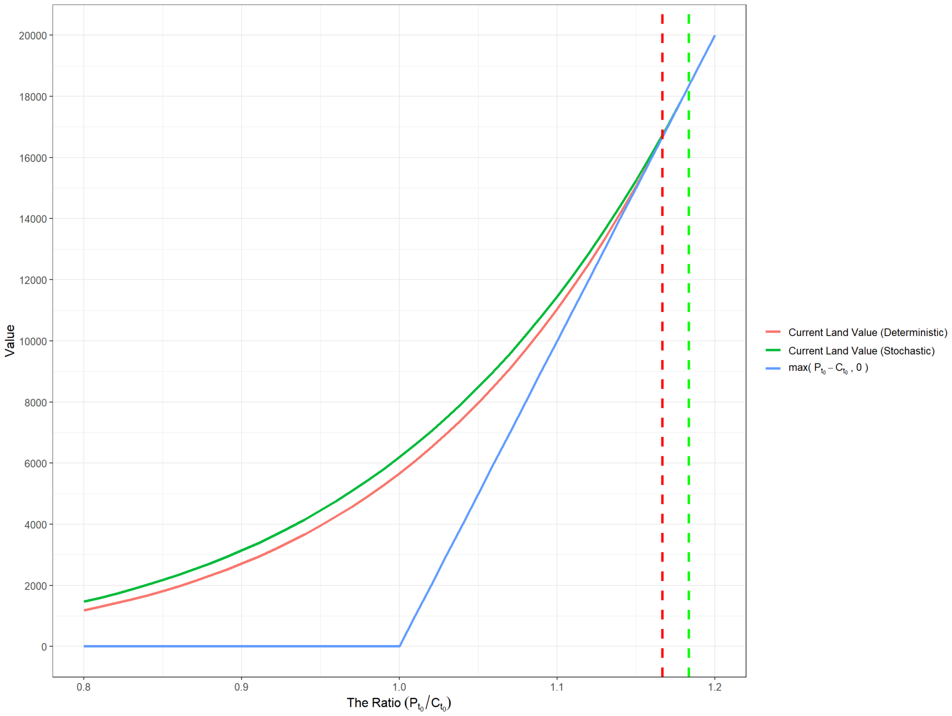

Figure 7.1: The land value graph: Deterministic process *VS* Stochastic process. The green line shows the current land value in the stochastic process under different initial ratios of price and construction cost (P_{t_0}/C_{t_0}). The red line shows the current land value in the deterministic process under different initial ratios of price and construction cost (P_{t_0}/C_{t_0}). The blue line shows the residual between the initial price and initial construction cost, if the difference is negative, the residual is equal to zero. The dashed lines depict the optimal R^* in the deterministic process and the stochastic process separately.

If assuming the future is certain, the optimal timing of development is determined and can be calculated using the available information. Otherwise, it is a random variable in the stochastic process. The simulation for the timing of development is set under the stochastic process. The settings are the same as the simulation of land value except for the initial ratio. The initial ratio (P_{t_0}/C_{t_0}) is equal to 1.08 in the simulation[6]. Meanwhile, Δt is the estimator of dt and is set to 0.01 to secure the accuracy of the simulation. The simulation runs $10,000$ times.

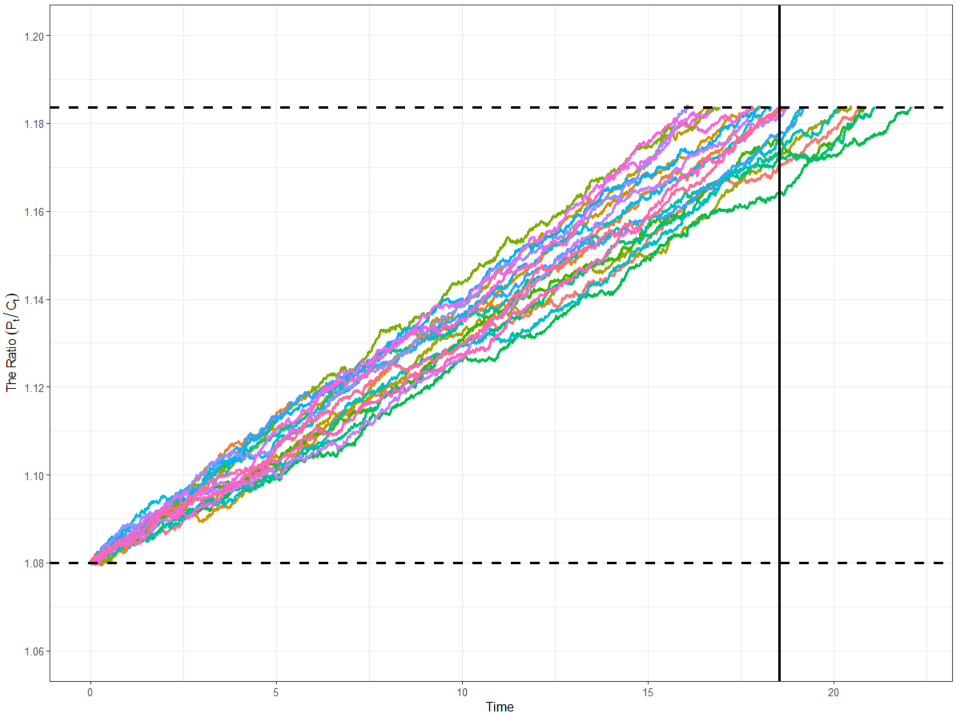

Figure 7.2: The simulation of expected stopping time (T^*). The graph only shows twenty trials of the simulation. The lower dashed line indicates the initial ratio of price and construction cost. The upper dash line shows the optimal R^*. When the ratio reaches the optimal one, the developer will convert the vacant land. The solid vertical line shows the expected value of stopping time (the expected timing of development).

[6] The lower initial ratio means the longer waiting time.

Figure 7.2 depicts the dynamics of the ratios via time for the first 20 trials. As this is a stochastic process, the vacant land could be converted early when the optimal ratio is reached. The simulated expected timing of development is 18.547 approximately. Comparing with 18.522, calculated by Eq. 7.8, the difference is acceptable[7].

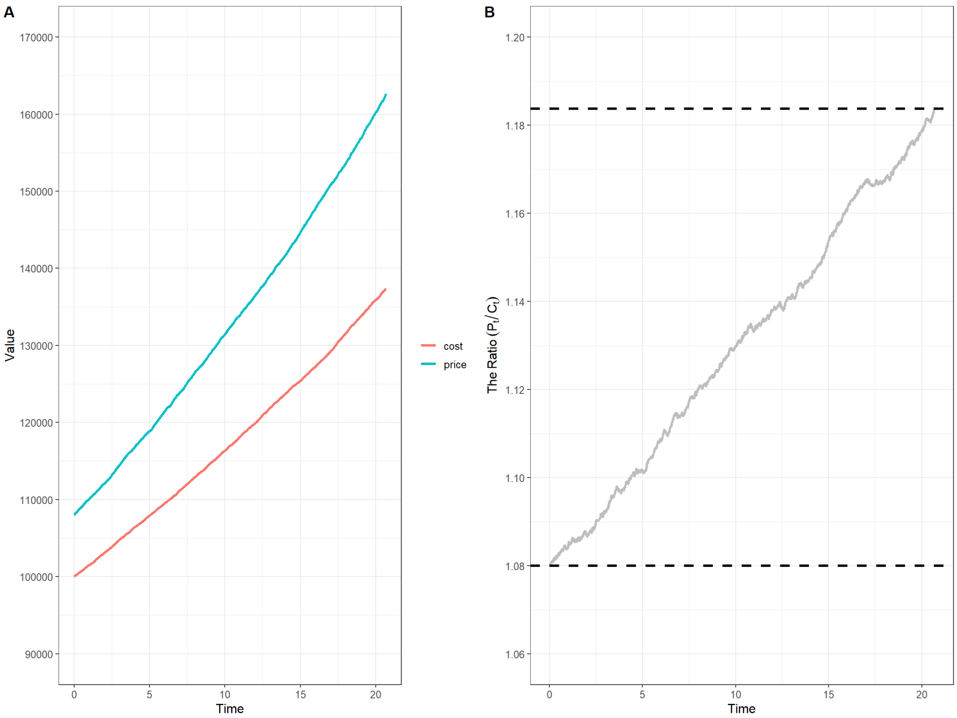

Figure 7.3: The dynamics of price, construction cost, and their ratio. The left graph depicts the dynamics of P_t and C_t. The right graph simply shows the path of the ratio, P_t/C_t.

Figure 7.3 depicts the dynamics of price, construction cost, and their ratios via time in one trial of the simulation, which clearly shows one possible path of the stochastic process. The Δt is set to 0.01, and the uncertainties are

[7] The accuracy of the simulation depends on the simulation times and the dt. The larger simulation times and the smaller Δt are chosen, the more accurate the expected timing of development is.

relatively small, thus, the lines of price and construction cost are smooth. The gap between price and construction is getting wider over time until the optimal ratio is reached, the duration is T^*. The ratio has a clear upward trend but fluctuated because of the uncertainties in price and construction cost.

From the theoretical proof and simulation, Table 7.1 summarizes the comparative statics results of the effects of determinants on the value of vacant land and the timing of development. The results are similar to those in the other real-option models like those of McDonald and Siegel (1986) or Capozza and Li (1994). The effects of determinants are theoretically explained, for example, the positive effects of volatilities on the value of vacant land. The empirical examination of these effects will be focused on and discussed in Section 7.5.

Table 7.1: The expected effects of key covariant on land value and timing of development. shows the effects of covariant on land value and timing of development. The signs in the table indicate the effect of a change in the covariant at the top column on the row variable. $*$ indicates the effect of a change in the covariant is not investigated. All the signs are theoretical under the prerequisites and assumptions, such as the growth of housing price is larger than the growth of construction cost.

Variable	Price level, P	Construction cost, C	Growth rate, g_p	Growth rate, g_c	Uncertainty, σ_p	Uncertainty, σ_c
Land value	$*$	$*$	$+$	$-$	$+$	$+$
Timing of development	$-$	$+$	$-$	$+$	$+$	$+$

7.3 Data for land development

There are two common land development scenarios when constructing a new dwelling for residential purposes in Australia (Australian Bureau of Statistics 2020b). One is that the land is purchased, and the owner enters into a building contract with a builder (includes "house and land" packages), the other is that the owner purchases a dwelling "off-the-plan" from a landowner (also known as a "turnkey" package). No matter which scenario, in Australia, the builders are required to apply for a building permit from the local council, then the land development project could be commenced. After the development is finished, the builders need to report the completion to the local council. Thus, the ideal data for analyzing land development are the building approval and the completion report for each project and each parcel, which contain some detailed information about the development project, such as expected spending and actual spending. However, these data are available at an aggregative level[8] rather than individual level. For privacy, the data for individual projects are not publicly accessible.

As discussed in Cunningham (2006), detailed transaction-level data could be an appropriate alternative to investigate the land value and timing of development. Two consecutive transactions for the same parcel could reflect the decision made for land development. If a piece of land is developed, the former transaction is a vacant land sale, and the latter is a sale for a house. If a vacant land is not developed, there will be two consecutive transactions for the same piece of land. Thus, the data are required to include information for parcel identification and the types of transacted properties. The data set described in Chapter 3 satisfies the requirements, the history of a parcel is composed of chronologically chained transactions, which could contain vacant land transactions only, residential dwelling transactions only, or both.

As mentioned previously, two data sets are prepared separately. For land

[8] The information of development projects are gathered monthly in each LGA.

valuation, all land transactions available in the data set, described in Chapter 3, are included. The land size, location, and transaction details are available in each record. For investigating the timing of development, firstly, the transactions of one parcel are chronologically chained. Then, the appropriate development projects are identified and picked out for composing a data set. The duration and an indicator for whether the vacant land is developed in the project are generated. Meanwhile, it is assumed that a house is considered in each decision for land development by default, which means only the land-to-house development is analyzed in this chapter[9]. More importantly, depending on what real options theory suggests in Eq. 7.8 and Eq. 7.9, several variables are generated and included in both data sets, such as the growth rates of house price and construction cost and the uncertainties. The variable-generating process is described in the posterior section.

In Section 7.2, dwelling prices and construction costs, and their growth rates and uncertainties contribute significant effects to land value and timing of development. These factors for individual development projects are not observed in the available data sources, such that, the "averages" in the local housing markets are used as the estimators. The series of "average" dwelling prices and construction costs are built in Section 7.3.1. Their growth rates and uncertainties are also generated in Section 7.3.2. Then, two data sets are described in Section 7.3.3 and Section 7.3.4.

[9] Because, initially, detached houses were the major property type in the Perth metropolitan area. Concentrating on this type of residential property will not lose many observations, and it is representative. Then, when only detached houses are considered, the relationship between vacant land and buildings is more straightforward, compared with some residential properties, such as duplexes, townhouses, and apartments, which share one piece of vacant land with the others.

7.3.1 Time series of dwelling prices and construction costs

As described in Eq. 7.8, P_{t_0} and C_{t_0} act significantly when evaluating land value and estimating the expected timing of development. They are measured by the quality-adjusted housing price indices and average construction costs respectively in each LGA.

Following the indexing process in Chapter 6, separate house price regressions are run with the following specification in each quarterly cross-section:

$$p_{i,t} = log(P_{i,t}) = f(\mathbf{x}_{i,t}, age_{i,t}) + \epsilon_{i,t}, \tag{7.10}$$

where $f(\cdot)$ is tree-structured GBM, $age_{i,t}$ is the age of the house, i, sold in quarter, t, $\mathbf{x}_{i,t}$ includes all the other structural characteristics and location characteristics, such as floor area, location coordinates and LGA factors, which are the same as the characteristics used in Chapter 5. To observe the price of a standard-quality house in each quarter t and each LGA j, the medians of characteristics are used instead of the means, because this could avoid fractions for some housing characteristics. In addition, the median may be more representative than the mean when the variable is skewed. There is a special case, the age of the house. It is set to 1 for the standard-quality house, because the output of a land development project is a new house, instead of an old house with price depreciation. Then, the log-price series of new standard-quality house, $p^{\bar{q}}_{j,t}$ is,

$$p^{\bar{q}}_{j,t} = f(med(\mathbf{x}_{j,t}),\ 1), \tag{7.11}$$

where 1 indicates all the houses are newly constructed, $med(\mathbf{x}_{j,t})$ includes the medians of all the other characteristics for each LGA, j, in quarter t.

Then, the average construction cost is calculated from the data sourced from ABS. The total construction costs of the building activities[10] in one LGA and the numbers of the building activities are documented quarterly. Thus, the quarterly average construction cost is simply calculated in each LGA,

$$C_{j,t} = \frac{TC_{j,t}}{N_{j,t}}. \tag{7.12}$$

Then, the logarithm of $C_{j,t}$ is computed, $c_{j,t}$ and $p_{j,t}$ are paired in each quarter. Thus, the series of these pairs is the estimator for the initial expected house price (P_{t_0}) and construction cost (C_{t_0}) for all development projects in LGA j, started in quarter t. These two time-series data are depicted in Figure 7.4 and Figure 7.5.

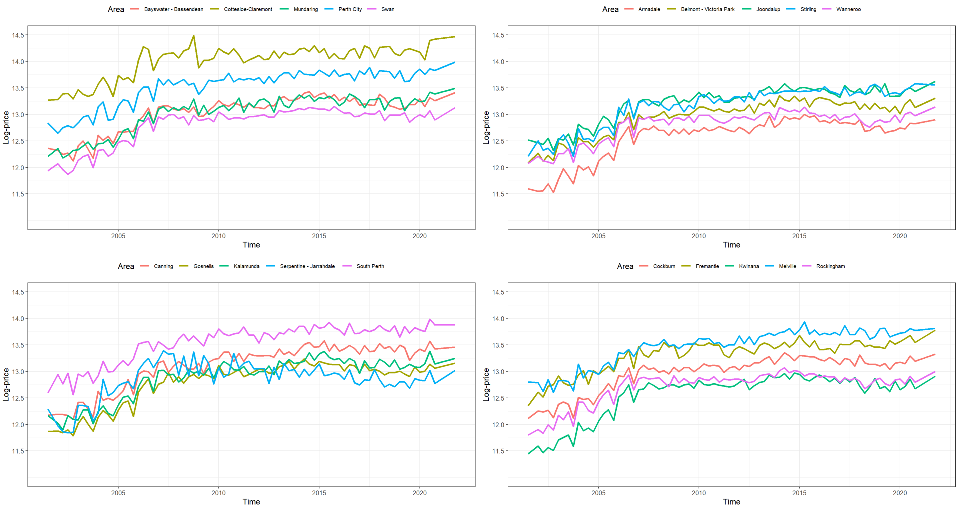

Figure 7.4: The time series plot of price in twenty *LGA*s. The x-axis is the time from 2003. The y-axis shows the logarithm of prices.

[10] This total number is the summation of the construction costs for all building activities in one LGA and in one quarter. The construction cost for one individual building activity is documented in the building approval. These approvals are not accessible to the public due to privacy issues.

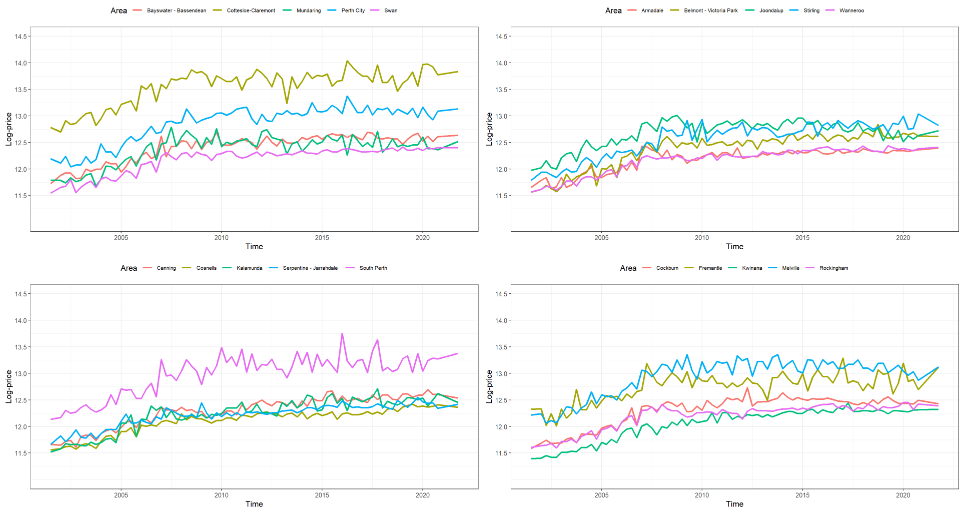

Figure 7.5: The time series plot of construction cost in twenty LGAs. The x-axis is the time from 2003. The y-axis shows the logarithm of averaged construction costs in one LGA.

7.3.2 Growth rates and uncertainties

As the time series of house price and construction cost are chronologically paired, the ideal approach for generating their growth rates and uncertainties is to apply vector autoregression (VAR). Before applying VAR, the stationarities of house price and construction cost series have to be checked. However, they are not stationary in this case, see details in the appendix. Instead, the growth rates for house price and construction cost are estimated, and the first-order differences of the prices of standard quality houses and average construction costs are regressed in VAR. The uncertainties are calculated using the recommended method in Cunningham (2006).

Growth rates

The first-order difference is calculated and paired as the following.

$$\begin{bmatrix} \Delta p_t \\ \Delta c_t \end{bmatrix} = \begin{bmatrix} log(P_t) - log(P_{t-1}) \\ log(C_t) - log(C_{t-1}) \end{bmatrix} = \begin{bmatrix} p_t - p_{t-1} \\ c_t - c_{t-1} \end{bmatrix}. \tag{7.13}$$

Then, the Δs are estimated in *VAR*. The specification could be expressed as

$$\begin{bmatrix} \Delta p_t \\ \Delta c_t \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix} + \begin{bmatrix} \alpha_{p,1} & \alpha_{c,1} \\ \beta_{p,1} & \beta_{c,1} \end{bmatrix} \begin{bmatrix} \Delta p_{t-1} \\ \Delta c_{t-1} \end{bmatrix} + \begin{bmatrix} \alpha_{p,2} & \alpha_{c,2} \\ \beta_{p,2} & \beta_{c,2} \end{bmatrix} \begin{bmatrix} \Delta p_{t-2} \\ \Delta c_{t-2} \end{bmatrix} + \cdots + \begin{bmatrix} \epsilon_{p,t} \\ \epsilon_{c,t} \end{bmatrix}. \tag{7.14}$$

To provide the up-to-date quarterly growth rate estimates, *VAR* is estimated in quarterly extending windows, and the optimal lag is selected for each estimation. The first window is from the third quarter of 2001 to the second quarter of 2006 through twenty quarters. It is to ensure that there are enough observations for initializing the *VAR*. The following window appends a new quarter in the previous window. In total, there are 77 observations in the last estimation, the window is from the third quarter of 2001 to the third quarter of 2020.

In this study, the unconditional expectations are considered estimates for growth rates. These expectations are described below. Assuming the lag of *VAR* is 1 (*VAR*(1)) as an example, the unconditional expectations of growth rates are calculated as

$$\begin{bmatrix} E(\Delta p) \\ E(\Delta c) \end{bmatrix} = \left(I_2 - \begin{bmatrix} \alpha_{p,1} & \alpha_{c,1} \\ \beta_{p,1} & \beta_{c,1} \end{bmatrix} \right)^{-1} \begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix}. \tag{7.15}$$

I_2 is identity matrix with rank 2. These two types of expected growth rates capture the trend of house price and construction cost in the short run and long run. As described in Section 7.2, the growth rates of property price and construction cost are assumed as constants. However, the growth rates may vary during the long term. Thus, in addition, the vector of unconditional variances for growth rates is calculated as

$$
\begin{aligned}
vec\Big(\begin{bmatrix} Var(\Delta p) & Cov(\Delta p, \Delta c) \\ Cov(\Delta p, \Delta c) & Var(\Delta c) \end{bmatrix}\Big) \\
= \Big(I_4 - \begin{bmatrix} \alpha_{p,1} & \alpha_{c,1} \\ \beta_{p,1} & \beta_{c,1} \end{bmatrix} \otimes \begin{bmatrix} \alpha_{p,1} & \alpha_{c,1} \\ \beta_{p,1} & \beta_{c,1} \end{bmatrix}\Big)^{-1} \\
vec\Big(\begin{bmatrix} Var(\epsilon_p) & Cov(\epsilon_p, \epsilon_c) \\ Cov(\epsilon_p, \epsilon_c) & Var(\epsilon_c) \end{bmatrix}\Big)
\end{aligned}
\tag{7.16}
$$

I_4 are identity matrices with rank 4. $vec()$ is to verctorize the matrix and $\otimes$ is the Kronecker product of matrices. The variance-covariance matrix of VAR is used for the calculation in Eq. 7.16 as two series are stationary. Meanwhile, the eigenvalue of the parameter matrix (the αs and βs) should be less than one. These variances of the growth rates for house prices and construction costs capture the uncertainties of growth rates when the assumption about constant growth rate is not true in Eq. 7.6. This measure captures the landowners' expectations for growth rates and their uncertainties in the long term, which are not considered in other empirical studies.

The uncertainties of house prices and construction costs

The estimates of uncertainties for house prices and construction costs used in this study apply the same measure in Cunningham (2006). The variance of housing price, $p^{\bar{q}}_{j,t}$, and construction cost, $c_{j,t}$ are computed and applied as the estimate, notated as $\hat{\sigma}_{p_{j,t}}^{2}$ and $\hat{\sigma}_{c_{j,t}}^{2}$. To measure the degree of uncertainties, the current price of the standard quality house and the current average construction cost are estimated separately, as a function of an intercept and the one-year-ago price of the standard quality house or average construction cost, $p^{\bar{q}}_{j,t-4}$ or $c_{j,t-4}$.

$$p^{\bar{q}}_{j,t} = \alpha_{p0,j} + \alpha_{p1,j} p^{\bar{q}}_{j,t-4} + e_{pj,t} \tag{7.17}$$

and

$$c_{j,t} = \alpha_{c0,j} + \alpha_{c1,j} p^{\bar{q}}_{j,t-4} + e_{cj,t} \tag{7.18}$$

where t is the quarter, and j denotes the LGA. Four quarters is the shortest time horizon a landowner might consider, given that the construction time for a new house in Perth is approximately one year. Meanwhile, the predicted potential house prices are estimated using Eq. 7.17, which present the potential house prices in one year ahead, notated as $E(p^{\bar{q}}_{j,t})$.

The estimates of price and construction cost uncertainty, $\hat{\sigma}^2_{pj,t}$ and $\hat{\sigma}^2_{cj,t}$, then is calculated as the four-quarter moving variance of residuals from the previous two equations,

$$\hat{\sigma}^2_{pj,t} = \frac{1}{4} \sum_{s=1}^{4} (\hat{e}_{pj,t-s} - \bar{e}_{pj,t}) \tag{7.19}$$

and

$$\hat{\sigma}^2_{cj,t} = \frac{1}{4} \sum_{s=1}^{4} (\hat{e}_{cj,t-s} - \bar{e}_{pj,t}), \tag{7.20}$$

where

$$\bar{e}_{pj,t} = \frac{1}{4} \sum_{s=1}^{4} \hat{e}_{pj,t-s} \tag{7.21}$$

and

$$\bar{e}_{cj,t} = \frac{1}{4} \sum_{s=1}^{4} \hat{e}_{cj,t-s} \tag{7.22}$$

This specification follows the same procedure of uncertainty measurement in Cunningham (2006). There are other methods available for the uncertainties, Cunningham had described and investigated them in the robustness check of his initial findings. He concludes that the variance of residuals is an appealing measure, comparing with the variance of prices directly, the sum of squared residuals and $GARCH(1,1)$. Using this method could ensure that the uncertainty's estimates only increase when a price forecast for a particular quarter deviates from the other three price forecasts in one year.

7.3.3 Land transaction data set

To test whether the uncertainties raise land prices because high-risk exposure provides a risk premium, a *land transaction data set* of vacant land sales is directly extracted from the full transaction data, introduced in Chapter 3. A total of 336,677 vacant land transactions were recorded from 1988 to 2020. Due to the construction costs being available after the second quarter of 2001, 227,336 observations were picked for this topic. Of these sales, 4,726 are bundle sales. And 5,149 observations are duplicated. In addition, 423 are dropped because they are not arm-length transactions, and the transaction prices of those sales do not reflect the market value of vacant lands.

Table 7.2: The details of the data preparation process for vacant land transactions. shows the number of observations deleted and the reasons why the observations were removed. The overlaps happen because some observations could match multiple washing rules.

		No. of obs.	Filtered obs.
Land transactions after *2001Q2*		227,336	
	Bundle sales		-4,726
	Duplicate records		-5,149
	Non-arm-length sales		-423
The available dataset		217,326	
	After *2006Q2* due to growth and uncertainty measurement		-71,467
The *land transaction dataset*		145,858	

Due to the measurement of growth rates and uncertainties, there are twenty quarters sacrificed. Thus, the *land transaction data set* excludes the transactions observed before the third quarter of 2006. In total, 145,858 vacant land transactions are available in the *land transaction dateset* from the third quarter of 2006 to the end of 2020. The details about the data set are summarized in Table 7.3 for investigating the effect of potential determinants for land valuation.

Table 7.3: Summary statistics for variables used to test the determinants of land price. summarizes the statistics for the available variables in the *land transaction dataset*. The growth rates and their uncertainties in the short term and long term are summarized separately in the table. The property use is a dummy variable, 1 means the vacant land is for residential purposes, 0 means the other purposes.

Variable	Mean	Median	S.D.
Vacant land price (log)	12.481	12.409	0.458
Growth of house price	0.026	0.024	0.011
Uncertainty of house price growth	0.010	0.008	0.008
Uncertainty of house price	0.006	0.004	0.008
Growth of construction cost	0.016	0.015	0.008
Uncertainty of construction cost growth	0.007	0.005	0.007
Uncertainty of construction cost	0.006	0.002	0.010
Lot size (log, m^2)	6.088	6.087	0.563
Distance to CBD (km)	25.350	22.777	12.084
Property use	0.069		
N	145,858		

7.3.4 Land development data set

Similar to normal survival analysis, vacant land development is an interesting event in this topic, and the optimal timing and its hazard are keen to be evaluated. Before applying survival analysis, two declarations should be made when a data set is generated for the timing of development. The first is that each parcel of land is considered as a "patient" in a clinical trial. If the parcel is developed, then, it is treated as a "dead" land. Otherwise, it survives from the development event. The status of a parcel could be 1 (“dead”) or 0 (“survived”). Secondly, informative censoring is considered as default. This censoring method is commonly used in survival analysis when the participants are lost to follow-up due to some reasons. Thus, the development status of each parcel could be decided by its latest transaction. Additionally, this censoring method could confirm that the estimated waiting times are always longer than the actual waiting time. The follow-up period from the date of the latest sale to the end of 2020 is unobserved and unknown, then, the potential status changes are censored.

Land development identification

To investigate land development, initially, it should figure out when vacant lands are at risk of developing. The *data set* described in Chapter 3 only contains the transaction records, thus, the changes in development status could only be identified between two consecutive transactions for the same parcel. Simply, if there are two consecutive “land” transactions, it means the vacant land is not developed. If the former is a “land” sale, and the latter is a “house” transaction, then, it reflects that this parcel is development. For this, the *data set* has been modified to a series of chained transactions to clarify the transaction history for each parcel. The transactions for the same parcel are linked by the same “Land ID” and “Parcel ID”, and chained in chronological order.

There are generally five situations about the land development in the *data* *set*[11]. The possible situations of two consecutive transactions are

1. The ideal series of chained transactions contain both land transaction(s) and house transaction(s), meanwhile, both of them are observed from 2006Q3 to 2020Q4. The parcel was developed after its last land transaction in the investigation period (*Land-to-house 1* in Figure 7.6).
2. Only house transaction(s) is(are) observed in the period from 2006Q3 to 2020Q4. The land transaction(s) should be observed before 2006Q3, but not shown in the investigation period. The parcel was developed in the investigation period (*Land-to-house 2* in Figure 7.6).
3. Only house transaction(s) is(are) observed in the period from 2006Q3 to 2020Q4. But the parcel was developed before 2006Q3, which is out of the investigation period. (*Land-to-house 3* in Figure 7.6).
4. Only one land transaction is observed in the investigation period. Not both of the two consecutive land transactions are observed, there should be another land transaction(s) observed before 2006Q3. In this situation, the parcel was not developed. Because the latter of two consecutive transactions is still a sale for vacant land. The land could be developed after the last land transaction, but there is no follow-up information available in the *data set*. Then, the vacant land is believed as undeveloped. This status is updated. (*Land-to-land 1* in Figure 7.6).
5. Only land transactions (at least two transactions) are observed in the investigation period. It is reflected that the parcel is undeveloped because these two transactions are the latest for the parcel. The land could

[11] If a parcel is not in stock, it is treated as not-at-risk, and is not considered in this study.

be developed after the last land transaction, but there is no follow-up information available in the *data set*. Then, the vacant land is believed as undeveloped. This status is updated. (*Land-to-land 2* in Figure 7.6).

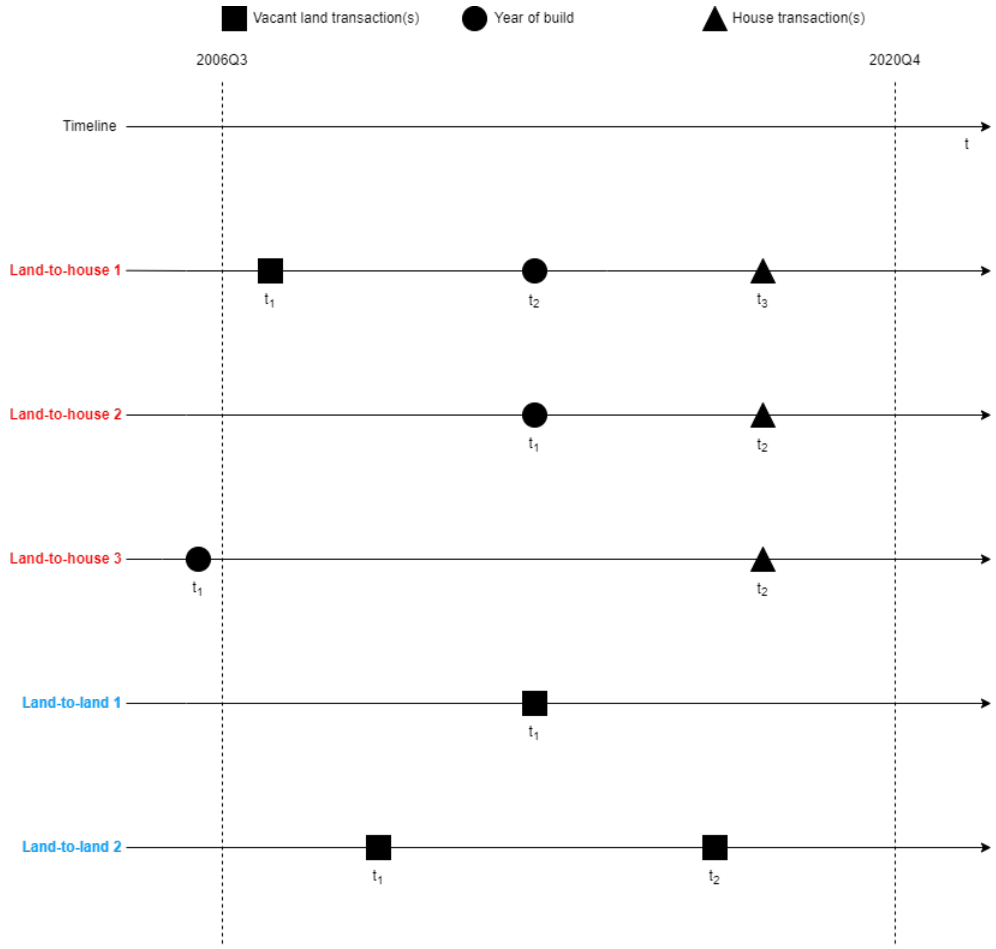

Figure 7.6: The explanation about the normal cases that could happen and what are observed. The red color means the developed cases are observed in the full dataset. The blue color means the information about the series is observed for pure land transactions (undeveloped cases). The history before the last transaction is known and the future after the last transactions are not observed. The development status for each available parcel is confirmed by checking its latest two consecutive transactions.

As declared previously, informative censoring is applied, thus the follow-up information from the date of the latest transaction to the end of 2020 is avoided. For *land-to-house 1*, *land-to-house 2*, and *land-to-house 3*, the follow-up information of each parcel is not that necessary, as the development status has been marked as developed or "dead". For *Land-to-land 1* and *Land-to-land 2*, the land could be developed after the latest transaction. The status of the parcel could change in the follow-up period. However, firstly, it is uncertain whether the land will be developed in the follow-up period. Secondly, the decision on land development made in the follow-up period is independent of the decision made previously. Because the vacant land has been sold to others, the correlation between these two decisions of land development made by different landowners is tiny and avoidable. Such that, for *land-to-land* situations, the follow-up period is not considered. Besides these five general situations, there are also some special cases, such as subdivision and demolition, which are discussed below.

Subdivision & Demolition

Subdivision and demolition are two common special situations. Figure 7.7 depicts all the possible situations of subdivision and demolition.

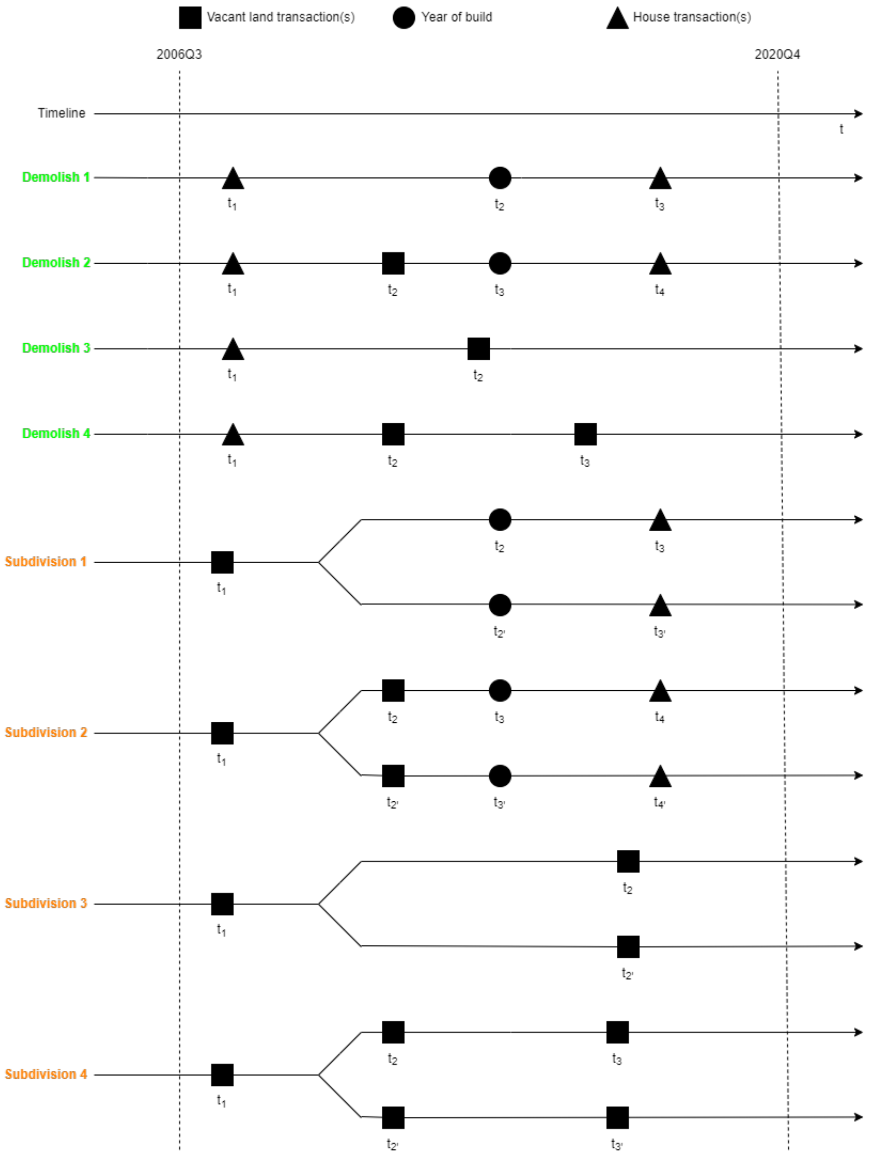

Figure 7.7: The explanation about the special cases (subdivision and demolition) could happen and what is observed. The green color means the possible demolition cases. The orange color means the information about the subdivision cases. In this study, only the development decision is considered, which means the interesting process is from vacant land to a built house. The composite decisions and projects are included in the analysis, such as the first case in subdivision and demolition.

For the demolition cases, the "Parcel ID" and "Land ID" are unchanged, thus, the full history of the parcel could be observed. *Demolish 1* shows two consecutive house transactions, however, the "year of build" records are different, and the latest "year of build" is after the sold date of the former transaction. Meanwhile, normally, the housing characteristics documented in two consecutive house transactions are different also[12]. In this case, it is believed that the old house was demolished and then, the owner built a new house on the land. Thus, there are two decisions are made during the period between two consecutive transactions, one is to decide whether demolition is appropriate, and the other is to decide whether the vacant land after demolition needs to be developed. This situation is not considered. Because, firstly, these two decisions are composite, the land development decision highly relies on the demolition decision. And then, the land development process is not complete, there is no information about when the parcel becomes vacant land after demolition.

For *Demolish 2*, after demolition, the vacant land was sold. Then, this parcel is developed as a normal land development process. In this case, the demolition and land development are two separate and independent decisions by different landowners. Only the land development part is considered, and the observations under this situation are similar to *Land-to-house 1.* Similarly, if the vacant land after demolition is resold at least once (*Demolish 3* and *Demolish 4*), these situations could be treated as *land-to-land 1* and *land-to-land 2* respectively. However, *Demolish 3* is excluded in the *land development dataset.* Because this situation only reflects that a parcel is demolished, there

[12] There are some chances the housing characteristics are not changed. The owner may build a new house with the same number of facilities. However, this case is rare.

is no development information available. But, under *Demolish 4*, the two consecutive land transactions indicate that the land remains undeveloped, which is of interest.

For the subdivision situations, the "Parcel ID" and "Land ID" are changed, thus, the full history of the parcel before and after subdivision can not be smoothly connected by the IDs. However, the land plan type could be helpful to locate them, which involves information in the "Parcel ID" and "Land ID". The IDs are generated by numbers and letters. If the IDs include the letter, "S", it means the parcel is one of the subdivided pieces of land. Then, the lands after the subdivision are concerned.

For *Subdivision 1*, the former land sale has different "Parcel ID" and "Land ID" from the following house transactions. When the original parcel is subdivided is unknown. Such that, this situation is not included in the *land development dataset*, due to that the land development process is incomplete, even the subdivided parcel(s) is(are) developed. For *Subdivision 2*, the subdivided vacant lands are sold after subdivision. After, each separate land could be developed or resold. Under this situation, the series of transactions for the subdivided parcels can be tracked because of the same "Parcel ID" and "Land ID". Such that, *Subdivision 2* is a special case of *land-to-house 1.* If the subdivided vacant lands are sold only once after subdivision, they are under *Subdivision 3* and treated as *land-to-land 1.* This situation is excluded from the *land development dataset* because there is no land development information available. If the subdivided vacant lands are sold at least twice (*Subdivision 4*), they are considered as new vacant lands in the market. The observations under *Subdivision 4* are similar to the observations under *Land-to-land 2.*

In total, there are **346,136** parcels in stock after 2006Q2. For *Land-to-house 1* with *Demolish 2* and *Subdivision 2*, there are **26,140** parcels that have a full history from land to house (**1,240** observations for *Subdivision 2* and **19** observations for *Demolish 2*). In *Land-to-house 2*, there are **20,110**. And in *Land-to-house 3*, there are **183,337** parcels. Meanwhile, there are **1,783** parcels that only observe house transactions but no information about the year of build in *Land-to-house 2* and *Land-to-house 3*. For land-to-land cases, *Land-to-land 1* has **85,973** parcels. For *Land-to-land 2* with *Demolish 4* and *Subdivision 4*, there are **7,735** parcels (**794** observations for *Subdivision 4* and **47** observations for *Demolish 4*). The selection of diverse situations is summarized in Table 7.4

Table 7.4: The details of data preparation process for different land development situations. shows the number of observations under each situation. The situations with an asterisk are included in the *land development dataset*.

	Situations	The reason why not include	Number of obs.
Total 346,136 parcels in stock			
Common situations			
	Land-to-house 1*		24,881
	Land-to-house 2*		20,110
	Land-to-house 3	Land developed before 2006Q2	183,337
	Land-to-land 1*		85,973
	Land-to-land 2*		6,894
Demolition & Subdivision			
	Demolish 1	Development process incomplete	644
	Demolish 2*		19
	Demolish 3	Demolition only	452
	Demolish 4*		47
	Subdivision 1	Development process incomplete	7,460
	Subdivision 2*		1,240
	Subdivision 3	Subdivision only	12,475
	Subdivision 4*		794
Other		Missing “year of build”	1,783
		Data updating error	27
The *land development dateset**			139,579

From 2006Q3, there are **139,579** parcels that are included in the data set for the timing of development, **46,226** parcels are developed, **93,353** parcels are still vacant and undeveloped. The available information for the analysis of the timing of development is summarized in Table 7.5.

Table 7.5: Summary statistics for variables used to test the determinants of land development. summarizes the statistics for the available variables in the *timing of development* data set.

Variable	Mean	Median	S.D.
Dummy if parcel developed	0.331		
Predicted potential house price (log)	13.163	13.124	0.263
Construction cost (log)	12.323	12.297	0.239
Growth of house price	0.034	0.029	0.017
Uncertainty of house price growth	0.011	0.008	0.009
Uncertainty of house price	0.007	0.005	0.007
Growth of construction cost	0.019	0.016	0.014
Uncertainty of construction cost growth	0.007	0.005	0.007
Uncertainty of construction cost	0.005	0.002	0.008
Lot size (log, m^2)	6.162	6.114	0.587
Distance to CBD (km)	26.363	23.696	11.984
Property use	0.021		
N	139,579		

7.4 Analysis design

As mentioned in the Chapter 4, the average marginal effects could be calculated using *ALE*. However, those marginal effects lack statistical inference. Thus, in this chapter, the parametric models are applied, which are as simple as linear models, in case the machine learning implementations are hard to interpret. The model specifications described below are for parametric models.

7.4.1 Vacant land valuation

As discussed in Eq. 7.8, the determinants for the valuation of vacant lands include the current value of investment output (the house price), the current investment cost if developed (the construction cost), and a constant α which is determined by the growth rates of price and cost, discount rate and the uncertainties at the beginning of the development project. A reduced form of vacant land valuation is established as a function of the growth expectations and the uncertainties of house price construction cost and individual land characteristics.

$$\begin{aligned} l_{it} &= log(L_{it}) \\ &= \beta_0 + \beta_1 \tilde{\sigma}^2_{pj,t} + \beta_2 \tilde{\sigma}^2_{cj,t} + \beta_3 \tilde{g}_{pj,t} + \beta_4 \tilde{g}_{cj,t} \\ &\quad + \beta_5 lotsize_i + \beta_6 distCBD_i + \beta_7 \mathbf{D}_j \\ &\quad + \boldsymbol{\beta}_8 \mathbf{D}_t + \beta_9 D_{purpose,i} + \varepsilon_{it}. \end{aligned} \tag{7.23}$$

where $i \in j$, which means the observation i belongs to the LGA j. Notes that, the expected growth rates and the uncertainties are measured in each quarter (t) and each LGA (j). For individual transactions, the distance to *CBD*, $distCBD_i$, the land size, $lotsize_i$, the purpose of the land ownership, $D_{purpose,i}$, and coordinates, lon_i and lat_i, only vary across observations. The reduced form does not include the house price level and construction cost level in one LGA due to they may cause the collinearity problem. In addition, including the discount rate is not appropriate in the valuation model, due to it is a temporal-only variable. In lieu of discount rate, quarter dummy variables D_i are included to measure the fixed effect in each quarter. $D_{purpose,i}$ controls the use of a vacant land. If the parcel of vacant land is for residential purposes, $D_{purpose,i}$ is equal to 1. This variable may investigate the fixed effect of the purpose when buying land. If the βs are positive, this indicates that the higher value of the factor boosts the price of vacant land. Otherwise, the factors harm the land prices.

According to Table 7.1, growth rates (g_p and g_c) may have completely different effects on vacant land value. The growth rate of house price level will show a positive effect on the land value, because of the increase of P_t. On the contrary, the increase in the growth rate of construction cost, g_c, raises C_t, which results in a decrease in vacant land value. Beforehand, there is a prerequisite that needs to be satisfied. The growth rate of construction cost should not be larger than the discount rate ($r - g_c \geq 0$). The estimates of growth rates are calculated using vector autoregression which is described in Section 7.3.2. The estimates, $\tilde{g}_p$ and $\tilde{g}_c$, are used as the unconditional expectations of the growth rates (or return rates) which could reflect the trend of long-term growth. The growth rates for a parcel located in the LGA j and sold in the quarter t are $\tilde{g}_{p_{j,t}}$ and $\tilde{g}_{c j,t}$.

As the tendency of people is risk-averse, the higher uncertainties may indicate the higher value of vacant land. Because people need a risk premium to cover the uncertainties they may face, the price of vacant land is equal to the summation of the risk-neutral land value and risk premium. Thus, this risk premium makes vacant land more "valuable". In addition, the higher uncertainties (or higher risk) may lead to a higher risk premium (a higher price of vacant land). An increase in either σ_p^2 or σ_c^2 will increase the price of vacant land. The uncertainties affect the dynamics of the house price level and construction cost level, P_t and C_t, these fluctuations exposed may need to be covered by premiums. The estimate for σ_p^2, $\tilde{\sigma_p}^2$, and the estimate for the uncertainty of the cost (σ_c^2), $\tilde{\sigma_c}^2$ are calculated in Section 7.3.2, following Cunningham's suggestion. The uncertainties for a parcel located in the LGA j and sold in the quarter t are $\tilde{\sigma}_{p_{j,s}}^2$ and $\tilde{\sigma}_{c_{j,s}}^2$.

7.4.2 Timing of development

To test the determinants for the timing of development, one method for survival analysis is estimated, in which vacant land is graduated or "dead" while a building is constructed. The duration is measured from the history of transactions, it is explained in Section 7.3.4. If a parcel is developed, the previous record of its two consecutive transactions would document it as vacant land, the recent one will be a house transaction. The duration is equal to that of the year of build subtracted from the year of the previous vacant land sale. Otherwise, both of the two consecutive records are vacant land sales. The duration is the gap period between the year of two vacant land sales. The longest duration could be 14 years, from 2006 to 2020. The theoretical determinants described in Eq. 7.8 include the current house price level, the current construction cost level, the growth rates for house price and construction cost, and their uncertainties. Besides them, the effect of location is also involved, such as the distance to CBD. Utilizing these determinants, the Accelerated Failure Time model (AFT) is applied. It is an alternative to the commonly used proportional hard models, but it estimates the "failure time" directly rather than the "hazard of failure". The AFT is specified as

$$\lambda(t) = \lambda_0(t_0)e^{f(\mathbf{X})} \tag{7.24}$$

where the baseline hazard, $\lambda_0(t_0)$, is shifted by a vector of covariates, $\mathbf{X}$. To achieve this survival analysis and investigate the timing of development, the duration and the status of development collaborate as the dependent variable ($\lambda(t)$) in the AFT model[13]. The covariates are presented in the following:

$$\begin{aligned} f(\mathbf{X}) &= \mathbf{X}'\beta \\ &= \beta_0 + \beta_1 E(p^{\bar{q}}_{j,t}) + \beta_2 \bar{C}_{j,t} + \beta_4 \tilde{\sigma}^2_{p_{j,t}} + \beta_5 \tilde{\sigma}^2_{cj,t} + \beta_7 \tilde{g}_{p_{j,t}} + \beta_8 \tilde{g}_{cj,t} \\ &\quad + \beta_9 distCBD_i + \beta_{10} lotsize_i + \beta_{11}\mathbf{D}_t + \beta_{12}\mathbf{D}_j. \end{aligned} \tag{7.25}$$

Note that the current levels of house price and construction cost, the growth rates for house price and construction cost, and their uncertainties vary by quarter t and by LGA j. Thus, the effects of these determinants on the timing of development are identified by intertemporal and cross-sectional variation. Parcel's distance from the CBD, $distCBD_i$, and coordinates, lon_i and lat_i, only vary across parcels i. If the βs are positive, this indicates that the optimal timing is delayed. If the βs are negative, the factors encourage the landowners to covert the vacant lands.

According to Eq. 7.8, assuming α is a constant, house price level (P_{t_0}) increases, the optimal timing t^* will be brought forward. Thus, $E(p^{\bar{q}}_{j,t})$ should have a negative effect on duration. This indicates that the landowners are more likely to convert the owned vacant lands immediately or just wait for a short time. The house price level of newly constructed buildings is to represent

[13] The duration and the status of development are combined by the function, "Surv". The combination is treated as the dependent variable in the model. The parameters (βs) reflect the effect of determinants on the duration of the development project.

P_{t_0}, described in Section 7.3.2. To capture the changing demand for different locations, the price levels are compiled in each LGA. If a parcel is in the LGA j and decided to convert in the quarter t, the price level of a newly constructed house for this parcel is $E(p^{\bar{q}}_{j,t})$.

The effect of the construction cost, C_{t_0}, on the optimal timing is contrary to the influence of the price level. It means the higher construction cost may delay the optimal timing to convert the vacant lands, which will show a positive sign in the hazard model. These costs are observed in the building approvals reports issued by the Australian Bureau of Statistics (ABS). The building approvals report includes the total values of the construction cost and the number of approved dwellings, they are aggregated in each quarter and each LGA. The average in LGA j is used to estimate the construction cost for a parcel, that is in the same LGA and to be developed in the quarter t, denoted as $C_{j,t}$. C_{t_0} in Eq. 7.8 is the construction cost at the beginning of the development project, thus, the current construction cost level in the local region, $C_{j,t}$, is its estimate. The dynamics of construction cost could be captured by its growth rate and uncertainty, g_c and σ_c^2.

Recall Eq. 7.8, the higher house price growth will reduce the duration. In contrast, the higher construction growth will extend the duration. Because, the optimal timing of development is the moment when the ratio of house price and construction cost reaches the optimal ratio, $R^* = P_t/C_t = P_{T^*}/C_{T^*}$. If the optimal ratio is not reached at the beginning, the duration will be accelerated when the house price growth is higher; but, the optimal ratio will be achieved late when the construction cost growth is higher. Meanwhile, the higher risk exposure to house price may lead to a lower willingness to develop immediately; the higher risk exposure to construction cost may lead to a higher willingness to develop immediately. Because landowners intend to choose the opportunity that potentially has lower risk. In other words, the higher house price uncertainty forces landowners not to develop vacant lands, which means the longer optimal conversion time t^* (Capozza and Li 1994). Under this situation, quitting land development could avoid the risk of future house price fluctuation. However, when construction cost risk potentially increases, it may lead landowners to convert vacant lands. Because they want to avoid potential cost fluctuations in the future. But, the truth might be that the risk appetites of project landowners are one of the actual determining factors for the timing of development. If one landowner is risk-loving, he may be encouraged to convert the vacant land when the house price uncertainty is high. The effect of uncertainties on the timing of development will be opposite to the suggestions from the real option theory (Eq. 7.8). The actual effects of uncertainties will be investigated, the uncertainty estimates used in the hazard model are the same as the estimates used in the vacant land valuation, $\tilde{\sigma}_{p_{j,s}}^2$ and $\tilde{\sigma}_{c_{j,s}}^2$, described in Section 7.3.2.

7.5 Empirical results

7.5.1 The results for valuation

Land valuation (Eq. 7.23) is conducted via linear model and *GBM*. For investigating the accuracy of valuation, the full data set is divided into ten folds. One of the ten folds is picked as the test sample, the rest is for training. The result for accuracy is summarized in Table 7.6. Overall, *GBM* still shows better accuracy than the linear model in *RMSPE* and *MAPE*, roughly 7% more accurate. It is consistent with what is found in Chapter 5. However, whether *GBM* is explainable and what the effects of determinants are on land values need further investigation.

Table 7.6: The errors in 10-fold cross-validation on the full dataset. All numbers are in percentage.

Error (%)		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
RMSPE	*LM*	28.52	28.56	28.93	28.93	28.67
	GBM	21.52	22.25	22.64	22.10	22.61
MAPE	*LM*	18.44	18.61	18.77	18.76	18.59
	GBM	11.13	11.44	11.71	11.38	11.47
Error (%)		Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
RMSPE	*LM*	28.46	29.58	28.97	29.02	28.74
	GBM	21.70	23.06	22.42	22.58	22.68
MAPE	*LM*	18.77	18.98	18.82	18.67	18.75
	GBM	11.38	11.74	11.69	11.49	11.51

Table 7.7 shows the coefficients of the determinants estimated from the linear model and *GBM*. From the results of inherent land factors, three outcomes can be concluded. Firstly, a larger land size indicates a higher land estimated value. If the land is 1% larger, the value is around 0.4% higher estimated in the linear models and around 0.5% higher estimated in the *GBM*s. Secondly, the land is estimated cheaper if it is far away from the city downtown. The value will decrease by around 1.3% for every additional kilometer far away from the city center using the linear models. The land value drops around 1.6% for the

GBM estimators. Then, the residential vacant land is slightly more expensive than the similar vacant land for other purposes. On average, the gap of the values is around 7%. For *GBM*, the same positive effect is found in the "residential purpose", however, the magnitude is dramatically high. It shows that the value of the vacant land bought for residential purposes is average 39% higher than the value of vacant land for other purposes.

Recall the real option theory (Eq. 7.8), if real options exist, then greater uncertainties about future prices and construction costs should raise the option premium and result in a higher observed sale price. Meanwhile, a higher rate of price growth and a lower rate of construction cost growth could lead to a higher land value. In column 1 of Table 7.7, the coefficient for price growth is positive and significantly greater than zero, suggesting that the land value will dramatically increase around 4.6% when the growth rate of house prices increases around 1%. Meanwhile, the coefficient for construction cost uncertainty is also positive and statistically significant. It reflects that a one-standard-deviation increase in construction cost uncertainty results in a modest increase in the land price of AUD\$698 in 2020Q4. However, the coefficient for the price uncertainty is not statistically significant, which means the price uncertainty does not significantly affect the land value in Perth. The potential reason could be that there is a high level of confidence in Perth's housing market and a high tolerance for temporary price drops, as the quarterly growth rates of house prices are always positive throughout the investigation period. Surprisingly, the coefficient for the growth of construction cost is different from the suggestion in the real option theory. Possibly, because the assumptions may be not always valid during the period, the discount rate is not always larger than the growth rate of construction cost and the growth rates of construction costs are not always positive.

Table 7.7: The factors' effect on land value. summarizes the effect of determinants on the vacant land price. The marginal effect for each determinant in the *GBM* is the average that is calculated in the full determinant's range. The numbers in the parentheses are the standard errors of parameters. The standard errors for *GBM* are generated using bootstrapping methods (20,000 times). These standard errors are only for reference and comparison, there is no inference purpose. Asterisk (*) means the variable is statistically significant at 0.05 significant level (only applicable for linear model).

Explanatory variable	*LM*		*GBM*	
House price growth	4.5685 *	6.1943 *	1.0064	1.1383
	(0.3696)	(0.3870)	(0.0298)	(0.0272)
Construction cost growth	0.7612 *	0.9445 *	0.6898	0.3500
	(0.2563)	(0.2591)	(0.0435)	(0.0358)
House price uncertainty	-0.0427	-0.2037	1.1401	0.5814
	(0.1283)	(0.1432)	(0.0411)	(0.0366)
Construction cost uncertainty	0.2848 *	0.8727 *	-2.5092	-0.0851
	(0.1048)	(0.1113)	(0.0462)	(0.0360)
Uncertainty of price growth	-	-1.0280 *	-	-7.9151
	(-)	(0.5156)	(-)	(0.0952)
Uncertainty of cost growth	-	-5.1922 *	-	-5.4805
	(-)	(0.3297)	(-)	(0.1103)
Land size	0.3919 *	0.3924 *	0.5132	0.5057
	(0.0014)	(0.0014)	(0.0009)	(0.0009)
Distance to CBD	-0.0132 *	-0.0132 *	-0.0172	-0.0155
	(0.0001)	(0.0001)	(0.0004)	(0.0004)
Residential purpose	0.06619 *	0.0706 *	0.3943	0.3855
	(0.0040)	(0.0040)	(0.0006)	(0.0006)
Constant	11.0885 *	11.1244 *	-	-
	(0.0277)	(0.0317)	(-)	(-)
Quarter fixed effect	Yes	Yes	-	-
LGA fixed effect	Yes	Yes	-	-
N	145,858	145,858	145,858	145,858

In column 2 of Table 7.7, the uncertainties of growth rates are taken into account, when the landowners do not believe the growth rates are constant. Both of the coefficients are negative and statistically significant, it reflects that the higher risks on growth rates result in land price drops. Because the expected potential of the housing market is at risk if the growth rates are uncertain. When evaluated at the median land value, a one-standard-deviation increase in the uncertainties of house price growth and construction cost growth leads to AUD\$2,015 and AUD\$8,905 decrease in land price respectively in 2020Q4. These changes in value are equivalent to 0.8% and 3.6% drops in land value. The rest of coefficients in the column 2 show the same or similar characters of the coefficients in column 1 of Table 7.7. It looks like these two uncertainties play important roles in the "real option premium" evaluation.

For *GBM*, compared to the coefficients from the linear model, the signs of marginal effects are the same as the signs of coefficients in the linear model except for the coefficients for the uncertainties of house price and construction cost. The magnitude of the marginal effects is slightly different. The marginal effects suggest that the land value will increase by around 1% and 0.7% when the growth rates of house price and construction cost increase by around 1% respectively. Meanwhile, for the uncertainties of growth rates, a one-standard-deviation increase in these uncertainties leads to a 6.3% and 3.8% decrease in land price respectively in 2020Q4. However, because of the lack of statistical inference, there is less confidence to believe whether the average marginal effects reflect the influence on the value of vacant lands for the determinants and whether the effects are statistically significant. Indirectly, this reflects that the ability to interpret *GBM* is limited using *ALE*.

7.5.2 The results for the timing of development

The survival analysis is conducted via accelerated failure time models (*AFT*) with linear model specification and *GBM*. It is more straightforward due to the duration of waiting for development is directly estimated, rather than the hazard of development. For the *AFT* model with linear specification, two distributions, Exponential distribution, and Weibull distribution, are assumed as they commonly model the random variables like time to failure or time between events. Initially, the performance of these two models is tested using ten-fold cross-validation. Table 7.8 summarizes the results of the performance. Overall, the *AFT* with *GBM* still shows a better accuracy than the *AFT* with linear specification in *RMSE* and *MAE*. However, both of the models perform considerably, the difference is 0.3, when the errors show the difference between predicted duration and real duration.

Table 7.8: The errors in 10-fold cross-validation on the full dataset. The errors for *AFT* model with linear specification are shown in the results when the model applies exponential distribution. The errors depict the difference between the predicted duration and the real duration.

Error (%)		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
RMSE	*LM*	2.1713	2.0834	2.1129	2.1226	2.1124
	GBM	1.8243	1.7466	1.8452	1.8027	1.8083
MAE	*LM*	1.5154	1.4719	1.4770	1.4896	1.4870
	GBM	1.1244	1.0950	1.1273	1.1208	1.0999
Error (%)		Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
RMSE	*LM*	2.1257	2.0863	2.1173	2.0792	2.0442
	GBM	1.8191	1.9407	1.8082	1.8123	1.7649
MAE	*LM*	1.4778	1.4615	1.4839	1.4658	1.4550
	GBM	1.1051	1.0955	1.1116	1.1012	1.0826

The effects of determinants on the timing of development are summarized in Table 7.9. The *AFT*s using linear specification are presented from column 1 to column 4. The marginal effects of determinants presented in column 5 and column 6 result from the *AFT*s using *GBM* specification. For *GBM*, the signs of marginal effects are the same as the signs of coefficients in linear models, but there are significant quantitative differences. For example, a 1% increase in potential future house price shortens the waiting time by 0.56 times; and a 1% increase in construction cost extends the waiting duration by 1.30 times. These results still show that the higher potential future house price leads to a shorter waiting period for development and the higher construction cost results in the longer duration. However, the effects of these determinants are surprisingly magnified. Meanwhile, it is unsure whether the average marginal effects statistically reflect the effects of determinants on the duration. Thus, hard to interpret is still a shortcoming for *GBM*. Such that, the coefficients from the *AFT* using the linear specification are focused when explaining the results below.

For general land characteristics, three outcomes could be summarized. Initially, a larger vacant land needs more time to wait for development. If the land size is 2.7 times larger, the waiting time for development is extended by around 1.04 times. Then, the duration of waiting to be developed is longer when the vacant land is located far away from the city. The waiting time is 1.09 times longer for every additional 10 kilometers away from the Perth city center. Thirdly, the purpose of ownership for the vacant land does not significantly affect the duration of waiting time.

Table 7.9: The factors' effect on the timing of development. summarizes the effect of factors on the timing of development. The coefficients of *GBM-AFT* model are average marginal effects, calculated by *ALE* method. The estimation is under the accelerated failure time model. The numbers in the parentheses are the standard errors of coefficients. The standard errors for *GBM* are generated using bootstrapping methods (20,000 times). These standard errors are only for reference and comparison, there is no inference purpose. Asterisk (*) means the variable is statistically significant at 0.05 significant level (only applicable for linear model).

Explanatory variable	Timing of development					
	AFT (linear)				AFT (*GBM*)	
	Exponential		Weibull			
Predicted potential house price	-4.4339 *	-4.9525 *	-2.6296 *	-2.8784 *	-65.7804	-51.8912
	(0.8043)	(1.4095)	(0.6717)	(1.2233)	(0.4721)	(0.3223)
Current construction cost	19.1128 *	31.2699 *	12.1482 *	20.1735 *	43.9133	8.8639
	(1.2585)	(3.0409)	(1.0598)	(2.6100)	(1.4611)	(1.2442)
House price growth	-452.6110 *	-766.6322 *	-314.8630 *	-515.1150 *	-5561.1939	-4864.7260
	(37.8182)	(105.5890)	(31.5763)	(91.5159)	(116.8959)	(99.3753)
Construction cost growth	-203.8520 *	-409.2445 *	-132.3460 *	-268.4950 *	-3468.0873	-4068.0423
	(20.2984)	(62.2245)	(16.7069)	(53.9293)	(44.7964)	(56.8244)
House price uncertainty	-15.4744	-37.0344	-14.3494	-26.7802	-1697.9376	-148.8489
	(17.0619)	(26.0892)	(14.6115)	(22.8797)	(67.7714)	(43.9798)
Construction cost uncertainty	-0.7509	-14.2346	-1.3684	-11.4720	-9594.2261	1770.8731
	(10.9344)	(25.4213)	(9.1630)	(21.2374)	(76.2907)	(33.7680)
Uncertainty of price growth	-	-360.5466	-	-206.5190	-	-9473.6935
	(-)	(199.4370)	(-)	(178.2780)	(-)	(114.0737)
Uncertainty of cost growth	-	381.5885 *	-	255.9110 *	-	16468.0157
	(-)	(110.5760)	(-)	(95.6967)	(-)	(169.4111)
Land size	0.0387 *	0.0368 *	0.0371 *	0.0366 *	13.1049	30.6118
	(0.0082)	(0.0082)	(0.0051)	(0.0051)	(0.8565)	(1.0409)
Distance to CBD	0.0085 *	0.0085 *	0.0082 *	0.0082 *	3.1470	1.1892
	(0.0008)	(0.0008)	(0.0005)	(0.0005)	(0.2364)	(0.2719)
Residential purpose	16.1522	15.7631	9.6331	9.3141	74.7233	76.3312
	(214.5840)	(207.9720)	(125.6150)	(120.6580)	(0.1626)	(0.1638)
Constant	-160.8000 *	-293.2088 *	-102.4450 *	-192.2870 *	-	-
	(10.1920)	(34.3656)	(8.5682)	(29.9555)	(-)	(-)
Quarter fixed effect	Yes	Yes	Yes	-	-	-
LGA fixed effect	Yes	Yes	Yes	-	-	-
N	139,579	139,579	139,579	139,579	139,579	139,579

Recall the real option theory (Eq. 7.8), the greater future sale prices of houses will boost the likelihood of developing vacant lands. Thus, the duration of waiting will be reduced. Meanwhile, the higher cost of construction harms the willingness of land development. A longer waiting time could be expected. From the first two rows in Table 7.9, all coefficients for predicted price present negative effects on the duration of waiting for development; all coefficients for current construction cost show positive effects. In column 1 and column 2, a 1% increase in predicted potential house price leads to around 0.95 times shorter waiting period of development. On the contrary, a 1% increase in current construction cost results in a longer duration of waiting by 1.29 times. For growth rates, both house price growth and construction cost growth would shorten the waiting time of development, the negative effects on duration are significant compared with other variables. The waiting time will be shortened by around 0.006 times when the growth rate of house price increase by 1%. This indicates that the willingness of landowners for development is extremely boosted when the expected long-term growth rate of house prices increases. Mostly, the landowners will immediately build a house on the owned vacant land without any doubts. Because they believe that house prices will increase in the long run. Meanwhile, when the growth rate of construction cost increases by 1%, the waiting duration will be made 0.08 times shorter. This reflects that the landowners will also develop their vacant land as soon as possible when the expected long-term growth rate of construction cost increases. Because they want to avoid the high construction cost when they believe the construction cost will be more and more expensive in the future.

For the timing of development, the uncertainties don't show significant effects on the duration, except for the uncertainty of cost growth. When there is a one-standard-deviation increase in the uncertainty of cost growth, the waiting time will be extended by 9.31 times. It means that the landowners are sensitive to the changes in their budgets for land development. When the uncertainties of cost growth are high, the construction process may not accurately follow the developing plan, and the spending may not be the same as the expected budget. Thus, assuming the landowners are rational and risk-averse, the higher risk of unstable construction cost growth leads to a longer waiting time for land development. The rest uncertainties shown in Table 7.9 don't present any statistically significant effects on the timing of development.

7.5.3 Robustness

As discussed in Section 7.3.2, unconditional mean and variance of growth rates for house prices and construction costs are the estimates for investigating the effects of growth rates on vacant land value and timing of development. Because the growth rates are assumed as constants in the theory of real options (Eq. 7.6). The unconditional mean measures the potential long-term growth rates. And the unconditional variance measures the potential uncertainties, the growth rates may expose in the long run. For thoroughly considering the measures of growth rates and their uncertainties, conditional mean and variance are taken into account as alternatives, that could present what will happen in the short future.

When the landowners are unable to have a full foresight of the market trends for house prices and construction costs, the potential growth rates and their uncertainties in the coming year may be the best measure. Because the landowners mostly will worry about the market trends and the uncertainties during the construction period, if they decide to develop their lands now. One-step-ahead predictions and their variances are calculated in Eq. 7.26 and

Eq. 7.27, those are the conditional means and variances of growth rates for house prices and construction costs.

$$\begin{bmatrix} E(\Delta p_{t+1}|t) \\ E(\Delta c_{t+1}|t) \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix} + \begin{bmatrix} \alpha_{p,1} & \alpha_{c,1} \\ \beta_{p,1} & \beta_{c,1} \end{bmatrix} \begin{bmatrix} \Delta p_t \\ \Delta c_t \end{bmatrix} \tag{7.26}$$

and

$$\begin{aligned} vec\Bigg(&\begin{bmatrix} Var(\Delta p_{t+1}|t) & Cov(\Delta p, \Delta c|t) \\ Cov(\Delta p, \Delta c|t) & Var(\Delta c_{t+1}|t) \end{bmatrix} \Bigg) \\ &= vec\Bigg(\begin{bmatrix} Var(\epsilon_p|t) & Cov(\epsilon_p, \epsilon_c|t) \\ Cov(\epsilon_p, \epsilon_c|t) & Var(\epsilon_c|t) \end{bmatrix} \Bigg) \end{aligned} \tag{7.27}$$

Table 7.10 includes all tests for vacant land valuation and timing of development when the alternatives, conditional means, and variances, are applied. *GBM* is not examined, due to lack of interpretability. Several findings are consistent, including, for valuation, the increases in the growth rate and uncertainty of construction cost boost the value of vacant land, meanwhile, the uncertainties of growth rates show the negative effects on land values; for the timing of development, the higher potential house price and the price growth shorten the waiting period for land development, however, higher current construction cost and uncertainty of construction cost growth delay the timing of development.

There are also some inconsistent results. For valuation, the growth of house prices doesn't have a statistically significant effect on vacant land value when conditional means and uncertainties of growth rates are applied, no matter whether the uncertainties of growth rates are included. Compared to applying unconditional means and uncertainties, the growth of house prices presents a significant positive effect. This may indicate that the landowner may care more about the growth of house prices in the long run rather than in the short run when building a house is a long-term investment. For timing of development, the growth rate of construction cost doesn't consistently shorten the waiting time. Additionally, house price uncertainty extends the waiting time, which means it decreases the likelihood of development. This finding is in line with the results in *Cunningham2006*. Meanwhile, higher uncertainty of construction cost delays the timing of development, and higher uncertainty of price growth leads to covert the land earlier. Overall, some determinants present significant effects on land value or timing of development when the measure for growth rates and their uncertainties is on the short-term aspect; some are not.

In summary, the key findings are still highly robust. For valuation, the parameter estimates associated with the growth of construction cost and the uncertainty of construction cost are positive and significant for different measures of growth and its uncertainty. In addition, the uncertainties of growth rates for house prices and construction costs are essential to be considered, which indicates the assumption of constant growth rates may not be valid. For the timing of development, four variables are essential. Potential house prices and growth of house prices shorten the waiting time. Current construction cost and uncertainty of construction cost growth deter the development of vacant land. The significance of house price growth reflects that the landowners care about the future growth that shows the potential benefits from the investment output in the long term. Meanwhile, the landowners worry about the uncertainty involved in the construction growth during the building process when they decide to convert the lands, which is reflected by the significance of uncertainty of construction cost growth. Because the construction cost is not a one-off payment during the building process, the uncertainty in the cost growth may be more likely to represent the uncertainty involved on the cost side.

Table 7.10: The factors' effect for checking robustness. summarizes the effect of factors on the valuation and the timing of development for robustness.

Explanatory variable	Valution		Timing of development			
	Linear		AFT			
			Exponential		Weibull	
Predicted potential house price	-	-	-13.4307 *	-15.0976 *	-8.7403 *	-10.0102 *
	(-)	(-)	(0.3124)	(0.5342)	(0.2130)	(0.3745)
Current construction cost	-	-	24.2784 *	26.2838 *	15.5534 *	17.0890 *
	(-)	(-)	(0.5319)	(0.8183)	(0.3583)	(0.5710)
House price growth	-0.3020	-0.1295	-143.3950 *	-161.9490 *	-107.4970 *	-119.6850 *
	(0.1524)	(0.1542)	(2.6357)	(4.0815)	(1.8247)	(2.8147)
Construction cost growth	1.0067 *	0.9879 *	9.0438 *	0.7329	3.5350	-3.4478
	(0.1316)	(0.1316)	(3.2857)	(5.0329)	(2.1168)	(3.3839)
House price uncertainty	-0.0124	0.4800 *	66.3328 *	61.9250 *	45.1415 *	40.4386 *
	(0.1280)	(0.1443)	(4.5214)	(4.6775)	(3.0310)	(3.1272)
Construction cost uncertainty	0.3124 *	0.8481 *	-7.4524	39.3260 *	-2.9165	27.4231 *
	(0.1033)	(0.1116)	(3.1004)	(5.9072)	(2.0184)	(3.8802)
Uncertainty of price growth	-	-5.0617 *	-	-531.9470 *	-	-334.6020 *
	(-)	(0.6645)	(-)	(27.7047)	(-)	(18.6950)
Uncertainty of cost growth	-	-6.7689 *	-	234.8970 *	-	131.9550 *
	(-)	(0.5170)	(-)	(16.6421)	(-)	(10.9128)
Land size	0.3923 *	0.3927 *	0.0404 *	0.0380 *	0.0037 *	0.0366 *
	(0.0014)	(0.0014)	(0.0082)	(0.0082)	(0.0005)	(0.0051)
Distance to CBD	-0.0131 *	-0.0132 *	0.0084 *	0.0084 *	0.0082 *	0.0082 *
	(0.0001)	(0.0001)	(0.0008)	(0.0008)	(0.0005)	(0.0005)
Residential purpose	0.0689 *	0.0719 *	16.6971	16.4185	10.4348 *	10.2725
	(0.0040)	(0.0040)	(204.3130)	(209.0300)	(147.3330)	(155.0070)
Constant	11.3179 *	11.4845 *	-130.2360 *	-127.2200 *	-80.3468 *	-78.6708 *
	(0.0159)	(0.0218)	(4.1168)	(5.1362)	(2.7371)	(3.4127)
N	145,858	145,858	139,579	139,579	139,579	139,579

7.6 Conclusions

This chapter applies the theory of real options with different models, including machine learning techniques, that is to study the research questions about vacant land valuation and optimal timing of development. The empirical analysis considers all determinants involved in the theory on both sides of investment output and investment cost. The cost side has not gotten enough attention before. Moreover, this chapter also takes the uncertainties of growth rates into account, when the assumption of constant growth rates is not valid. The results are summarized for the vacant land valuation and the optimal timing of development respectively. Generally, the results present that the effects from the side of investment cost (construction cost) should be considered on land value and optimal timing.

For valuation, valuation models with a reduced form are applied with the variables suggested from the theory of real options, such as growth rates and uncertainties on the investment output and investment cost sides. Meanwhile, the common land characteristics are involved, for instance, land size and distance from the city center. The results robustly show that higher uncertainty of construction cost could lead to a higher land value initially, which is consistent with the real option theory. However, the finding for the statistical insignificancy of house price uncertainty is not similar to that of (Cunningham 2006) and (Bulan et al. 2009). Most importantly, the uncertainties for growth rates for house prices and construction costs present significant negative effects on the value of vacant land. This indicates that the unstable growth will lead to a drop in land value, no matter whether the unstable growth is on the investment output side or the investment cost side. This finding complements the results from the theory of real options when the assumption of constant growth rates is not valid.

For the timing of development, the duration of waiting time is directly estimated by the accelerated failure time model, rather than estimating the hazard of development. The greater potential future sale prices of houses lead to a shorter waiting duration. Meanwhile, a longer waiting time could be expected, when the current construction costs are higher. These two findings are in line with the theoretical findings in real options, and they are robust to different measures for growth and uncertainty. For growth rates, only house price growth would consistently and significantly shorten the waiting time of development, which means the landowners will immediately build a house on the owned vacant land without any doubts when the house price growth is high. For uncertainties, the uncertainties don't consistently show significant effects on the duration, except for the uncertainty of cost growth. It means that the landowners are sensitive to the changes in their budgets for land development. When the uncertainties of cost growth are high, the construction cost may not accurately follow the plan, and the expected budget may not cover the spending. Due to that building a house is a term project, the budget becomes more sensitive to cost growth. All these results suggest that landowners do account for real options. The presence of real options in the Perth land market has a broader implication that the consideration of the growth changes should be included on the investment output and investment cost sides.

Chapter 8

Conclusion

8.1 Introduction

There has been a growing recognition that machine learning techniques could benefit the valuation of residential properties, especially for valuation accuracy. However, the potential issues are neglected when implementing the valuation process. For instance, in recent literature, real estate data are trimmed beautifully. However, the data issues are rarely discussed; and missing values and outliers are simply avoided sometimes, which means a loss of observations or housing characteristics. Consequently, the results in some academic research are well "pruned". However, in practice, these data issues can not be neglected. Such that, are those results in some previous research the truth? This is the motivation of the first topic for this thesis. It starts with solving the data issues in transaction-level data for residential properties.

In Chapter 5, the best combination of the estimation model and the solution for data issues is found. The implementation does improve the accuracy for valuation and meanwhile, it solves the data issues perfectly. At this stage, additional questions come out, "whether the accurate price prediction could benefit the other fields" and "whether the time effects could be interpreted in machine learning models". To address these two questions, the application of machine learning techniques is extended to the field of residential property price index using the best combination in Chapter 5. Because, the residential property price index ($RPPI$) is one of the derivatives of the valuation models, and it requires explaining the time effects in the valuation models, such as the time dummy approach. Chapter 6 primarily presents the combination of the valuation model with machine learning techniques and classic hedonic indexing approaches. Meanwhile, the ability to interpret machine learning techniques is examined, as the time effects must be extracted from the machine learning implementation for compiling a price index.

After thoroughly investigating the valuation and price index for residential buildings, it is the turn of vacant lands. Chapter 7 discusses the valuation of vacant lands via real option theory. Land development is treated as a real option of an investment, and construction cost is the investment cost, the sale price of the built dwelling is the investment output. At this stage, the land value could be roughly equal to the present value of the subtraction of the house price and the construction cost when the vacant land is developed. Because of this, besides the characteristics of vacant lands, the valuation of land could also take the changes in house prices and construction costs into account. However, in previous studies, the contribution of construction cost is rarely discussed. In addition, according to the findings of the Australian Bureau of Statistics, the final cost of construction can differ from initial expectations at the building approval stage and the start of construction when building a new dwelling. Thus, it is necessary to comprehensively study the valuation of vacant lands. The optimal timing of development is evaluated as well.

Overall, this thesis addresses the research gaps in three fields: 1. The solutions for the faced data issues when conducting valuation models; 2. Testing the benefits of the prediction accuracy on indexing and machine learning interpretation; 3. Valuation of vacant lands including the effects of house prices and construction costs. The research findings are summarized firstly in this chapter (Section 8.2). It then highlights the strengths and contributions of this thesis (Section 8.3). The limitations of this research are discussed in Section 8.4, along with the potential direction for future studies.

8.2 Summary of research findings

The research gaps in the three topics addressed in this thesis are progressively related. The models applied in Chapter 6 for compiling indices are the best combination of the estimation model and the solution for data issues found in Chapter 5. The changes in house prices used in Chapter 7 are measured by the indexing methods studied in Chapter 6. Thus, the research findings are summarized respectively. Additionally, the findings for machine learning techniques, which are concluded from three topics, are summarized separately in this section.

8.2.1 Price prediction with data issues

One of the most headache data issues is solved using missing value node strategy in Chapter 5. Thus, missing values may not have to be arbitrarily or experientially removed or imputed. The missing value node strategy allows not only missing values in the training sets for building models but also allows missing values to appear in the testing sets after the estimation model is built. Such that, this strategy achieves that the users only need to regularly maintain the estimation model, which truly implements the *RICS*'s definition of *AVM*, "... without or with less human intervention". Compared with the complete case strategy, it prevents a potential sizeable data loss. Such that, the automated valuation model applying this strategy could comprehensively evaluate the price of residential properties, rather than only estimate the information from the complete case. Compared with the imputation strategy, it has a simple estimation process and prediction process. Because, the imputation strategy does not only require imputing missing values in training sets, but also fulfilling the missing information using the same imputation model in the testing sets. In addition, the missing value node strategy could also robust a special case, when missing values initially appear in the testing sets. It means, there is no missing value when the model was established, but there are missing values when applying the built model. In this situation, normal *AVM*s may not provide any predictions due to missing values. But *AVM* using the missing value node strategy could overcome this disadvantage. Even if there are missing values in the coming cases, the predictions could be provided. However, the accuracy of the predictions drops, because, the cases with missing haven't been well studied before. Overall, the missing value node strategy is reliable when missing values occur in transaction-level real estate data.

8.2.2 Residential property price indices

The examination of five indices conducted in Chapter 6 presents that more accurate predictions could benefit the accuracy of indices when the hedonic imputation approach is applied. However, the improvement is not as significant as the improvement in the accuracy of price predictions. This answers one of the research questions for the second topic, "whether the accurate price prediction could benefit the other fields". This finding is similar to the results in Hill and Scholz (2018). The results present that *GBM* provides the most accurate price predictions, *GAM* closely follows, and *LM* is the last. This is the same as the results exhibited in Chapter 5. However, the differences between the qualities of indices are not that significant. This was proven in the investigation of contemporaneous correlation between indices. Quarterly growth rates are calculated from each index, and they are pairwise compared. All the correlation coefficients between indices are around or above 0.9, and statistically significant. Such that, the difference in quarterly growth is negligible, which confirms the results by Shimizu, Nishimura and Watanabe (2010). For index revision, it is still an issue for indices compiled by the time dummy approach. The issue of index revision, however, is not fatal. The index compiled by the machine learning technique suffers slightly more significant revision than the index compiled by the linear model. Meanwhile, the index compiled by the machine learning technique has a slightly downward revision tendency, which means the revised index number is normally less than the index number before revision. Taking all into account, machine learning techniques are empirically pursuable for indexing purposes, using the hedonic imputation approach or the time dummy approach, however, the drawbacks of the time dummy approach can not be overcome.

8.2.3 Vacant land valuation and timing of development

In Chapter 7, for valuation, a reduced form approach is applied with the variables suggested in the real option theory, such as growth rates and uncertainties on the “investment output” side and the “investment cost” sides. The option premium seems to rely on the “investment cost” side more. The results robustly show that a higher uncertainty of construction costs could lead to a higher land value initially, which is consistent with the real option theory. However, the uncertainty of house prices does not present a statistically significant effect on the value of vacant land, which is not similar to the findings of (Bulan et al. 2009) and (Quigg 1993). But different from the previous studies, the uncertainties of growth rates for house prices and construction costs are included in the reduced form. These uncertainties present negative effects on land values. This indicates that the unstable growth will lead to a drop in land value, no matter whether the instability is on the "investment output" side or the "investment cost" side. This finding complements the results from real options theory when the assumption of constant growth rates is not valid.

For the timing of development, the greater potential future sale prices of houses lead to a shorter waiting duration. Meanwhile, a longer waiting time could be expected, when the current construction costs are higher. These two findings are in line with real options theory, and they are robust to different measures for growth and uncertainty. For growth rates, only house price growth would consistently and significantly shorten the waiting time of development. It indicates that the landowners will immediately build a house on the owned vacant land without any doubts when the house price growth is high. The uncertainties don't consistently show significant effects on the duration, except for the uncertainty of cost growth. It means that the landowners are sensitive to the changes in their budgets for land development, especially when the construction cost is unpredictable. All these results also present that the "investment cost" side is necessary to be considered, no matter for vacant land valuation or optimal timing of development. The presence of real options in land markets is further evidence for the necessary inclusion of real options in models of capital investment and has broader implications for including the consideration of the growth changes on the “investment output” and “investment cost” sides.

8.2.4 Machine learning applications

Machine learning techniques are applied through the three topics in this thesis for solving data issues, compiling price indices, and evaluating the land value and optimal timing of development. They do present their potential. Firstly, machine learning techniques could significantly improve the accuracy of predictions, no matter whether they are applied for valuation analysis or survival analysis. The reason for this improvement could be the recursive partitioning approach of the tree-based machine learning model. This approach allows to investigation of more complex interactions between variables, that the classic parametric model can not. Meanwhile, the tree structure could isolate missing values or outliers into some specific nodes, and these values are “learned” within the nodes.

Additionally, the ability of interpretation for machine learning techniques is examined in Chapter 6 and Chapter 7. Compiling the price index requires the time effects to be extracted from the model in Chapter 6. The accumulated local effects method is applied to calculate the average marginal effects of the temporal variable. The price index compiled by the marginal effects is competent, its quality is as good as the official price index published by the Australian Bureau of Statistics. However, in Chapter 7, the same approach is applied to explain the effects of growth and uncertainties that could affect the land value and optimal timing of development. The signs of the coefficients for the independent variables are mostly consistent with the signs for the same variables in the linear models. However, the magnitude of the coefficients is significantly different from the outcomes from the linear models. Moreover, those marginal effects lack statistical inference, so it is hard to investigate whether the marginal effects are statistically significant. Thus, from those two applications, it could be concluded that the machine learning techniques could be interpreted by extracting the average marginal effects of target independent variables. Those marginal effects could be a competent alternative in some cases depending on purposes. However, the application of this model-interpretation method is limited, one important reason is the absence of statistical inference in the outcomes.

8.3 Research contributions

This thesis serves as a comprehensive piece of progressively related topics for addressing different research gaps, applying the recommended machine learning method from the recent literature, and using transaction-level data with the common issues of real estate data. Overall, there are three main contributions.

Firstly, this thesis contributes to the valuation aspect of residential properties. The data issues, that may impede the valuation process, are focused and investigated. The diverse combination of the recommended machine learning method, missing value strategies, and loss functions are implemented. The solution is provided, it has the best accuracy of predictions, and the whole valuation process is automated and the most efficient, when data issues occur. Meanwhile, it provides the foundation for the second topic.

Then, the residential property price index can be compiled using the machine learning technique, especially using the time dummy approach proposed previously. This fills the research gap in the price index field. At the same time, data issues are perfectly solved. Thus, there is no need to conduct a complex process for indexing, such as using a batch of similar models in (Hill and Scholz 2018). In practice, this thesis provides a competent alternative for the *RPPI* published by *ABS*, which is more efficient and easier to maintain. Additionally, the ability to interpret the machine learning model is examined initially. Calculating the average marginal effect for the target independent variable could be a considerable solution. This topic tests and presents the indexing methods that could be applied in the third topic for measuring the growth and uncertainty of market prices for houses. Also, it gives confidence for calculating the average marginal effect as an alternative method to interpret machine learning techniques.

Thirdly, another prevailing contribution of this thesis is to investigate the land development including vacant land valuation and optimal timing of development. Land development is the beginning of a parcel's life cycle. The factors that could affect the decision and timing of land development are considered on the two sides, both of the investment output side (house prices) and the construction cost side (construction cost). The thesis includes the effects from the construction cost side in the reduced form for evaluating vacant land and optimizing the timing of development. In previous literature, the effects from the construction cost side and the uncertainties of growth rates are not well and comprehensively discussed. Thus, this thesis contributes to the empirical analysis of land development, and the findings could more reliably and accurately corroborate the real option theory.

8.4 Research limitations and future research

This thesis is subject to several limitations, which could be enhanced in future studies. The first limitation is related to that only one of the tree-based machine learning techniques is applied through three topics. With the development of generative artificial intelligence, the other machine learning or artificial intelligence models could be more competent and more explainable than the gradient boosting machine. Therefore, those machine learning models should be considered in future research. The second is about missing value strategies. The missing value node strategy is only designed for tree-based machine learning models. The other types of machine learning techniques may require a data adjustment or restructure or other specific model designs. Thus, this missing value strategy could not be commonly used, it has its limitations, compared with the strategies of using complete cases or imputation. A more general missing value strategy may be worthwhile to investigate in the future. The third limitation is about land development identification in the third topic. The whole process of building activities includes applying for building approval and reporting the construction work done. Simply checking two consecutive transactions of the same lot may not be rigorous.

Apart from the aforementioned limitations for future research, another interesting topic related to text and image analysis using machine learning is also worth investigating. Because, except transaction-level data, there is also a lot of information available in the official documents, such as valuer reports, blueprints of house design, and images of the dwellings. Those types of information can not be used under the classic approaches, such as the hedonic pricing approach. However, they are more detailed and relevant than the cold numbers, especially the numbers may be missing for some reasons. An innovation in this field may probably change the way of valuation for residential properties.

Appendix A

For Chapter 3: Data

A.1 ***Statistical Area* VS *LGA***

Generally, there are two structures of geographical classifications available, that are documented in the Australian Statistical Geography Standard (ASGS)[1]. The ASGS is split into two parts, the *Australian Bureau of Statistics* (*ABS*) and Non *ABS* structures. The *ABS* structures are geographies that the ABS designs specifically for the release and analysis of statistics. This means that the statistical areas are designed to meet the requirements of statistical collections as well as geographic concepts relevant to those statistics. The *ABS* structures include several geographies that approximate urban areas, and these may differ from official or commonly accepted definitions. The following lists the levels of statistical areas within one city metropolitan area.

- Mesh Blocks (*MB*s) are the smallest geographic areas defined by the ABS and form the building blocks for the larger regions of the ASGS. Most Mesh Blocks contain 30 to 60 dwellings.
- Statistical Areas Level 1 (*SA1*s) are designed to maximize the geographic detail available for Census of Population and Housing data while maintaining confidentiality. Most *SA1*s have a population of from 200 to 800 people.
- Statistical Areas Level 2 (*SA2*s) are medium-sized general-purpose areas built to represent communities that interact together socially and

[1] The ASGS is a classification of Australia into a hierarchy of statistical areas. It is a social geography, developed to reflect the location of people and communities, and is updated every five years to account for growth and change in Australia's population, economy, and infrastructure.

economically. Most *SA2*s have a population range of 3,000 to 25,000 people.

- Statistical Areas Level 3 (*SA3*s) are designed for the output of regional data and most have populations between 30,000 and 130,000 people.
- Statistical Areas Level 4 (*SA4*s) are designed for the output of a variety of regional data, and represent labour markets and the functional area of Australian capital cities. Most *SA4*s have a population of over 100,000 people.

The Non *ABS* structures generally represent administrative regions that are not defined or maintained by the *ABS*, but for which the *ABS* is committed to directly providing a range of statistics. All Non *ABS* structures are approximated using ASGS regions. Because of this, they are only cartographic representations of legally (or otherwise) designated boundaries, for example, Local Government Areas.

- Local Government Areas (LGA) are an *ABS* Mesh Block representation of gazetted local government boundaries as defined by each state and territory.
- Suburbs and Localities (formerly State Suburbs) is an *ABS* Mesh Block approximation of gazetted localities.

For larger regions in Perth, such as submarkets, it is decided that *SA4*s are used for defining the geographical areas. Firstly, *SA4*s are designed for this purpose. Additionally, this is also suggestions about Perth's housing submarkets from *REIWA*, which is Western Australia's real estate institute, the peak body for the real estate profession in the state. For smaller regions in Perth, *LGA*s and suburbs are used as location identifiers. Because, initially, they are available in the *Landgate* data set. Then, they are more widely acceptable for residents in Perth, rather than *SA2*s and *SA3*s, those the residents are not familiar with the definitions and the boundaries. *SA4*s are the larger areas that include several LGAs. The relationship of *SA4*s and *LGA*s in Perth are defined in *ABS* and *REIWA* as follows. **Central**: Cambridge, Claremont, Cottesloe, Mosman Park, Nedlands, Peppermint Grove, Perth City, Subiaco and Vincent. **Northeast**: Bayswater, Bassendean, Mundaring, and Swan. **Northwest**: Joondalup, Stirling and Wanneroo. **Southeast**: Armadale, Belmont, Canning, Gosnells, Kalamunda, Serpentine-Jarrahdale, South Perth and Victoria Park. **Southwest**: Cockburn, East Fremantle, Fremantle, Kwinana, Melville and Rockingham.

A.2 Price boundaries plots

Figure A.1 depicts the minimum and maximum prices for houses and units in the other four *SA4*s from 1994 to 2020, plot **A** to plot **D** show the minimum and maximum prices in the **Northeast**, **Northwest**, **Southeast** and **Southwest** respectively. All four plots show the same characteristics. The difference in minimum prices between houses and units is negligible, both of them have an upward trend. For the maximum prices, the climbing trend is still clear. The gap between houses and units is much larger, which means houses could be much more expensive than units. In addition, the minimum and maximum prices are not significantly different among *SA4*s. The logarithm of the lowest transaction price in all places was around 10 in 1994 and rose to 12 in 2020. The highest prices are within the interval from 12 to 16 through the 27-year period.

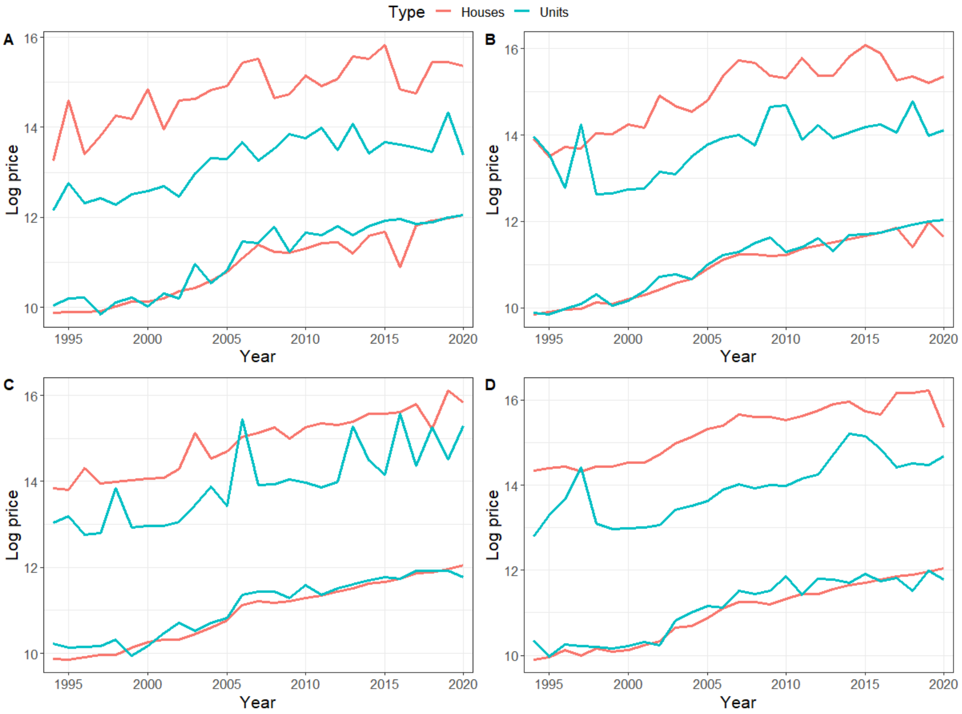

Figure A.1: The minimum and maximum of transaction prices in the other four *SA4*s from 1994 to 2020. To depict the minimums and maximums in one graph, the transaction prices are under a logarithm scale.

A.3 Property groups and classes

Table A.1 shows three main groups of properties, vacant lands, residential dwellings, and non-residential buildings, available in the *Landgate* data. The definitions of vacant lands, residential dwellings, and non-residential buildings are clarified in the Functional Classification of Building Structure 1999 (FCB, revision 2011) and the Census Dictionary (2016). All the property classes are listed in Table A.1, and the property types with an asterisk belong to the residential properties, the property types with two asterisks are vacant lands, and the rest are the non-residential properties, such as commercial properties.

Table A.1: The summary of property classes and groups. reports the property classes in the *Landgate* data. The property class with one asterisk means the residential property class we pick. The property class with two asterisks means vacant land. The number of total buildings excludes vacant lands and blanks.

Residentials (*)					
Vacant lands (**)					
Types					
(Blanks)	Add service	Aged care	Amenities	Ancillary	Apartment (*)
Army	Art gallery	Bakery	Bank	Bed sit	Boat shed
Bore	Bulk conveyor	Business	Camp	Car bay	Caraven park
Caretaker	Cement works	Center	Chalet	Church	Clinic
Club	Cluster house	Cold store	College	Community center	Converted house
Convent	Cottage	Dance studio	Daycare center	Dental surgery	Depot
Detached (*)	Disused building	Drain	Dry cleaner	Duplex (*)	Factory
Farm	Fastfood outlet	Fire station	Flat (*)	Foodhall	Foreshore
Foundry	Function center	Funeral parlor	Garage	Garden	Golf course
Group house (*)	Hairdresser	Hall	Health studio	Holiday unit (*)	Home unit (*)
Hostel	Hotel	House (*)	Ice rink	Indoor sport	Infant health
Joinery	Kennel	Kiosk	Laboratory	Laundry	Liquor store
Lookout	Manse	Market	Medical center	Mission school	Motel
Motor wrecks	Multiplex (*)	Museum	Nursery	Office	Orchard
P.A.W.	Pad mount	Paddock	Park	Patio house (*)	Penthouse (*)
Pharmacy	Physio	Piggery	Pipe line	Playground	Playing field
Police station	Post office	Poultry farm	Pre Primary	Presbytery	Primary school
Psych hostel	Public hall	Pump station	Quadruplex (*)	Quarantine	Quarry
Quarters	R.O.W.	Raaf institute	Recreation center	Rectory	Regional park
Reserve	Rest home	Restaurant	Retirement village (*)	River bank	Road widening
Row house	Sale yard	School	Semi-detached (*)	Service station	Shed
Shoping center	Shops	Showroom	Slipway	Squash court	Stables
Stand pipe	Storage tank	Storeroom	Studio	Sub station	Sump
Supermarket	Seperseded	Surgery	Tab agency	Tavern	Temporary service
Tennis club	Tennis court	Terrace house (*)	Town house (*)	Transportable	Triplex (*)
Turf farm	Vacant land (**)	Vet	Villa house (*)	Vineyard	Warehouse
Water supply	Welfare center	Workshop	Yard	Zoo	

Appendix B

For Chapter 5: Residential property valuation

B.1 The comparison of predictive accuracy

This section briefly discussed the predictive accuracy for the commonly used models, that are mentioned in the previous literature (Mayer et al. 2019, Schulz and Wersing 2021, for example). In this comparison, the gradient boosting machine (GBM), random forests (RF), the generalized additive model with spline GAM, and the neural network (NN) are included. For each model, we back-transform log predictions to a natural scale and calculate the percentage error.

The generalized additive model allows a spline to measure the geographical information more precisely. Hill and Scholz (2018) discover that the geospatial spline improves the predictive performance of valuation. In this section, the baseline spline model is applied, and the geographical information (longitudes and latitudes) is captured by the geospatial spline. The other housing characteristics are in the parametric part with an assumption of linearity. The specification of GAM is

$$p_i = \mathbf{x}_i\boldsymbol{\beta} + f(x_{lon}, x_{lat}) + \epsilon, \tag{B.1}$$

where $\mathbf{x}_i$ excludes the coordinates.

The tree-based machine learning models are easier to build by R packages, *ranger* is for Random Forest; *gbm* is for Gradient Boosting Machine. The Random Forest has two pruning variables, the number of trees and the number of randomly selected predictors (mtry). Because, Random Forest takes a long time to provide results, and the extremely large number of trees does not benefit the predictive accuracy. Such that, 2000 trees are set as default in

Random Forest. Similar to Random Forest, *GBM* requires tuning three main hyperparameters, interactive depth, number of trees, and shrinkage separately. The Neural Network is achieved by the R package (**h2o**). There are much more hyperparameters, that need to be tuned, than the tree-based models. The most influenced two of them are the learning rate and the architecture of hidden layers. The learning rate determines the size of the correction step to adjust for errors when the model is learning from a new observation. A high learning rate means an aggressive correction, which will shorten the learning time, but with lower accuracy; and vice versa. The architecture of hidden layers defines the complexity of a neural network. It contains the number of hidden layers and the number of neurons on each hidden layer. All the hyperparameters will be tuned by K-fold cross-validation[1] (R package, *caret*). The optimal hyperparameters for each model are selected.

The metrics of percentage errors are used to compare the predictive accuracy of models, such as Mean percentage error (*MeanPE*), Mean Absolute Relative Error (*MAPE*), Root Mean Squared Percentage Error (*RMSPE*), and Percentage Error Ranges (10 and 20). The quarterly rolling window outputs and the weekly rolling window outputs are presented separately, see Table B.1.

Table B.1 shows the summary of errors in each model during the full testing period (2016 – 2020, 68,188 observations[2]). The tree-based models (*GBM* and *RF*) are generally more accurate than the generalized additive model with spline and neural network, and all of them outperformed the benchmark.

[1] The default setting of K-fold cross-validation is 10 folds of data and 5 repeats of the procedure.

[2] This is for testing the accuracy only. Thus, only residential detached houses are kept in the data, meanwhile without any missing values.

Table B.1: The Summary of the predictive performance of models compares the mean absolute percentage error (*MAPE*), root mean squared percentage error (*RMSPE*), and percentage error ranges (*PER(10)* and *PER(20)*) for residential houses (68,188 observations).

Model		MAPE	RMSPE	PER(10)	PER(20)
Linear	Weekly	29.47	40.96	75.83	54.54
	Quarterly	29.63	41.21	75.85	54.53
GBM	Weekly	13.62	20.42	48.72	19.70
	Quarterly	13.71	20.59	48.95	19.90
RF	Weekly	13.82	20.81	48.57	20.41
	Quarterly	13.94	21.00	48.99	20.59
NN	Weekly	16.21	24.06	55.85	26.81
	Quarterly	16.26	24.08	55.85	26.81
GAM	Weekly	15.30	22.85	53.82	24.44
	Quarterly	15.34	23.30	53.83	24.42

Moreover, there are around 49% and 20% of predictive errors in the tree-based models, those are within the 10% and 20% of predicted prices. In the tree-based models, *GBM* is slightly better than Random Forest. It shows lower errors in all metrics.

B.2 Preparation for *AVM* implementations

B.2.1 Tuning parameters

The hyperparameters of the gradient boosting machine (*GBM*) are tuned by grid searching on ten-fold cross-validation. The optimal hyperparameters are selected depending on the comprehensive performance ranking (mean absolute error and root mean squared error) when *GBM* applies different loss functions, the squared error loss function (*LS*) and the absolute deviation loss function (*LAD*). The most conservative model is chosen when two sets of hyperparameters have the same ranking. Table B.2 reports the tuning parameters for *GBM*.

Table B.2: The overview of tuning parameters. shows the hyperparameters for GBMs. Tuning criteria gives the estimator for MAE and RMSE which are minimized to choose the optimal hyperparameters. The other parameters in the model use the default settings in R package, such as the default value of bag.fraction is 0.5 in the package and recommended in Friedman (2002). The definition of parameters could refer to the R document of package, it is available on https://cran.r-project.org/.

Method	Hyper-parameter	Tuning criteria

GBM	n.minobsinnode = {5, 10, 15},
	shrinkage = {0.01, 0.05, 0.1},
	interaction.depth = {4, 6, 8, 10, 12, 14},
	n.trees = 2,000.
		Cross-Validation

B.2.2 R packages

The analysis in Chapter 5 is conducted using various R packages in Ubuntu 18.04 LTS, see Table B.3. greybox (Svetunkov 2021) and gbm (Greenwell et al. 2020) are commonly used R packages for linear models and *GBM*s using different loss functions. missRanger (Mayer 2021) is an alternative implementation of the “missForest” algorithm using ranger to build chaining random forests for imputing missing values. The multiple imputation strategy is achieved using that package. There are five imputed datasets generated. leaps (Lumley

2020) is to implement the subset selection for selecting the optimal form of linear hedonic model. All packages can be downloaded from https://cran.r-project.org/.

Table B.3: The overview of functions and packages used in R. shows R packages and the functions used to implement models or strategies.

Method	Package	Function
Hedonic linear	greybox	alm
GBM	gbm	gbm
Multiple imputation	missRanger	missRanger
Subset selection	leaps	regsubsets

B.3 Additional Results

For comprehensively exploring GBM potential, the predictive performance of non-stochastic GBMs ((LS, $Non-stochastic$) & (LAD, $Non-stochastic$)) are investigated, which do not sub-sample training sets in each iteration. In addition, for comprehensively comparing the GBMs ($LS\&LAD$) using missing value node strategy and the GBMs ((LS, C) and (LAD, C)) using complete case strategy, the predictive accuracy of the complete cases in the test samples are summarized in Table B.4, that could explain why there is an accuracy drop when the GBMs ((LS, C) and (LAD, C)) predict testing samples with missing values.

In Table B.4, it shows that the GBMs ((LS, $Non-stochastic$) & GBM (LAD, $Non-stochastic$)) also have a competent predictive accuracy. Such that, it does not matter whether using the non-stochastic GBM or using the

Table B.4: The predictive performance of complementary GBMs. presents the evaluation metrics of each model. The performance ranking follows the rule, the lower the error value the system has, the more accurate and better it is.

Model	MeanPE	MedianPE	MAPE	RMSPE	PER(10)	PER(20)
All test samples ($N = 113,104$)						
GBM (*LS*, *Nonstochastic*)	0.70	0.28	11.81	17.72	42.82	15.57
GBM (*LAD*, *Nonstochastic*)	0.79	0.12	11.90	18.67	42.29	15.48
All test samples ($N = 71,944$) without missing values						
GBM (*LS*)	0.50	0.20	11.05	16.43	40.33	13.39
GBM (*LAD*)	0.60	0.15	10.91	17.04	38.71	13.00
GBM (*LS*, *C*)	0.51	0.19	11.06	16.73	40.22	13.47
GBM (*LAD*, *C*)	0.64	0.16	10.94	17.17	38.70	13.11

stochastic *GBM*. When the test samples have no missing value, the predictive accuracy of the *GBM*s ((LS, C) and (LAD, C)) is also considerable. The accuracy gap is negligible when they are compared with the *GBM*s (LS&LAD). Thus, the drop in predictive accuracy for *GBM*s ((LS, C) and (LAD, C)) in Table 5.4 is caused by the cases that contain missing values in the test sample. This proves that the *GBM*s using missing value node strategy could not accurately provide the predictions for the testing samples containing missing values if they have not learned missing values thoroughly before (training samples). Overall, the additional results complement the analysis in Chapter 5, and they won't affect the conclusions, the *GBM* (*LAD*) is still the most recommended *AVM* implementation.

Appendix C

For Chapter 6: Residential Property Price Index

C.1 The other benefit of machine learning techniques

The TD_{GBM} index applied ALE to explain the time effects. The performance of the compiled index is impressive. It could measure the market dynamics as the official $RPPI$ published by ABS (Chapter 6). Meanwhile, the implementation also has a benefit because of ALE. Normally, the time dummy approach allows the time trend of the market prices to be measured at an aggregative level. It is hard to compile a sequence of $RPPI$s for submarkets, such as in specific LGAs or within property classes.

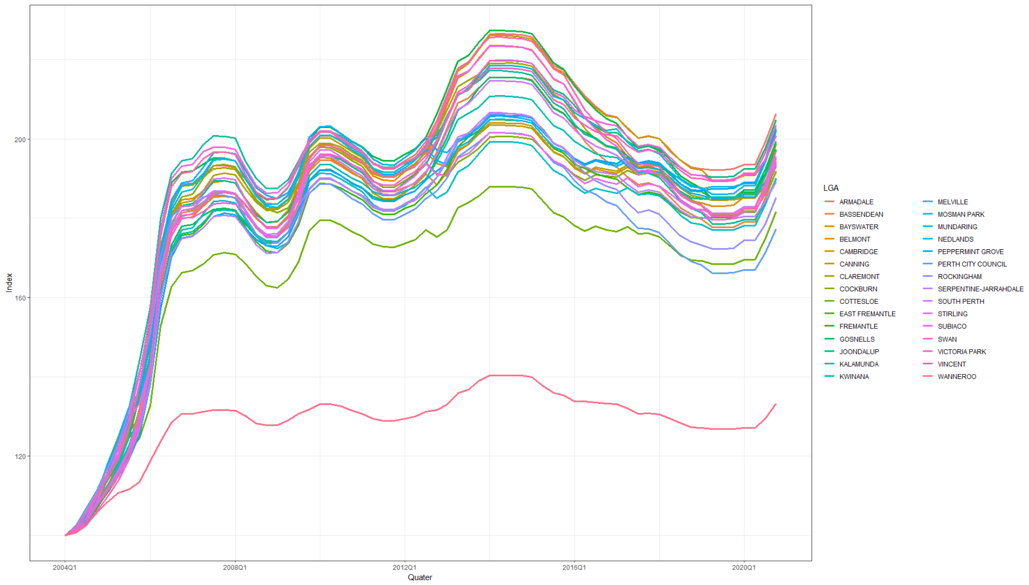

Figure C.1: The sequence of *RPPI* for each *LGA*, Perth.

ALE improves this by explaining the interactions between the temporal variable and the other variables. For instance, Figure C.1 and Figure C.2 depict the $RPPI$s for all LGAs and all property classes respectively. In these two figures, the price dynamics in each LGA or property class follow the whole market trend roughly. The implementation process is much more simple and convenient than compiling submarket indices using the classic method. This could benefit the practical application of machine learning techniques for residential real estate analysis.

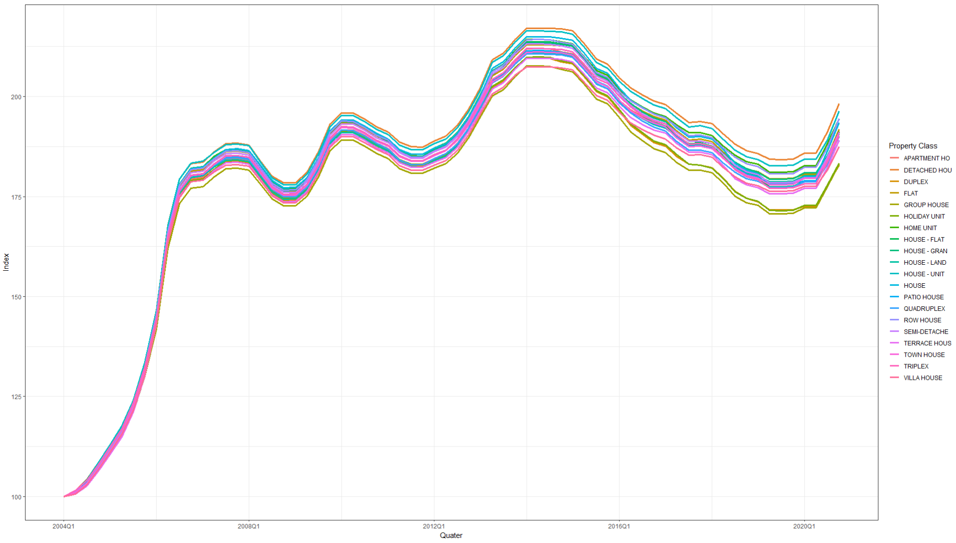

Figure C.2: The sequence of *RPPI* for all available property classes in Perth.

Appendix D

For Chapter 7: Real Option Theory

D.1 The deterministic process

To define the problems mathematically, suppose initially that P_t and C_t follow a deterministic process without any uncertainties, given constant growth rates, g_p and g_c, respectively.

$$\begin{aligned} \frac{dP_t}{P_t} &= g_p\, dt, \\ \frac{dC_t}{C_t} &= g_c\, dt, \end{aligned} \tag{D.1}$$

Eq. 7.5 suggests that the optimization in the deterministic process is

$$\begin{aligned} L_{t_0} &= \max_t (V_t e^{-r_f(t-t_0)}) \\ &= \max_T (V_{T+t_0} e^{-r_f(T)}), \\ L_{t_0}(T^*) &= (P^* - C^*) e^{-r_f T^*} \\ &= (R^* - 1) C^* e^{-r_f T^*}, \end{aligned} \tag{D.2}$$

where r_f is the risk-free return rate as the future is certain, T denotes the duration from t_0 to future time t, and T^* is the optimal duration. Let R^* represent the ratio of the optimal price (P^*) and the optimal construction cost (C^*). At any time, if the ratio of the P_t and C_t reaches R^* ($R_t = P_t/C_t \geq R^*$), the development is undertaken, and deferred otherwise.

Then, let L_{t_0} is a function of P_{t_0} and C_{t_0},

$$L_{t_0} = f(P_{t_0}, C_{t_0}), \tag{D.3}$$

due to that, the information is only available at time t_0. Thus, the total derivative equation of L_{t_0} is:

$$\begin{aligned} dL_{t_0} &= \frac{\partial L_{t_0}}{\partial P_{t_0}} dP_{t_0} + \frac{\partial L_{t_0}}{\partial C_{t_0}} dC_{t_0}, \\ &= \frac{\partial L_{t_0}}{\partial P_{t_0}} g_p P_{t_0}\, dt + \frac{\partial L_{t_0}}{\partial C_{t_0}} g_c C_{t_0}\, dt, \\ &= \left(\frac{\partial L_{t_0}}{\partial P_{t_0}} g_p P_{t_0} + \frac{\partial L_{t_0}}{\partial C_{t_0}} g_c C_{t_0} \right) dt. \end{aligned} \tag{D.4}$$

Then, the rate of return on this development project must be equal to the rate of return on any other riskless projects; otherwise, there would be arbitrage opportunities. Such that,

$$r_f L_{t_0} = \frac{\partial L_{t_0}}{\partial P_{t_0}} g_p P_{t_0} + \frac{\partial L_{t_0}}{\partial C_{t_0}} g_c C_{t_0}. \tag{D.5}$$

Eq. D.5 explains that the return the vacant land generates in each period should be equal to the growth for the payoff of the development project in each period when the land value is equal to the maximized present value of the development project.

According to McDonald and Siegel (1986), the form of L_{t_0} could be

$$L_{t_0} = k(R^* - 1) C_{t_0}^{\mu} R_{t_0}^{\alpha}, \tag{D.6}$$

where k is a constant and R_{t_0} is the ratio of P_{t_0} and C_{t_0}. Then, the solution of Eq. D.5 must satisfy certain boundary conditions: (i) $L_{t_0} = (R_{t_0} - 1)C_{t_0} = (R^* - 1)C^*$ when $R_{t_0} = P_{t_0}/C_{t_0} = P^*/C^* = R^*$ at time t_0; (ii) $L_{t_0} \to 0$ when $P_{t_0}/C_{t_0} \to 0$. Based on condition (i),

$$L_{t_0} = k(R^* - 1) C^{*\mu} R^{*\alpha} = (R^* - 1)C^*. \tag{D.7}$$

Such that, $\mu = 1$ and $k = 1/R^{*\alpha}$. Substituting μ, k and Eq. D.6 to Eq. D.5,

$$\begin{aligned} r_f C_{t_0} R_{t_0}^{\alpha} &= \alpha C_{t_0} R_{t_0}^{\alpha-1} C_{t_0}^{-1} g_p P_{t_0} + R_{t_0}^{\alpha} g_c C_{t_0} - \alpha C_{t_0} R_{t_0}^{\alpha-1} \frac{P_{t_0}}{C_{t_0}^2} g_c C_{t_0} \\ r_f &= \alpha g_p + (1-\alpha) g_c \\ \alpha &= \frac{r_f - g_c}{g_p - g_c}. \end{aligned} \tag{D.8}$$

Thus, the value of vacant land could be

$$\begin{aligned} L_{t_0} &= \frac{R^* - 1}{R^{*\alpha}} C_{t_0} R_{t_0}^{\alpha} \\ &= (R^* - 1) C_{t_0} \left(\frac{R_{t_0}}{R^*} \right)^{\alpha}, \\ R_{t_0} &= \frac{P_{t_0}}{C_{t_0}}, \\ \alpha &= \frac{r_f - g_c}{g_p - g_c}. \end{aligned} \tag{D.9}$$

Depending on the condition (ii),

$$\alpha = \frac{r_f - g_c}{g_p - g_c} > 0. \tag{D.10}$$

To derive the optimal timing of development, the first order condition with respect to T^* is applied on Eq. D.2,

$$P_{t_0} e^{(g_p - r_f)T^*}(g_p - r_f) - C_{t_0} e^{(g_c - r_f)T^*}(g_c - r_f) = 0. \tag{D.11}$$

By solving the previous equation, the optimal timing of development (T^*) is

$$T^* = ln\left(\frac{C_{t_0}}{P_{t_0}} \times \frac{r - g_c}{r - g_p} \right) / (g_p - g_c). \tag{D.12}$$

In addition, the second-order condition is needed to ensure that the extreme value is the maximum. The second order derivative needs to be less than 0 when $T = T^*$, then,

$$\begin{aligned} P_{t_0}e^{(g_p-r_f)T^*}(g_p-r_f)^2 - C_{t_0}e^{(g_c-r_f)T^*}(g_c-r_f)^2 &< 0 \\ \frac{P_{t_0}}{C_{t_0}}e^{(g_p-g_c)T^*}\frac{(g_p-r_f)^2}{(g_c-r_f)^2} &< 1 \\ \frac{g_p-r_f}{g_c-r_f} &< 1. \end{aligned} \tag{D.13}$$

Eq. D.13 gives some restrictions on g_p and g_c. When growth rates, g_p and g_c, are larger than the real risk-free discount rate, r_f, then, the growth rate of the price should be less than the growth rate of construction cost, $g_p < g_c$; Otherwise, $g_p > g_c$. After T^* is calculated, R^*, then, is

$$\begin{aligned} R^* &= \frac{P^*}{C^*} \\ &= \frac{P_{t_0}}{C_{t_0}}e^{(g_p-g_c)T^*} \\ &= \frac{P_{t_0}}{C_{t_0}}e^{(g_p-g_c)ln(\frac{C_{t_0}}{P_{t_0}}\times\frac{r_f-g_c}{r_f-g_p})/(g_p-g_c)} \\ &= \frac{r_f-g_c}{r_f-g_p} \\ R^* &> 1. \end{aligned} \tag{D.14}$$

Given the restrictions,

$$g_c > g_p > r_f \ OR \ r_f > g_p > g_c, \tag{D.15}$$

the solution of the deterministic process at time t_0 is summarized as the following:

- When $P_{t_0}/C_{t_0} \leq R^*$,

$$\begin{aligned} L_{t_0} &= (R^* - 1)C_{t_0}\left[\frac{P_{t_0}/C_{t_0}}{R^*}\right]^{\frac{r_f - g_c}{g_p - g_c}}, \\ T^* &= ln\left(\frac{C_{t_0}}{P_{t_0}}R^*\right)/(g_p - g_c), \\ R^* &= \frac{r_f - g_c}{r_f - g_p} > 1. \end{aligned} \tag{D.16}$$

- Otherwise, when $P_{t_0}/C_{t_0} > R^*$,

$$\begin{aligned} L_{t_0} &= P_{t_0} - C_{t_0} \\ T^* &= 0, \\ R^* &= \frac{r_f - g_c}{r_f - g_p} > 1. \end{aligned} \tag{D.17}$$

The analogy to the financial option, the development project under the deterministic process is analogous to a European call option on a common stock. It gives the landowners the right to spend the construction cost (the exercise price of the option, C^*) and receive a property (a share of stock) which is worth P^* at the expiration time $t_0 + T^*$. The value of the option (the land value) is derived above. The solution to this problem under the deterministic process is a special case of the problem solved below.

D.2 The stochastic process

Now consider the same problem, except that the future price and construction cost are stochastic. Thus, P_t and C_t include risk premium, because of uncertainties, and they are assumed following the geometric Brownian motion of the form,

$$\begin{aligned}\frac{dP_t}{P_t} &= g_p\,dt + \sigma_p\,dz_p,\\ \frac{dC_t}{C_t} &= g_c\,dt + \sigma_c\,dz_c.\end{aligned} \tag{D.18}$$

g_p and g_c are the growth rates, σ_p and σ_c are the volatility or standard deviations, all of which are time-invariant. Additionally, dz_p and dz_c are the increments of a standard Wiener process. The problem is formulated as a first passage problem. The same argument establishes, that development should occur when the ratio, P_t/C_t, reaches the critical barrier, R^*. Then, Eq. 7.5 could be rewritten, and the expected land value (or the expected present value of the payoff) is

$$\begin{aligned}L_{t_0} &= V_{t_0}\\ &= \max_t E_{t_0}[(P_t - C_t)e^{-r'(t-t_0)}]\\ &= E_{t_0}[(R^* - 1)C^* e^{-r'T^*}]\end{aligned} \tag{D.19}$$

E_{t_0} is the conditional expectation on the available information at time t_0, r' is the return rate with a risk premium to discount the future values[1]. T^* is the duration from t_0 to the first-passage time ($t^* \geq t_0$) when P_{t^*}/C_{t^*} first reaches the critical ratio, R^*, and $V_{t^*}e^{-r'(t^*-t_0)}$ is the maximum.

Thus, at time t_0, the land value could be expressed as a function of P_{t_0} and C_{t_0}.

$$L_{t_0} = L(P_{t_0}, C_{t_0}) = E_{t_0}[(R^* - 1)C^* e^{-r'T^*}] \tag{D.20}$$

[1] Because the future is uncertain. Thus, the discount rate should have a risk premium. The Capital Asset Pricing Model (*CAPM*) can be used to determine the risk-adjusted discount rate, $r' = r_f + \beta[E(r_m) - r_f]$, where $E(r_m)$ is the expected discount rate of the market.

It is noted that L_{t_0} must satisfy the following equation[2]:

$$\begin{aligned}
dL_{t_0} &= \frac{\partial L_{t_0}}{\partial P_{t_0}} dP_{t_0} + \frac{\partial L_{t_0}}{\partial C_{t_0}} dC_{t_0} + \frac{1}{2}\left(\frac{\partial^2 L_{t_0}}{\partial P_{t_0}^2} dP_{t_0}^2 + 2\frac{\partial L_{t_0}}{\partial P_{t_0}}\frac{\partial L_{t_0}}{\partial C_{t_0}} dP_{t_0} dC_{t_0} \right. \\
&\quad \left. + \frac{\partial^2 L_{t_0}}{\partial C_{t_0}^2} dC_{t_0}^2 \right) \\
dL_{t_0} &= \left[\frac{\partial L_{t_0}}{\partial P_{t_0}} g_p P_{t_0} + \frac{\partial L_{t_0}}{\partial C_{t_0}} g_c C_{t_0} + \frac{1}{2}\left(\frac{\partial^2 L_{t_0}}{\partial P_{t_0}^2} \sigma_p^2 P_{t_0}^2 + 2\frac{\partial L_{t_0}}{\partial P_{t_0}}\frac{\partial L_{t_0}}{\partial C_{t_0}} \rho_{pc}\sigma_p\sigma_c P_{t_0} C_{t_0} \right.\right. \\
&\quad \left.\left. + \frac{\partial^2 L_{t_0}}{\partial C_{t_0}^2} \sigma_c^2 C_{t_0}^2 \right)\right] dt + \frac{\partial L_{t_0}}{\partial P_{t_0}} \sigma_p P_{t_0} dz_p + \frac{\partial L_{t_0}}{\partial C_{t_0}} \sigma_c C_{t_0} dz_c.
\end{aligned} \tag{D.21}$$

Then, the rate of return on this development project must be equal to the rate of return on any other risky projects; otherwise, there would be arbitrage opportunities. Using Malliaris and Brock (1982, Theorem 7.5), it is arrived at the partial differential equation PDE:

$$\begin{aligned}
r' L_{t_0} &= \frac{\partial L_{t_0}}{\partial P_{t_0}} g_p P_{t_0} + \frac{\partial L_{t_0}}{\partial C_{t_0}} g_c C_{t_0} + \frac{1}{2}\frac{\partial^2 L_{t_0}}{\partial P_{t_0}^2} \sigma_p^2 P_{t_0}^2 + \frac{\partial L_{t_0}}{\partial P_{t_0}}\frac{\partial L_{t_0}}{\partial C_{t_0}} \rho_{pc}\sigma_p\sigma_c P_{t_0} C_{t_0} \\
&\quad + \frac{1}{2}\frac{\partial^2 L_{t_0}}{\partial C_{t_0}^2} \sigma_c^2 C_{t_0}^2.
\end{aligned} \tag{D.22}$$

As stated before, r' is a risk-adjusted return rate. The solution of Eq. D.22 must satisfy certain boundary conditions as shown in the deterministic process: (i) $L_{t_0} = (R_{t_0} - 1)C_{t_0} = (R^* - 1)C^*$ when $R_{t_0} = P_{t_0}/C_{t_0} = P^*/C^* = R^*$ at time t_0; (ii) $L_{t_0} \to 0$ when $P_{t_0}/C_{t_0} \to 0$.

According to McDonald and Siegel (1986), the form of L_{t_0} could be

$$L_{t_0} = k(R^* - 1)C_{t_0}^{\mu} R_{t_0}^{\alpha} \tag{D.23}$$

[2] This equation follows Itô's lemma and the conditions of the Wiener process.

with k is a constant and R_{t_0} is the ratio of P_{t_0} and C_{t_0}. This guess satisfies Eq. D.22. The boundary condition (i) requires that $k = \frac{1}{(R^*)^{-\alpha}}$ and that $\mu = 1$. With these conditions, Eq. D.22 can be written as

$$0 = \frac{1}{2}\alpha(\alpha-1)[\sigma_p^2 + \sigma_c^2 - 2\rho_{pc}\sigma_p\sigma_c] + \alpha g_p + (1-\alpha)g_c - r'. \tag{D.24}$$

Define $\sigma^2 = \sigma_p^2 + \sigma_c^2 - 2\rho_{pc}\sigma_p\sigma_c$ and ρ_{pc} is the instantaneous correlation between the changes of the property price and construction cost, and simplify the equation,

$$0 = \frac{1}{2}\sigma^2\alpha^2 + (g_p - g_c - \frac{1}{2}\sigma^2)\alpha + g_c - r'. \tag{D.25}$$

As long as $r' - g_c > 0$, Eq. D.25 will have both a positive and a negative results. Boundary condition (ii) requires that $\alpha > 0$, so the positive solution is the correct one. α is equal to

$$\alpha = \frac{1}{\sigma^2}\left[-\left(g_p - g_c - \frac{1}{2}\sigma^2\right) + \sqrt{(g_p - g_c - \frac{1}{2}\sigma^2)^2 + 2\sigma^2(r' - g_c)}\right]. \tag{D.26}$$

Then, the value of vacant land, L, is

$$\begin{aligned} L_{t_0} &= (R^* - 1)C_{t_0}\left(\frac{P_{t_0}/C_{t_0}}{R^*}\right)^{\alpha}, \\ \alpha &= \frac{1}{\sigma^2}\left[-\left(g_p - g_c - \frac{1}{2}\sigma^2\right) + \sqrt{(g_p - g_c - \frac{1}{2}\sigma^2)^2 + 2\sigma^2(r' - g_c)}\right]. \end{aligned} \tag{D.27}$$

This solution works for arbitrary R^*. Choosing the optimal R^* amounts to imposing an additional boundary condition, variously known as "high contact" or "smooth pasting". Optimal R^* is calculated using the first-order condition on L_{t_0}.

$$\begin{aligned} \frac{dL_{t_0}}{dR^*} = C_{t_0}\left(\frac{P_{t_0}}{C_{t_0}}\right)^{\alpha} R^{*-\alpha} + (R^* - 1)C_{t_0}\left(\frac{P_{t_0}}{C_{t_0}}\right)^{\alpha}(-\alpha)R^{*-(\alpha+1)} = 0 \\ 0 = R^{*-\alpha} + (R^* - 1)(-\alpha)R^{*-(\alpha+1)} \\ 0 = R^{*-\alpha} - \alpha R^{*-\alpha} + \alpha R^{*-(\alpha+1)} \\ \alpha - 1 = \alpha R^{*-1} \\ R^* = \frac{\alpha}{\alpha - 1} \end{aligned} \tag{D.28}$$

After deriving the value of the option, the timing of development in the stochastic process is also determined by the available information at time t_0. As explained previously, P_t and C_t are assumed following geometric Brownian motion, such that, their analytic solutions under Itô's interpretation could be expressed as

$$\begin{aligned} P_{t^*} &= P_{t_0+T^*} = P_{t_0} e^{(g_p - \frac{1}{2}\sigma_p^2)T^* + \sigma_p z_{p,T^*}}, \\ C_{t^*} &= C_{t_0+T^*} = C_{t_0} e^{(g_c - \frac{1}{2}\sigma_c^2)T^* + \sigma_c z_{c,T^*}}. \end{aligned} \tag{D.29}$$

Because t^* is the optimal time to develop when P_{t^*}/C_{t^*} reaches the critical ratio (R^*) at the first time, then, the optimal duration is T^*. Such that,

$$\begin{aligned} \frac{P_{t^*}}{C_{t^*}} &= R^* \\ &= \frac{P_{t_0}}{C_{t_0}} e^{[g_p - \frac{1}{2}\sigma_p^2 - (g_c - \frac{1}{2}\sigma_c^2)]T^* + \sigma z_{pc,T^*}}, \end{aligned} \tag{D.30}$$

where z_{pc,T^*} is a Wiener process[3]. As the Wiener process is a martingale, it has a property given the optional stopping theorem, $E(z_{pc,T^*}) = E(z_{pc,t_0}) = 0$. Then,

$$0 = E_{t_0}\left\{ ln\left[\frac{P_{t^*}}{C_{t^*}} \times \frac{C_{t_0}}{P_{t_0}}\right] - \left[g_p - \frac{1}{2}\sigma_p^2 - (g_c - \frac{1}{2}\sigma_c^2)\right]T^* \right\} \tag{D.31}$$

Because $P_{t^*}/C_{t^*} = R^*$, the above equation could be rewritten as

$$0 = E_t\left\{ ln\left[R^* \times \frac{C_{t_0}}{P_{t_0}}\right] - \left[g_p - \frac{1}{2}\sigma_p^2 - g_c + \frac{1}{2}\sigma_c^2\right]T^* \right\}. \tag{D.32}$$

[3] Because,P_{t^*}/P_{t_0} and C_{t^*}/C_{t_0} follow log-normal distribution. Thus, $log(P_{t^*}/P_{t_0})$ and $log(C_{t^*}/C_{t_0})$ follow normal distribution with mean $(g_p - \frac{1}{2}\sigma_p^2)t^*$, variance $\sigma_p^2 T^*$ and mean $(g_c - \frac{1}{2}\sigma_c^2)T^*$, variance $\sigma_c^2 T^*$ separately. $log(P_{t^*}/C_{t^*}C_{t_0}/P_{t_0}) = log(P_{t^*}/P_{t_0}) - log(C_{t^*}/C_{t_0})$, then, it should follow the convolution distribution of $N[(g_p - \frac{1}{2}\sigma_p^2)T^*,\ \sigma_p^2 T^*]$ and $N[(g_c - \frac{1}{2}\sigma_c^2)T^*,\ \sigma_c^2 T^*]$, which is $N\{[g_p - \frac{1}{2}\sigma_p^2 - (g_c - \frac{1}{2}\sigma_c^2)]T^*,\ \sigma^2 T^*\}$, where $\sigma^2 = \sigma_p^2 + \sigma_c^2 - 2\rho_{pc}\sigma_p\sigma_c$. The derivation of Eq. D.30 is from the definition of log-normal distribution, $log(P_{t^*}/C_{t^*} \times C_{t_0}/P_{t_0}) \sim N\{[g_p - \frac{1}{2}\sigma_p^2 - (g_c - \frac{1}{2}\sigma_c^2)]T^*,\ \sigma^2 T^*\}$.

Rearrange the equation, the expected value of optimal time (t^*) could be derived,

$$E_{t_0}(T^*) = ln\left[R^* \times \frac{C_{t_0}}{P_{t_0}}\right] / \left[g_p - g_c - (\frac{1}{2}\sigma_p^2 - \frac{1}{2}\sigma_c^2)\right]. \tag{D.33}$$

To summarize, the solution of the problem at time t_0 for the stochastic process is

- When $P_{t_0}/C_{t_0} \leq R^*$,

$$\begin{aligned} L_{t_0} &= (R^* - 1)C_{t_0}\left(\frac{P_{t_0}/C_{t_0}}{R^*}\right)^{\alpha}, \\ E_{t_0}(T^*) &= ln\left[\frac{C_{t_0}}{P_{t_0}}R^*\right] / \left[g_p - g_c - (\frac{1}{2}\sigma_p^2 - \frac{1}{2}\sigma_c^2)\right], \\ R^* &= \frac{\alpha}{\alpha - 1}, \\ \alpha &= \frac{1}{\sigma^2}\left[-\left(g_p - g_c - \frac{1}{2}\sigma^2\right) + \sqrt{(g_p - g_c - \frac{1}{2}\sigma^2)^2 + 2\sigma^2(r' - g_c)}\right]. \end{aligned} \tag{D.34}$$

- Otherwise, when $P_{t_0}/C_{t_0} > R^*$,

$$\begin{aligned} L_{t_0} &= P_{t_0} - C_{t_0} \\ T^* &= 0, \\ R^* &= \frac{\alpha}{\alpha - 1}, \\ \alpha &= \frac{1}{\sigma^2}\left[-\left(g_p - g_c - \frac{1}{2}\sigma^2\right) + \sqrt{(g_p - g_c - \frac{1}{2}\sigma^2)^2 + 2\sigma^2(r' - g_c)}\right]. \end{aligned} \tag{D.35}$$

The analogy to the financial option, the development project under stochastic process is analogous to a perpetual American call option on a common stock.

It gives the landowners the right to spend the construction cost (the exercise price of the option, C^*) and receive a property (a share of stock) that is worth P^* at the optimal time t^*. The option will not expire, as the optimal time t^* is a random variable. The value of the option (the land value) is derived above.

Compared with the solution for the deterministic process, the formulas of L_{t_0} and $E_{t_0}(T^*)$ have a similar form. In the stochastic process, however, the uncertainties play important roles. Risk premiums are taken into account when the option is evaluated and the timing of development is decided. The development project under these two processes could be represented by two different types of call options. In the deterministic process, the expired time of the option is fixed, and it is the time to exercise. On the contrary, there is no expired time of option in the stochastic process, the option could be exercised at any time if it is optimal. In addition, there is one more restriction in the stochastic process, $r' - g_c > 0$, which is not required in the deterministic process. The common prerequisite is r_f $(or\ r') > g_p > g_c$.

D.3 Stationarity of time series

The following plots show the time series of price, construction cost, and their first lag difference in twenty LGAs.

In this case, before applying these time series, their stationarities have to be checked via ADF (Augmented Dickey-Fuller) test. The time series of price and construction cost are not stationary, but their first differences are. The

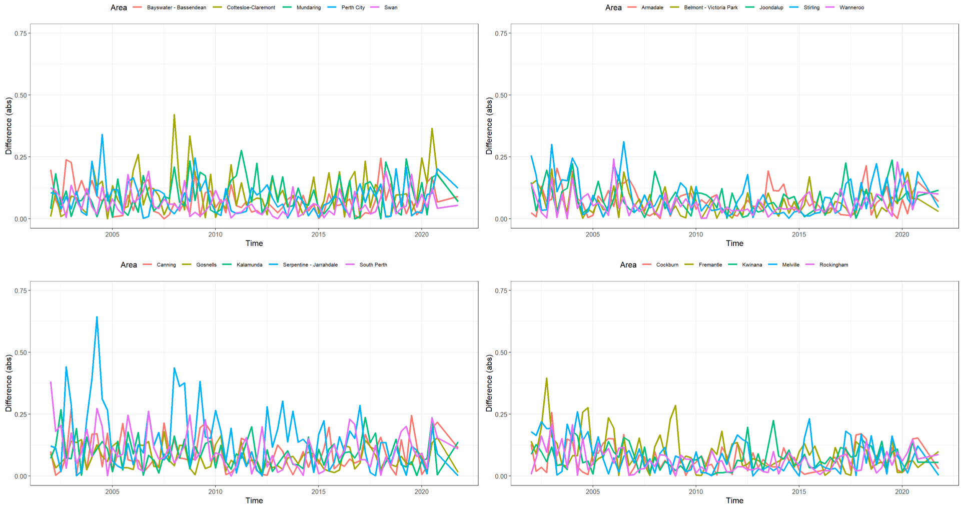

Figure D.1: The time series plot of the first difference of price in twenty *LGA*s. The x-axis is the time from 2003. The y-axis shows the first differences in the logarithm of prices.

ADF test has three specifications, shown in the following.

- None:

$$\Delta y_t = \alpha_1 y_{t-1} + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + \epsilon_t$$

$$H_0: \ \alpha_1 = 0 \ (\tau 1)$$

- Drift:

$$\Delta y_t = \alpha_1 y_{t-1} + \alpha_2 + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + \epsilon_t$$

$$H_0: \ \alpha_1 = 0 \ \& \ \alpha_2 = 0 \ (\phi 1)$$

$$H_0: \ \alpha_1 = 0 \ (\tau 2)$$

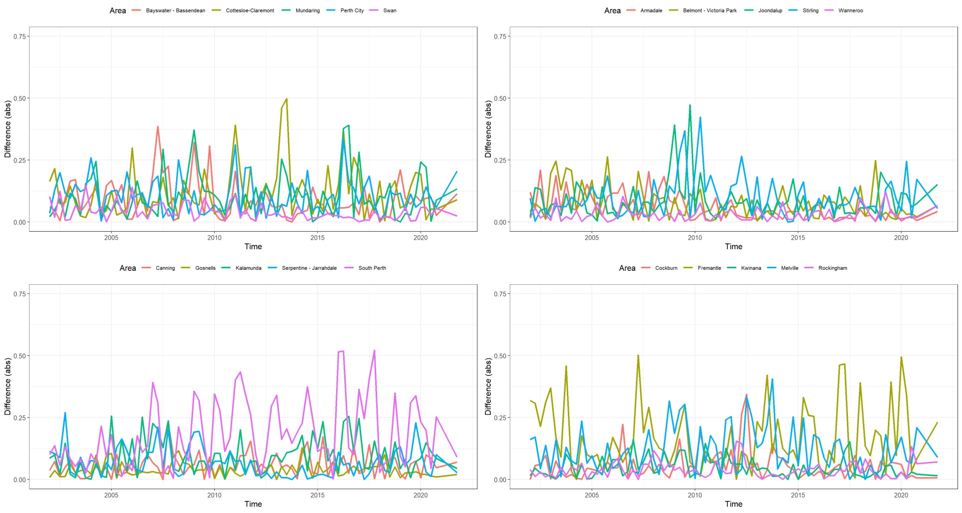

Figure D.2: The time series plot of the first difference of construction cost in twenty *LGA*s. The x-axis is the time from 2003. The y-axis shows the first differences for the logarithm of averaged construction costs in one *LGA*.

- Trend:

$$\Delta y_t = \alpha_1 y_{t-1} + \alpha_2 + \alpha_3 t + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + \epsilon_t$$

$$H_0: \ \alpha_1 = 0 \ \& \ \alpha_2 = 0 \ \& \ \alpha_3 = 0 \ (\phi 2)$$

$$H_0: \ \alpha_1 = 0 \ \& \ \alpha_2 = 0 \ (\phi 3)$$

$$H_0: \ \alpha_1 = 0 \ (\tau 3)$$

The test outputs are in the table below.

Table D.1: The outputs of *ADF* test using the time series of price and construction cost in one *LGA*. shows the ADF test statistics and the critical value when the significant level is 0.01, 0.05 and 0.1.

Test type		Test statistics	Critical values		
			$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Price, None	$(\tau 1)$	1.2081	-2.6	-1.95	-1.61
Price, Drift	$(\tau 2)$	-2.2066	-3.51	-2.89	-2.58
	$(\phi 1)$	3.2634	6.70	4.71	3.86
Price, Trend	$(\tau 3)$	-2.7492	-4.04	-3.45	-3.15
	$(\phi 2)$	3.1998	6.50	4.88	4.16
	$(\phi 3)$	3.9492	8.73	6.49	5.47
Construction Cost, None	$(\tau 1)$	1.2091	-2.6	-1.95	-1.61
Construction Cost, Drift	$(\tau 2)$	-2.1614	-3.51	-2.89	-2.58
	$(\phi 1)$	3.1682	6.70	4.71	3.86
Construction Cost, Trend	$(\tau 3)$	-2.2712	-4.04	-3.45	-3.15
	$(\phi 2)$	2.5546	6.50	4.88	4.16
	$(\phi 3)$	2.9962	8.73	6.49	5.47
Price Diff, None	$(\tau 1)$	-8.2758	-2.6	-1.95	-1.61
Price Diff, Drift	$(\tau 2)$	-8.4899	-3.51	-2.89	-2.58
	$(\phi 1)$	36.0684	6.70	4.71	3.86
Price Diff, Trend	$(\tau 3)$	-8.5337	-4.04	-3.45	-3.15
	$(\phi 2)$	24.3369	6.50	4.88	4.16
	$(\phi 3)$	36.4764	8.73	6.49	5.47
Construction Cost Diff, None	$(\tau 1)$	-8.5395	-2.6	-1.95	-1.61
Construction Cost Diff, Drift	$(\tau 2)$	-8.7876	-3.51	-2.89	-2.58
	$(\phi 1)$	38.6109	6.70	4.71	3.86
Construction Cost Diff, Trend	$(\tau 3)$	-8.9069	-4.04	-3.45	-3.15
	$(\phi 2)$	26.5028	6.50	4.88	4.16
	$(\phi 3)$	39.7541	8.73	6.49	5.47

One statement could be summarized from the above table clearly, that the time series of price and construction cost are not stationary, but their first differences are. Meanwhile, in addition, *HEGY* test (Hylleberg, Engle, Granger, and Yoo test) is applied for the seasonal unit root. When the significant level is set to 0.05, the all-time series for the price, construction cost, and growth rates don't present any seasonal unit roots. Also, for applying vector autoregression (*VAR*), the *ACF* (auto-correlation function) is checked in the first differences of prices and construction costs. The graphs are depicted below.

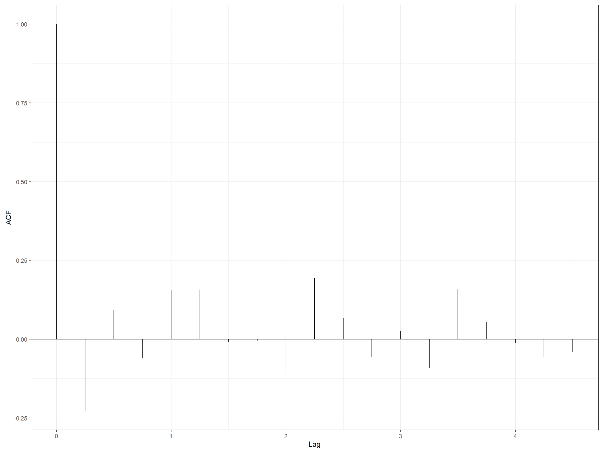

Figure D.3: The *ACF* plot for the first difference of prices. The graph shows the first differences of price in one *LGA* as an example.

D.4 The expectations and variances

The forecast for the growth rates of price and construction cost could be calculated. The expectation and the variance are calculated as the estimators for growth rates and volatility (uncertainties). For explaining the calculation process, the $AR(1)$ models for the first difference of price and construction cost are used as an example.

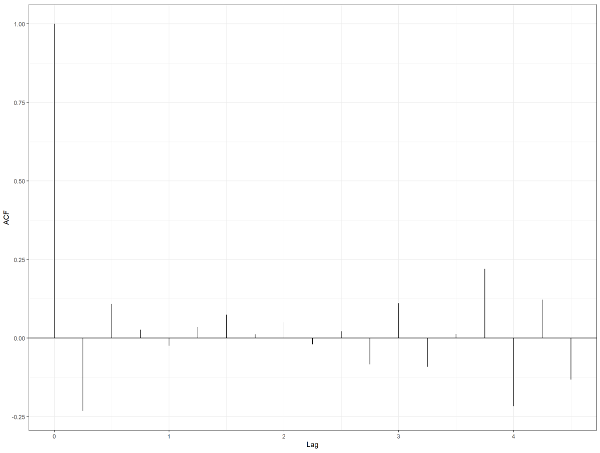

Figure D.4: The *ACF* plot for the first difference of construction cost. The graph shows the first differences in construction cost in one LGA as an example.

The two time-series data of the first differences for price and construction follow $AR(1)$ models:

$$\begin{aligned} \Delta p_t &= \alpha_{p,0} + \alpha_{p,1}\Delta p_{t-1} + \epsilon_{p,t} \\ \Delta c_t &= \alpha_{c,0} + \alpha_{c,1}\Delta c_{t-1} + \epsilon_{c,t} \end{aligned} \tag{D.36}$$

with $|\alpha_{p,1}| < 1,\ |\alpha_{c,1}| < 1$ and $\epsilon_{p,t} \sim (0, \sigma_p^2),\ \epsilon_{c,t} \sim (0, \sigma_c^2)$. At time t, all the information is available before t. When steps ahead forecasts are needed with horizon $h \geq 1$, a set of equations could be consecutively substituted for obtaining the predictions.

$$\begin{aligned} \Delta p_{t+1} &= \alpha_{p,0} + \alpha_{p,1}\Delta p_t + \epsilon_{p,t+1} \\ \Delta p_{t+2} &= \alpha_{p,0} + \alpha_{p,1}(\alpha_{p,0} + \alpha_{p,1}\Delta p_t + \epsilon_{p,t+1}) + \epsilon_{p,t+2} \\ \Delta p_{t+3} &= \alpha_{p,0} + \alpha_{p,1}[\alpha_{p,0} + \alpha_{p,1}(\alpha_{p,0} + \alpha_{p,1}\Delta p_t + \epsilon_{p,t+1}) + \epsilon_{p,t+2}] + \epsilon_{p,t+3} \\ \Delta p_{t+4} &= \alpha_{p,0} + \alpha_{p,1}\{\alpha_{p,0} + \alpha_{p,1}[\alpha_{p,0} + \alpha_{p,1}(\alpha_{p,0} + \alpha_{p,1}\Delta p_t + \epsilon_{p,t+1}) + \epsilon_{p,t+2}] \\ &\qquad + \epsilon_{p,t+3}\} + \epsilon_{p,t+4} \end{aligned} \tag{D.37}$$

Similarly, for the first difference of construction cost, we have

$$\begin{aligned}\Delta c_{t+1} &= \alpha_{c,0} + \alpha_{c,1}\Delta c_t + \epsilon_{c,t+1} \\ \Delta c_{t+2} &= \alpha_{c,0} + \alpha_{c,1}(\alpha_{c,0} + \alpha_{c,1}\Delta c_t + \epsilon_{c,t+1}) + \epsilon_{c,t+2} \\ \Delta c_{t+3} &= \alpha_{c,0} + \alpha_{c,1}[\alpha_{c,0} + \alpha_{c,1}(\alpha_{c,0} + \alpha_{c,1}\Delta c_t + \epsilon_{c,t+1}) + \epsilon_{c,t+2}] + \epsilon_{c,t+3} \\ \Delta c_{t+4} &= \alpha_{c,0} + \alpha_{c,1}\{\alpha_{c,0} + \alpha_{c,1}[\alpha_{c,0} + \alpha_{c,1}(\alpha_{c,0} + \alpha_{c,1}\Delta c_t + \epsilon_{c,t+1}) + \epsilon_{c,t+2}] \\ &\quad + \epsilon_{c,t+3}\} + \epsilon_{c,t+4}\end{aligned} \tag{D.38}$$

In the long term, the calculation process is

$$\begin{aligned}\Delta p_{t+h} &= \alpha_{p,0}\sum_{i=0}^{h-1}\alpha_{p,1}^i + \alpha_{p,1}^h\Delta p_t + \sum_{i=0}^{h-1}\alpha_{p,1}^i\epsilon_{t+h-i}, \\ \Delta c_{t+h} &= \alpha_{c,0}\sum_{i=0}^{h-1}\alpha_{c,1}^i + \alpha_{c,1}^h\Delta c_t + \sum_{i=0}^{h-1}\alpha_{c,1}^i\epsilon_{t+h-i}.\end{aligned} \tag{D.39}$$

Such that the conditional expectations and variances are

$$\begin{aligned}E(\Delta p_{t+h}|t) &= \alpha_{p,0}\sum_{i=0}^{h-1}\alpha_{p,1}^i + \alpha_{p,1}^h\Delta p_t \\ Var(\Delta p_{t+h}|t) &= \sum_{i=0}^{h-1}\alpha_{p,1}^{2i}\sigma_p^2\end{aligned} \tag{D.40}$$

and

$$\begin{aligned}E(\Delta c_{t+h}|t) &= \alpha_{c,0}\sum_{i=0}^{h-1}\alpha_{c,1}^i + \alpha_{c,1}^h\Delta c_t \\ Var(\Delta c_{t+h}|t) &= \sum_{i=0}^{h-1}\alpha_{c,1}^{2i}\sigma_c^2\end{aligned} \tag{D.41}$$

For the unconditional expectations and variances, we obtain for the limits

$$\begin{aligned}\lim_{h\to\infty} E(\Delta p_{t+h}) &= \frac{\alpha_{p,0}}{1-\alpha_{p,1}} \\ \lim_{h\to\infty} Var(\Delta p_{t+h}) &= \frac{\sigma_p^2}{1-\alpha_{p,1}^2}\end{aligned} \tag{D.42}$$

and

$$\begin{aligned}\lim_{h\to\infty} E(\Delta c_{t+h}) &= \frac{\alpha_{c,0}}{1-\alpha_{c,1}}\\ \lim_{h\to\infty} Var(\Delta c_{t+h}) &= \frac{\sigma_c^2}{1-\alpha_{c,1}^2}\end{aligned} \tag{D.43}$$

Similarly, in *VAR*, the unconditional expectations and variances could be expressed, *VAR*(1) is shown as an example. The time series of the first differences' vector for price and construction cost that follows

$$\begin{bmatrix}\Delta p_t\\ \Delta c_t\end{bmatrix} = \begin{bmatrix}\alpha_0\\ \beta_0\end{bmatrix} + \begin{bmatrix}\alpha_{p,1} & \alpha_{c,1}\\ \beta_{p,1} & \beta_{c,1}\end{bmatrix}\begin{bmatrix}\Delta p_{t-1}\\ \Delta c_{t-1}\end{bmatrix} + \begin{bmatrix}\epsilon_{p,t}\\ \epsilon_{c,t}\end{bmatrix}. \tag{D.44}$$

The vector of unconditional expectations could be

$$\begin{bmatrix}E(\Delta p|t)\\ E(\Delta c|t)\end{bmatrix} = \left(\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} - \begin{bmatrix}\alpha_{p,1} & \alpha_{c,1}\\ \beta_{p,1} & \beta_{c,1}\end{bmatrix}\right)^{-1}\begin{bmatrix}\alpha_0\\ \beta_0\end{bmatrix}. \tag{D.45}$$

Meanwhile, the vector of unconditional variances is

$$\begin{aligned} &vec\left(\begin{bmatrix}Var(\Delta p|t) & Cov(\Delta p,\Delta c|t)\\ Cov(\Delta p,\Delta c|t) & Var(\Delta c|t)\end{bmatrix}\right)\\ &= \left(\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix} - \begin{bmatrix}\alpha_{p,1} & \alpha_{c,1}\\ \beta_{p,1} & \beta_{c,1}\end{bmatrix}\otimes\begin{bmatrix}\alpha_{p,1} & \alpha_{c,1}\\ \beta_{p,1} & \beta_{c,1}\end{bmatrix}\right)^{-1}\\ &\quad vec\left(\begin{bmatrix}Var(\epsilon_{p,t}) & Cov(\epsilon_{p,t},\epsilon_{c,t})\\ Cov(\epsilon_{p,t},\epsilon_{c,t}) & Var(\epsilon_{c,t})\end{bmatrix}\right)\end{aligned} \tag{D.46}$$

vec() is to verctorize the matrix and ⊗ is the Kronecker product of matrices. This shows the conditional variance in the long term for *VAR*, and the eigenvalue of the parameter matrix should be less than one.

Bibliography

Agarwal, S., Fan, Y., McMillen, D. P. and Sing, T. F.: 2021, Tracking the pulse of a city — 3d real estate price heat maps, *Journal of Regional Science* **61**(3), 543–569. (Cited on page 27.)

Anglin, P. M. and Gencay, R.: 1996, Semiparametric estimation of a hedonic price function, *Journal of Applied Econometrics* **11**, 633–648. (Cited on page 20.)

Apley, D. W. and Zhu, J.: 2020, Visualizing the effects of predictor variables in black box supervised learning models, *Journal of the Royal Statistical Society* **82**(4), 1059–1086. Series B, Statistical methodology. (Cited on page 78.)

Arnott, R. and Lewis, F.: 1979, The transition of land to urban use, *Journal of political economy* **87**(11), 161–169. (Cited on page 30.)

Australian Bureau of Statistics: 2020a, Building a new home: Construction cost changes.
URL: *https://www.abs.gov.au/articles/building-new-home-construction-cost-changes* (Cited on pages 37 and 145.)

Australian Bureau of Statistics: 2020b, Residential construction and the finance process. (Cited on pages 7 and 155.)

Bailey, M. J., Muth, R. F. and Nourse, H. O.: 1963, A regression method for real estate price index construction, *Journal of the American Statistical Association* **58**(304), 933–942. (Cited on page 24.)

Bao, H. X. H. and Wan, A. T. K.: 2004, On the use of spline smoothing in estimating hedonic housing price models: Empirical evidence using Hong Kong data, *Real Estate Economics* **32**(3), 487–507. (Cited on page 120.)

Bertsimas, D., Delarue, A. and Pauphilet, J.: 2021, Prediction with missing data. (Cited on page 85.)

Bourassa, S. C., Cantoni, E. and Hoesli, M.: 2010, Predicting house prices with spatial dependence: A comparison of alternative methods, *The Journal of Real Estate Research* **32**(2), 139–160. (Cited on page 20.)

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A.: 1984, *Classification and Regression Trees*, Belmont, Calif : Wadsworth International Group. (Cited on pages 68, 69, and 84.)

Bulan, L., Mayer, C. and Somerville, T.: 2009, Irreversible investment, realoptions, and competition: Evidence from real estate development, *Journalof Urban Economics* **65**, 237–251. (Cited on pages 34, 35, 194, and 202.)

Capozza, D. R. and Helsley, R. W.: 1989, The fundamentals of land prices and urban growth, *Journal of Urban Economics* **26**(3), 295–306. (Cited on pages 30 and 31.)

Capozza, D. R. and Li, Y.: 1994, The intensity and timing of investment: The case of land, *The American Economic Review* **84**(4), 889–904. (Cited on pages 32, 155, and 181.)

Capozza, D. R. and Li, Y.: 2001, Residential investment and interest rates:an empirical test of land development as a real option, *Real Estate Economics* **29**, 503–519. (Cited on page 34.)

Capozza, D. R. and Li, Y.: 2002, Optimal land development decisions, *Journal of Urban Economics,* **51**(1), 123–142. (Cited on pages 32 and 34.)

Case, B. and Quigley, J. M.: 1991, The dynamics of real estate prices, *Review of Economics and Statistics* **22**, 50–58. (Cited on page 26.)

Case, K. E. and Shiller, R. J.: 1987, Prices of single family homes since 1970: new indexes for four cities, *New England Economic Review* pp. 45–56. (Cited on page 25.)

Case, K. E. and Shiller, R. J.: 1989, The efficiency of the market for single-family homes, *The American Economic Review* **79**(1), 125–137. (Cited on page 25.)

Chau, K. W., Wong, S. K. and Yiu, C. Y.: 2005, Adjusting for non-linear age effects in the repeat sales index, *The journal of real estate finance and economics* **31**(2), 137–153. (Cited on page 120.)

Chu, Y. and Sing, T. F.: 2021, Intensity and timing options in real estate developments, *International Real Estate Review* **24**(1), 1–17. (Cited on page 33.)

Clapham, E., Englund, P., Quigley, J. M. and Redfearn, C. L.: 2006, Revisiting the past and settling the score: Index revision for house price derivatives, *Real Estate Economics* **34**(2), 275–302. (Cited on pages 25 and 136.)

Clapp, J. M. and Giaccotto, C.: 1992, Estimating price trends for residential property: A comparison of repeat sales and assessed value methods, *Journal of Real Estate Finance and Economics* **5**, 357–374. (Cited on page 117.)

Clapp, J. M. and Giaccotto, C.: 1998, Price indices based on the hedonic repeat sale method: application to the housing market, *Journal of Real Estate Finance and Economics* **16**(1), 5–26. (Cited on page 26.)

Clapp, J. M. and Giaccotto, C.: 1999, Revisions in repeat-sales price indexes: here today, gone tomorrow?, *Real Estate Economics* **27**(1), 79–104. (Cited on page 25.)

Crone, T. M. and Voith, R. P.: 1992, Estimating house price appreciation: A comparison of methods, *Journal of Housing Economics* **2**(4), 324–338. (Cited on pages 27 and 28.)

Cunningham, C. R.: 2006, House price uncertainty, timing of development, and vacant land prices: Evidence for real options in Seattle, *Journal of Urban Economics* **59**(1), 1–31. (Cited on pages 7, 10, 34, 156, 160, 162, 164, and 194.)

de Haan, J.: 2004, Direct and indirect time dummy approaches to hedonic price measurement, *Journal of Economic and Social Measurement* **29**(4), 427–443. (Cited on page 27.)

Dempster, A. P., Laird, N. M. and Rubin, D. B.: 1977, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1), 1–38. (Cited on page 80.)

Deng, Y., McMillen, D. P. and Sing, T. F.: 2012, Private residential price indices in singapore: A matching approach, *Regional Science and Urban Economics* **42**(3), 485–494. (Cited on page 25.)

Deng, Y. and Quigley, J. M.: 2008, Index revision, house price risk, and the market for house price derivatives, *The Journal of Real Estate Finance and Economics* **37**(3), 191–209. (Cited on page 135.)

Diewert, E. and Shimizu, C.: 2015, Residential property price indices for Tokyo, *Macroeconomic dynamics* **19**(8), 1659–1714. (Cited on pages 119 and 120.)

Diewert, W. E., Silver, M. and Heravi, S.: 2007, Hedonic imputation versus time dummy hedonic indexes, *IMF working paper* **7**(234), 1. (Cited on page 27.)

Dixit, A. K. and Pindyck, R. S.: 1994, *Investment under uncertainty*, Princeton University Press, Princeton, New Jersey. (Cited on page 31.)

Donald, S.: 1970, The optimal timing of urban land development., *Papers of the Regional Science Association* **25**, 35–44. (Cited on page 30.)

Duan, N.: 1983, Smearing estimate: A nonparametric retransformation method, *Journal of the American Statistical Association* **78**(383), 605–610. (Cited on page 79.)

Eurostat: 2013, *Handbook on Residential Propertty Prices Indices (RPPIs)*, European Commission, Eurostat, OECD and World Bank. (Cited on pages 6, 9, 22, and 119.)

Fan, G. Z., Pu, M., Sing, T. F. and Zhang, X.: 2022, Risk aversion and urban land development options, *Real Estate Economics* **50**(3), 767–788. (Cited on page 35.)

Friedman, J. H.: 2001, Greedy function approximation: a gradient boosting machine, *The Annals of Statistics* **29**(5), 1189–1232. (Cited on pages 69, 77, 89, and 104.)

Friedman, J. H.: 2002, Stochastic gradient boosting, *Computational Statistics & Data Analysis* **38**(4), 367–378. (Cited on pages 69, 70, and 218.)

Frisch, R.: 1981, From utopian theory to practical applications: The case of econometrics, *The American Economic Review* **71**(6), 1–16. Nobel Lectures and 1981 Survey of Members. (Cited on page 71.)

Gatzlaff, D. H. and Ling, D. C.: 1994, Measuring changes in local house prices: An empirical investigation of alternative methodologies, *Journal of Urban Economics* **35**, 221–244. (Cited on pages 27 and 28.)

Goetzmann, W. N. and Spiegel, M.: 1997, A spatial model of housing returns and neighborhood substitutability, *Journal of Real Estate Finance and Economics* **14**(1–2), 11–31. (Cited on page 26.)

Goh, M., Costello, G. and Schwann, G.: 2012, Accuracy and robustness of house price index methods, *Housing Studies* **27**, 643–666. (Cited on page 48.)

Greenwell, B., Boehmke, B., Cunningham, J. and Developers, G.: 2020, *gbm: Generalized Boosted Regression Models*. R package version 2.1.8.
URL: *https://CRAN.R-project.org/package=gbm* (Cited on page 218.)

Hastie, T., Tibshirani, R. and Friedman., J.: 2009, *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*, second edition edn, New York, NY : Springer New York. (Cited on pages 68, 84, and 95.)

Haupt, H., Schnurbus, J. and Tschernig, R.: 2010, On nonparametric estimation of a hedonic price function, *Journal of Applied Econometrics* **25**, 894–901. (Cited on page 20.)

Hill, R. C., Knight, J. R. and Sirmans, C. F.: 1997, Estimating capital asset price indexes, *Review of Economics and Statistics* **79**(2), 226–233. (Cited on page 26.)

Hill, R. J.: 2004, Superlative index numbers: not all of them are super, *Journal of Econometrics* **130**(1), 25–43. (Cited on page 27.)

Hill, R. J.: 2013a, Hedonic price indexes for residential housing: A survey, eevaluation and taxonomy, *Journal of Economic Surveys* **27**, 879–914. (Cited on page 119.)

Hill, R. J.: 2013b, Hedonic price indexes for residential housing: A survey, evaluation and taxonomy, *Journal of economic surveys* **27**(5), 879–914. (Cited on page 28.)

Hill, R. J. and Melser, D.: 2008, Hedonic imputation and the price index problem: An application to housing, *Economic Inquiry* **46**, 593–609. (Cited on pages 28 and 116.)

Hill, R. J. and Scholz, M.: 2018, Can geospatial data improve house price indexes? A hedonic imputation approach with splines, *Review of Income and Wealth* **64**, 737–756. (Cited on pages 5, 11, 55, 113, 120, 141, 200, 205, and 215.)

Hinrichs, N., Kolbe, J. and Werwatz, A.: 2021, Using shrinkage for data-driven automated valuation model specification – a case study from Berlin, *Journal of Property Research* **38**, 130–153. (Cited on page 90.)

Holland, A. S., Ott, S. H. and Riddiough, T. J.: 2000, The role of uncertainty in investment: An examination of competing investmentmodels using commercial real estate data, *Real Estate Economics* **28**, 33–64. (Cited on page 33.)

Huber, P. J.: 1964, Robust estimation of a location parameter, *The Annals of Mathematical Statistics* **35**(1), 73–101. (Cited on page 75.)

Hyndman, R. J. and Athanasopoulos, G.: 2018, *Forecasting: principles and practice*, second edn, OTexts: Melbourne, Australia.
URL: *https://otexts.com/fpp2/* (Cited on pages 72 and 74.)

James, G., Witten, D., Hastie, T. and Tibshirani, R.: 2017a, *An Introduction to Statistical Learning: With Applications in R*, New York, NY: Springer New York. (Cited on pages 67 and 69.)

James, G., Witten, D., Hastie, T. and Tibshirani, R.: 2017b, *An Introduction to Statistical Learning. With Applications in R*, Springer, New York NY. (Cited on page 70.)

Jamshidian, M. and Bentler, P. M.: 1999, ML estimation of mean and covariance structures with missing data ssing complete data routines, *Journal of Educational and Behavioral Statistics* **24**(1), 21–41. (Cited on page 85.)

Jansen, S., de Vries, P., Coolen, H., Lamain, C. and Boelhouwer, P.: 2008, Developing a house price index for the Netherlands, *Journal of Real Estate and Finance and Economics* **37**(2), 163–186. (Cited on page 25.)

Josse, J., Prost, N., Scornet, E. and Varoquaux, G.: 2020, On the consistency of supervised learning with missing values. (Cited on page 82.)

Kagie, M. and Van Wezel, M.: 2007, Hedonic price models and indices based on boosting applied to the Dutch housing market, *Intelligent Systems in Accounting, Finance and Management* **15**(3/4), 85–106. (Cited on pages 20, 55, 90, and 91.)

Knight, J. R., Dombrow, J. and Sirmans, C. F.: 1995, A varying parameters approach to constructing house price indexes, *Real Estate Economics* **23**(2), 187–205. (Cited on page 28.)

Knight, J. R., Sirmans, C. F., Gelfand, A. E. and Ghosh, S. K.: 1998, Analyzing real estate data problems using the gibbs sampler, *Real Estate Economics* **26**(3), 469–492. (Cited on page 90.)

Kok, N., Koponen, E.-L. and Martínez-Barbosa, C.: 2017, Big data in real estate? From manual appraisal to automated valuation, *Journal of Portfolio Management* **43**(6), 202–211. (Cited on pages 20, 21, and 89.)

Kolbe, J., Schulz, R., Wersing, M. and Werwatz, A.: 2021, Real estate listings and their usefulness for hedonic regressions, *Empirical Economics* **61**, 3239–3269. (Cited on page 90.)

Krause, A. and Lipscomb, C. A.: 2016, The data preparation process in real estate: Guidance and review, *Journal of Real Estate Practice and Education* **19**, 15–42. (Cited on pages 36, 55, 61, and 62.)

Leishman, C., Costello, G., Rowley, S. and Watkins, C.: 2013, The predictive performance of multilevel models of housing sub-markets: A comparative analysis, *Urban Studies* **50**, 1201–1220. (Cited on pages 48 and 96.)

Leishman, C. and Watkins, C.: 2002, Estimating local repeat sales house price indices for British cities, *Journal of Property Investment & Finance* **20**(1), 36–58. (Cited on page 25.)

Liao, W. and Wang, X.: 2012, Hedonic house prices and spatial quantile regression, *Journal of Housing Economics* **21**(1), 16–27. (Cited on page 20.)

Little, R. J. A. and Rubin, D. B.: 2002, *Statistical Analysis with Missing Data*, Wiley Series in Probability and Statistics, second edn, John Wiley & Sons, Hoboken NJ. (Cited on page 90.)

Liu, X.: 2012, Spatial and temporal dependence in house price prediction, *The Journal of Real Estate Finance and Economics* **47**(2), 341–369. (Cited on page 20.)

Lumley, T.: 2020, *leaps: Regression Subset Selection*. R package version 3.1.
URL: *https://CRAN.R-project.org/package=leaps* (Cited on page 218.)

Malliaris, A. G. and Brock, W. A.: 1982, *Stochastic methods in economics and finance*, Amsterdam: North-Holland Publishing Company. (Cited on page 231.)

Mayer, M.: 2021, *missRanger: Fast Imputation of Missing Values*. R package version 2.1.3.
URL: *https://CRAN.R-project.org/package=missRanger* (Cited on page 218.)

Mayer, M., Bourassa, S. C., Hoesli, M. and Scognamiglio, D.: 2019, Estimation and updating methods for hedonic valuation, *Journal of European Real Estate Research* **12**, 134–150. (Cited on pages 6, 20, 21, 36, 89, 91, 100, and 215.)

McDonald, R. and Siegel, D.: 1986, The value of waiting to invest, *The Quarterly Journal of Economics* **101**(4), 707–727. (Cited on pages 31, 32, 144, 155, 226, and 231.)

Morvan, M. L., Josse, J., Scornet, E. and Varoquaux, G.: 2021, What's a good imputation to predict with missing values? (Cited on pages 82 and 90.)

Nagaraja, C. H., Brown, L. D. and Wachter, S.: 2014, Repeat sales house price index methodology, *Journal of Real Estate Literature* **22**(1), 23–46. (Cited on page 118.)

Pakes, A.: 2003, A reconsideration of hedonic price indexes with an application to PCs, *The American Economic Review* **93**(5), 1578–1596. (Cited on page 117.)

Parmeter, C. F., Henderson, D. J. and Kumbhakar, S. C.: 2007, Nonparametric estimation of a hedonic price function, *Journal of Applied Econometrics* **22**, 695–699. (Cited on page 20.)

Pindyck, R. S.: 1991, Irreversibility, uncertainty, and investment, *Journal of Economic Literature* **29**, 1110–1148. (Cited on page 31.)

Prasad, N. L. and Richards, A.: 2008, Improving median housing price indexes through stratification, *Journal of Real EstateResearch* **30**(1), 45–71. (Cited on page 23.)

Quigg, L.: 1993, Empirical testing of real option-pricing models, *The Journal of Finance* **48**(2), 621–640. (Cited on pages 7, 32, 33, 37, and 202.)

Quigley, J. M.: 1995, A simple hybrid model for estimating real estate price indexes, *Journal of Housing Economics* **4**(1), 1–12. (Cited on page 26.)

R Core Team: 2021, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/* (Cited on page 95.)

Ribeiro, M. T., Singh, S. and Guestrin, C.: 2016, Model-agnostic interpretability of machine learning, *ICML Workshop on Human Interpretability in Machine Learning* . (Cited on page 76.)

RICS: 2017, The future of valuations, *Insight paper*, Royal Institution of Chartered Surveyors, London. (Cited on page 22.)

RICS: 2021, Automated valuation models, *Insight paper*, Royal Institution of Chartered Surveyors, London. (Cited on pages 19 and 89.)

Rosen, S.: 1974, Hedonic prices and implicit markets: product differentiation in pure competition, *Journal of Political Economy* **82**, 34–55. (Cited on pages 19 and 26.)

Rubin, D. B.: 1976, Inference and missing data, *Biometrika* **63**, 581–592. (Cited on page 80.)

Rubin, D. B.: 1978, Multiple imputations in sample surveys — A phenomenological bayesian approach to nonresponse, *Survey Research Methods*, American Statistical Association, pp. 20–34. (Cited on page 82.)

Rubin, D. B.: 1987, *Multiple imputation for nonresponse in surveys*, New York: Wiley. (Cited on pages 82 and 90.)

Schulz, R. and Wersing, M.: 2021, Automated valuation services: A case study for Aberdeen in Scotland, *Journal of Property Research* **38**(2), 154–172. (Cited on pages 6, 20, 21, 36, 79, 89, 94, 100, and 215.)

Schulz, R., Wersing, M. and Werwatz, A.: 2014, Automated valuation modelling: A specification exercise, *Journal of Property Research* **31**, 131–153. (Cited on pages 79 and 90.)

Shiller, R. J.: 1993, Measuring asset values for cash settlement in derivative markets: hedonic repeated measures indicesand perpetual futures, *Journal of Finance* **48**(3), 911–931. (Cited on page 26.)

Shimizu, C., Nishimura, K. G. and Watanabe, T.: 2010, Housing prices in Tokyo: A comparison of hedonic and repeat sales measures, *Journal of*

Economics and Statistics **230**(6), 792–813. (Cited on pages 27, 133, 142, and 201.)

Shimizu, C., Takatsuji, H., Ono, H. and Nishimura, K. G.: 2010, Structural and temporal changes in the housing market and hedonic housing price indices: A case of the previously owned condominium market in the Tokyo metropolitan area, *International Journal of Housing Markets and Analysis* **3**(4), 351–368. (Cited on page 120.)

Silver, M. and Heravi, S.: 2007, The difference between hedonic imputation indexes and time dummy hedonic indexes, *Journal of Business & Economic Statistics* **25**(2), 239–246. (Cited on page 119.)

Sing, T. F., Yang, J. J. and Yu, S. M.: 2022, Boosted tree ensembles for artificial intelligence based automated valuation models, *The Journal of Real Estate Finance and Economics* **65**(4), 649–674. (Cited on page 21.)

Somerville, C. T.: 2001, Permits, starts, and completions: Structural relationships versus real options, *Real Estate Economics* **29**, 161–190. (Cited on pages 33 and 34.)

Stekhoven, D. J. and Bühlmann, P.: 2012, Missforest—non-parametric missing value imputation for mixed-type data, *Bioinformatics* **28**(1), 112–118. (Cited on page 90.)

Steurer, M., Hill, R. J. and Pfeifer, N.: 2021, Metrics for evaluating the performance of machine learning based automated valuation models, *Journal of Property Research* **38**, 99–129. (Cited on pages 5, 55, 79, 86, and 90.)

Svetunkov, I.: 2021, *greybox: Toolbox for Model Building and Forecasting*. R package version 1.0.1.

URL: *https://CRAN.R-project.org/package=greybox* (Cited on page 218.)

Tang, F. and Ishwaran, H.: 2017, Random forest missing data algorithms, *Statistical Analysis and Data Mining* **10**(6), 363–377. (Cited on pages 85, 90, and 96.)

Titman, S.: 1985, Urban land prices under uncertainty, *The American Economic Review* **75**(3), 505–514. (Cited on pages 30 and 31.)

Triplett, J.: 2006, *Handbook on Hedonic indexes and Quality Adjustments in Price Indexes: Special Application to Information Technology Products*, Paris: OECD Publishing. (Cited on page 28.)

Wang, Y., Tang, W. and Jia, S.: 2016, Uncertainty, competition and timing of land development: theory and empirical evidence from Hang Zhou, China, *Journal of Real Estate Finance and Economics* **53**(2), 218–245. (Cited on pages 10, 35, and 37.)

Wilks, S. S.: 1932, Moments and distributions of estimates of population parameters from fragmentary samples, *The Annals of Mathematical Statistics* **3**(3), 163–195. (Cited on page 85.)

Williams, J. T.: 1991, Real estate development as an option, *Journal of RealEstate Finance and Economics* **4**, 191–208. (Cited on pages 30, 31, and 33.)

Wood, R.: 2005, A comparison of UK residential house price indices, *Real Estate Indicators and Financial Stability, BIS Papers No 21, Bank for International Settlements, Washington DC: The International Monetary Fund.* pp. 212–227. (Cited on page 23.)

Zhang, L., Shi, D., Chang, X. and Wen, H.: 2021, Timing decisions of housing sales and development based on real option theory, *International Journal of Strategic Property Management* **25**(2), 90–101. (Cited on pages 35 and 37.)

Zhu, J. and Apley, D. W.: 2018, *Accumulated local effects (ALE) and package ALEPlot*.
URL: *https://cran.r-project.org/web/packages/ALEPlot/vignettes* (Cited on page 77.)