ANALYSING HOUSING PRICE IN AUSTRALIA WITH DATA SCIENCE METHODS

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

Institute for Sustainable Industries and Liveable Cities

Victoria University

by

Jiaying Kou

March 2022

ABSTRACT

ANALYSING HOUSING PRICE IN AUSTRALIA WITH DATA SCIENCE METHODS

Jiaying Kou, Ph.D.

Victoria University 2022

Housing market price prediction is a major and important challenge in economics. Since the 2008 global financial crisis, researchers, economists, and politicians around the world have increasingly drawn attention to the need of better understanding housing market behaviour, since the failure to predict housing market crisis ahead of time had led to catastrophic global damage. On the other hand, around the same time, we have seen the revolution of information technology and artificial intelligence in the last two decades. The advent of powerful cloud and high performance computing systems, big data, and advanced machine learning algorithms have demonstrated new applications and advantages in cutting-edge research and technology areas such as pattern recognition, bioinformatics, natural language processing, and product recommendation systems.

Can we make the leap of improving our understanding of housing market behaviour by leveraging these recent advances in artificial intelligence and newly available big data? This is the main theme of the thesis. There is strong motivation to explore the application of data science methods, including new large datasets and advanced machine learning algorithm, to accelerate our understanding of housing market problems for the benefit of the common good.

In order to understand housing market behaviour, we divide the problem into two major steps: first, to improve understanding of housing appraisal (at micro

level), which is to predict housing price at the point level given a fixed timeframe; second, to improve understanding of the trend prediction (at macro level), which is to predict the housing price trend for a specific place during a time interval.

For these two major steps, we improve upon traditional economic modelling by:

- Adding new, non-traditional variables/features to our models, such as location-based Point of Interests, regional economic clusters, qualitative index, searching index, and newspaper articles
- Applying machine learning algorithms for data analysis, such as non-linear algorithms, K-Nearest-Neighbour, Support Vector Machine, Gradient Boost, and sentiment analysis

Specifically, in Chapter 3, we focus on the development of Location-Based Social Network (LBSN) for our micro-level housing appraisal modelling. A good location goes beyond the direct benefits from its neighbourhood. By leveraging housing data, neighbourhood data, regional economic cluster data and demographic data, we build a housing appraisal model, named HNED. Unlike most previous statistical and machine learning based housing appraisal research, which limit their investigations to neighbourhoods within 1km radius of the house, we expand the investigation beyond the local neighbourhood and to the whole metropolitan area, by introducing the connection to significant influential economic nodes, which we term *Regional Economic Clusters*. Specifically, we introduce regional economic clusters within the metropolitan range into the housing appraisal model, such as the connection to CBD, workplace, or the convenience and quality of big shopping malls and university clusters. When used with the gradient boosting algorithm XGBoost to perform housing price appraisal, HNED reached 0.88 in R^2. In addition, we found that the feature vector from Regional Economic Clusters alone

reached 0.63 in R^2, significantly higher than all traditional features. Chapter 3 focuses on the exploration and validation of HNED modelling.

In Chapter 4 and Chapter 5, we focus on macro-level housing price trend prediction. We fill the gap between the traditional macro-level housing market modelling and new developments of the concept of irrationality in microeconomic theories, by collecting and analysing economic behavioural data, such as real estate opinions in local newspaper articles, and people's web searching behaviour as captured by Google Trend Index. In Chapter 4, we discuss the usage of micro-level behavioural data for understanding macro-level housing market behaviour. We use sentiment analysis to examine local newspaper articles discussing real estate at a suburb level in inner-west Sydney, Australia. We then calculate the media sentiment index by using two different methods, and compare them with each other and the housing price index. The use of media sentiment index can serve as a finer-grained guiding tool to facilitate decision-making for home buyers, investors, researchers and policy makers. In Chapter 5, we discuss how new developments of behavioural economic theory indicate that the information from decision-making at the micro-level will bring a new solution to the age-old problem of economic forecasting. It provides the theoretical link between irrationality and big data methods. Specifically, Google Trend Index is included as a new variable in a time series auto-regression model to forecast housing market cycles.

To summarise the contributions of the thesis, we conclude that this is a successful early attempt to study housing price problems using data science methods, by leveraging newly available data sets and applying novel machine learning methods. Specifically, location-based social data improves the housing appraisal modelling. Human behaviour for housing market is analysed by introducing local newspaper articles and Google Trend Index into the modelling and analysis.

DOCTOR OF PHILOSOPHY DECLARATION

I, Jiaying Kou, declare that the PhD thesis entitled *Analysing Housing Price in Australia with Data Science Methods* is no more than 80,000 words in length including quotes and exclusive of tables, figures, appendices, bibliography, references and footnotes. This thesis contains no material that has been submitted previously, in whole or in part, for the award of any other academic degree or diploma. Except where otherwise indicated, this thesis is my own work.

I have conducted my research in alignment with the Australian Code for the Responsible Conduct of Research and Victoria University's Higher Degree by Research Policy and Procedures.

Signature  Date 11/03/2022

ACKNOWLEDGEMENTS

The journey of my Ph.D. is the journey of knowing myself. And like on many journeys, one must from time to time face the proverbial fork in the road. The poet Robert Frost once wrote these famous verses: "Two roads diverged in a wood, and I— I took the one less traveled by, And that has made all the difference."[1] In my Ph.D. journey, curiosity and instinct had led me to take the road "grassy and wanted wear",[2] just like in Frost's poem. I was not fully equipped with the required courage and ability at the beginning of the journey, but they were gained through the journey by enlightenment from myself and the unconditional trust, encouragement, guidance, support, and countless inspirational discussions with my supervisors, spouse, colleagues, friends, and family. My Ph.D. experience has been just like the African proverb, "It takes a village to raise a child".

If I try to describe the journey of knowing myself more precisely, it would be the journey of knowing my "growing" self. The process of enlightenment is the biggest reward of this journey. For example, to solve the problem of procrastination, I read lots of books related to psychology, brain function, jogging, and so on, and gained the habits of writing early in the morning, jogging, meditation, and keeping a diary. The rewards include reached 1,000 km cumulative jogging distance, writing more than 100 blog articles about self-growth, and becoming a healthier me. My Ph.D. journey is a great evidence of Immanuel Kant's definition of enlightenment. To me, the purpose of the Ph.D. journey is to become an independent researcher and thinker, who is able to find valuable research questions, and has the courage, means, and methods to answer these questions – *Sapere aude*![3] Luckily, I believe that I have achieved this purpose, to the extent that I trust my research journey

[1]Frost, Robert. "The road not taken." (1916): 1232-1233.

[2]Ibid.

[3]Kant, Immanuel. Answering the question: What is enlightenment?. Strelbytskyy Multimedia Publishing, 2019.

will continue beyond the completion of my Ph.D., and myself can continue to grow.

A seed cannot grow without the right timing, resources, and nurturing environment. I feel very lucky to have met so many great people who supported me to go through this journey. I cannot forget the day when I knocked on Prof. Yanchun Zhang's office door. Prof. Yanchun listened to my research plan, and kept supporting me since that day. A multi-disciplinary research project involving Data Science cannot go ahead without supervision from the Computer Science side. I have benefitted immensely from his wisdom, insights, and guidance regarding my general research direction. He also introduced me to my Principal Supervisor, Prof. Hua Wang, who is always calm, cheerful, and ready to provide great guidance, advice, and support. Prof. Hua taught me how to revise and refine the essence of my research questions, and drive the research project from there. He has also been a continued source of positive energy during my Ph.D. journey, even during my times of despair. He introduced me to the research lab's table tennis team, which helped me with work-life balance, which is essential for undertaking creative endeavours. I also would like to thank my Associate Supervisor, Prof. Xiaoming Fu, who brought me into the world of computational social science. Working with Prof. Xiaoming provided me with an excellent opportunity to look beyond where I am, and to challenge myself to the next level. Besides my supervisors, I would like to thank my spouse and collaborator, Geordie Zhang, who encouraged me to go on this journey, supported me by sharing his passion and understanding of research, his unconditional trust of me to complete this journey, and sustained real action of sharing the family housework workload. I feel very lucky to have had countless wonderful research discussions with him, to make our friendship extended to research collaboration. I am also indebted to Jiahua (Kevin) Du, who has been a great collaborator, and shared with me his experience on research and coding.

Many great memories of research discussions are always motivating me to continue exploring new research questions.

Besides my supervisors and collaborators, I would like to thank my friends from the research world who also provided countless support and encouragement: Yongqiang Li, Ye Wang, Yitong Li, Fan Liu, Jingyuan He, Hui Zheng, Siuly Siuly, Dinesh Pandey, Ravinder Singh, Sudha Subramani, Jiao Yin, Yongfeng Ge, Mingshan You, Kejing Du, Kun Wang, and many other fellow students and visiting fellows to our research lab at Victoria University.

I am also indebted to Victoria University, which provided me with the opportunity to work on a multidisciplinary research project, by being very supportive and flexible in organising supervision, and allowed me to transfer my Ph.D. from Business to Computer Science. Especially, I would like to thank Prof. Anne-Marie Hede and Prof. Ron Adams from the Office for Researcher Training, Quality and Integrity, and A/Prof. Randall Robinson from the Institute for Sustainable Industries and Liveable Cities, for providing valuable support, guidance and advice from a high level.

Besides the academic support, I am also lucky to have had support from the non-academic world. Reading books and articles by Ling Zhou[4], Rui Chen, Shanshan Zhang, Windy Liu[5], and personal discussions with them, regularly provided food for thought for different stages of the journey.

Finally, besides the fruitful inspiration from the great people I know, there are other inspirations, encouragement, and support from many great people whom I do not know personally. For example, reading Herbert Simon's works and autobiography[6] gave me a first hand account of how a scientist thinks and works through

[4]Zhou, Ling. "Renzhi Juexing [Cognitive Awakening]." Chinese, The People's Posts and Telecommunications Press (2020).

[5]Liu, Windy. "Xinzhi Tuwei [Mind Uplifting]." Chinese, Jiangxi People's Publishing House (2020)

[6]Simon, Herbert A. Models of my life. MIT press, 1996.

his life. Books by researchers such as Dan Ariely and Daniel Kahneman provided me with the opportunity to learn about how interesting thoughts become research outcomes. Haruki Murakami's book *Shokugyo Toshiteno Shosetsuka [Novelist as a profession]*[7] taught me how to become a writer.

Upon reflection, I realised that lots of the times when I fell down on this journey were because of a lack of courage. I would like to express my deepest gratitude to everyone in my life, who have been so generous and kind to give me positive affirmations, encouragement, and support.

Last but not least, I would like to express my sincere thanks to my parents and my daughter Audrey. Without their warm love and endless support, this journey would not have been possible.

I would like to finish this section with a quote from Rabindranath Tagore:

If you shed tears when you miss the sun, you also miss the stars.[8]

[7]Murakami, Haruki. "Shokugyo Toshiteno Shosetsuka [Novelist as a profession]." Japanese, Switch Publishing (2015).

[8]Tagore, Rabindranath. 1916. Stray Birds. New York: Macmillan.

PUBLICATIONS

1. **Kou, Jiaying**, Jiahua Du, Xiaoming Fu, Geordie Z. Zhang, Hua Wang, and Yanchun Zhang. "The Effect of Regional Economic Clusters on Housing Price." In Australasian Database Conference, Lecture Notes in Computer Science (LNCS), volume 12610, pp. 180-191. Springer, Cham, 2021.
2. **Kou, Jiaying**, and Yashar Gedik. "New Behavioural Big Data Methods for Predicting Housing Price." EAI Endorsed Transactions on Scalable Information Systems 6, no. 21 (2019).
3. **Kou, Jiaying**, Xiaoming Fu, Jiahua Du, Hua Wang, and Geordie Z. Zhang. "Understanding housing market behaviour from a microscopic perspective." In 2018 27th International Conference on Computer Communication and Networks (ICCCN), pp. 1-9. IEEE, 2018.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The application of current artificial intelligence technologies to research has transformed a wide range of scientific fields. Some recent breakthroughs include the use of AI generative adversarial networks (GANs) to explore the evolution of galaxies [122], and the AlphaFold algorithm fast-pacing the work of protein structure prediction for the human proteome [145]. Other research fields that have experience success with this approach include pattern recognition for medical image processing, natural language processing, human interactive design, and product recommendation systems.

Witnessing new breakthroughs and developments due to artificial intelligence in the physical and biomedical sciences, this naturally brings curiosity regarding the social sciences. A pertinent question arises from an empirical economist's perspective: will artificial intelligence make an impact on economic research, and how?

The short answer is yes. Recent developments in data science have brought new perspectives to economic problems – new approaches that involve new types of data, big data, and machine learning-based data analysis. These new developments provides a huge opportunity to discover new knowledge, to solve existing economic problems from a new angle, and even to solve new economic problems [105]. This also provides an exciting opportunity for fostering multi-disciplinary research questions, which can expand and benefit the understanding of economic problems. Since the joint research direction of economics and machine learning is relatively new, I will discuss in detail of the differences that machine learning can bring to econometrics in Section 1.2. This is especially important to provide a bird's-eye view to understand why such a multi-disciplinary approach is promising.

Given the view that introducing machine learning to tackle economic problems can generate fruitful outcomes, this thesis focuses on analysing housing price beyond the traditional economic modelling, by exploring new types of data, such as online searching index, text data from newspaper articles, location based data; and by applying machine learning algorithm, such as K-Nearest-Neighbour, Support Vector Machine, sentiment analysis, etc. The purpose is to explore and understand housing market related problems with machine learning tools.

1.1 Research Background and Motivation

Machine intelligence has widely used in many fields. In this section, the discussion is focused on what machine learning can bring to economic research, especially empirical economic study, why and how machine learning can generate a big impact on the economic research. The discussion is organised in four angles: the advantage of big data, new data, machine learning algorithms and the validation process.

There are lots of prolonged unsolved economic challenges. For example, how to reach relative accurate predictions for some economic trends, such as business cycles; how to understand economic growth; how to build a bridge between microeconomics and macroeconomics; how to understand the complex social and market behaviour. Economists tend to solve these problems by investigating top-down sophisticated mathematical models. However, as pointed out by Simon [128], who established the theory of bounded rationality, the painstaking microeconomic behaviour study is the solution for these decades-long economic challenges. Living in the digital age, it is natural to seek solutions of capturing micro-level economic behaviours through the methods of big data. The details are discussed in Section 1.3.

1.2 How Does Machine Learning Work With Economic Problems

This section focuses on the discussion of the impact of artificial intelligence on economic study. Econometrics, a field in economics study, is the science to understand economic problems by analysing real economic data with the support of statistic techniques and economic theories [132]. It has a long history to apply and to develop statistical modelling for economic problems. Linear Regression, non-linear regression, and time series regression have been widely used to understand economic variables since the 20th century. An interesting question here is: what are the breakthroughs when applying machine learning to econometrics? In the following section, the answer to this question is organised in four parts: big data, new data, algorithms, and validation methods.

1.2.1 Big Data

Large data sets have become possible and available due to the development of information technology in the last few decades. Cloud computing and storage systems are automatically collecting huge amount of data correctly and cheaply for online economic transactions, data generated through other sources such as during information searching, consuming, retrieval, generating, and processing behaviour. The amount of information and data generated in the modern era challenges the traditional methods of data processing and analysing. For example, YouTube channel's monthly users have reached two billion around the world. Within every minute, 500 hours of videos are uploaded (as of 2019)[1]. As the data continuously grows, information technology companies learned to cope with the increasing de-

[1] `https://blog.youtube/news-and-events/youtube-at-15-my-personal-journey`

mand of information storage and retrieval. In addition, other functions become essential, such as recommendation systems for electronic commerce, and real-time route optimisation systems for navigation software. New technologies tailored for big data processing have been developed because of the emerging applications of data science.

Economists who have the opportunity to face these dramatic changes first-hand, naturally came up with the idea to explore big data for economic studies [42, 148]. For example, they may encounter a spreadsheet with more than one million of rows. In [42], the author summarised the new challenges of big data. Firstly, modern data sets are available in real time, not as the traditional economic data sets which normally collected manually with high labour costs, long time consumption and delayed publication. New business and policy challenge is to process and respond to the real time data faster and better. For example, the current pandemic challenge needs prompt public health policy reaction based on the real time data collection on a daily basis. Secondly, data sets are larger than traditional data sets. Hundreds and thousands of rows are standard for economic empirical modelling. In comparison, one housing price modelling in this thesis used 161, 179 recorded sold properties in Melbourne in 2018. Thirdly, modern data sets are less structured and more complex. For example, data scraping technology can extract data from websites. This requires lots of data cleaning techniques to extract useful contents into the format ready for modelling.

To address these new data challenges, new tools are developed for manipulating and analysing big data. Varian in [148] discussed a few tools, such as MySQL, NoSQL, Hadoop file systems, Dremel, and BigQuery to open and analyse data sets that can't be loaded with normal Structured Query Language (SQL) system. The idea is to select sub samples with random sampling technique to do exploratory

data analysis. Due to the challenging demand, cloud computing systems gradually become the solution for data storage and computing. Large clusters of computers can be organised for computation to achieve low cost, high speed with reliability. In comparison, traditional economic modelling normally is performed using statistical software on a personal computer.

I believe in the future, big data analysis will become a common practice when economics merges with data science deeper with gradually available big data sets collected from commercial industries, government, IoT (The Internet of Things) and others.

1.2.2 New Data

The previous section focuses on the discussion of big data and especially the challenge of scale. This section focuses on the challenge and opportunity of novelty by introducing new types of data for economic modelling. New data includes the types of data are not commonly used in economics. The discussion focuses on *image data*, *text data*, *web related data*.

Image Data

In the recent decades, lots of exciting applications and developments of machine learning are primarily for pattern recognition. Real world applications include license plate recognition, face detection, medical diagnosis, automated driving, etc. Image pattern recognition used for economic study is relative new and in the exploratory stage. Most of the image applications are for economic value predictions. The motivation is that images haven't been used for economic value indexes. This can solve the problem of expensive, difficult and scarce data collection, because most of the socioeconomic data collection activities are organised by the

government periodically, which involve huge amount of costs and labour input. In addition, this creates dimensions beyond the traditional recognised variables, and increases the granularity of empirical data sets to discover patterns and behaviours not captured by the current data collection method. Leveraging existing image data, such as Google Street View, satellites, photos uploaded to the internet, can help data accessibility, real-time, scalable, low cost, high dimensional, high granular and high precision prediction.

The novelty of the typical methodology is to transform the pixel image into a classification problem, then to connect with the target of economic value prediction. For example, Gebru et al. [56] showed that by classifying cars in the Google Street View images with deep learning methods—specifically, Convolutional Neural Networks (CNN), socioeconomic attributes can be predicted. Fundamental economic values, such as income, education level, ethnic groups, election decision makings can be learned associated with types of cars on the street. Other studies such as [64, 71, 106] also showed similar prediction power for all kinds of important economic values. More detailed summary about satellite image research can be found in [39].

Text Data

Besides image data, text data is widely available online. The research about text analysis, called Natural Language Processing (NLP), has a long history. The recent development in machine learning and deep learning has achieved greatly towards the goal of machine "understanding" the language. One recent study discussed the development and future of deep learning for artificial intelligence in natural language processing [13]. There are a few natural language processing applications, such as speech recognition, machine translation, lexical semantics, automatic summarisation, natural language understanding, etc.

Most economic and financial natural language processing applications focus on the lexical semantics, at the level of word meaning analysis [57,112]. The data used for economic study can come from news, financial and economic reports, reviews and opinions on the websites, social media chats, etc. Such collected text is beyond traditional economic numerical data sets. People's opinions and perceptions about a social event or understanding of economic trends can be extracted from the text data. This is a new dimension of economic variables hard to be captured by the traditional methods, yet it is an important dimension to provide an opportunity to penetrate macro-level decision making behaviour of the population.

Most of the current applications of text data analysis are semantic and sentiment analysis. Researchers focus on the financial market prediction and have some positive results. Social media analysis related literature includes [5, 15, 58, 61, 123, 160]. These studies used social media data to generate trend prediction by analysing frequency of retweets, sentiment expressions, etc. Another direction is to analyse financial reports and other textual information using deep learning methods for stock price prediction, as described in [38, 49, 67, 154].

Web Related Data

Web related data includes all kinds of data collected from web resources using the technique as data scraping and crawling. The growing online resources have gradually become a rich source of data for economic studies. It has some unique characteristics and advantages compared to traditional economic variables. Firstly, similar to text data, it represents a collective opinion of people who use online resources. The most famous application is the study of Google trend index. People's searching popularity can be transformed into a search index and this has a near-term economic prediction ability [29]. Other popular predictions leverage reviews of all kinds of goods and services from websites. Secondly, similar to image data, it

is high dimensional, which contains information from other fields. It may contain geographic, demographic, business, transport, tourism information. Therefore, it needs multi-disciplinary inputs for joint study to reveal the new knowledge. Thirdly, comparing to the traditional economic data sets, web related data sets are usually unstructured and not cleaned. It needs a systematic data preprocessing stage before any statistical processing. This is a common practice and considered an important step in the data mining process.

1.2.3 Economic Modelling and Machine Learning Algorithms

The purpose of this section is to discuss the possible changes or integration of economic modeling with machine learning algorithms. The discussion follows the logic of showing the existing economic research problems and opportunities and how machine learning algorithms are relevant to these economic problems and opportunities. The detailed machine learning algorithms are discussed in the Literature Review Chapter, which is not the purpose of this section. Specifically, there are three important topics discussed in this section:

- First, the nature of socioeconomic problem is complexity, and how can machine learning algorithm help to understand the social and economic behaviour under the challenge of complexity?
- Second, why do I choose a real economic problem for this thesis with real data sets?
- Third, what is the role of economists in fostering machine learning in economic research?

Complexity

Physics Nobel Laureate Murray Gell-Mann once said, "Imagine how hard physics would be if electrons could think." (as reported by Page [110]) This described exactly the challenge of social science, including economics. If we could summarise this challenge in one word, it would be: complexity.

First, the complexity requires new data processing tools – specifically machine learning – to meet this challenge. The study of complexity can't be progressed without the leverage of new research data sets from either new types of data, such as text, images, videos, etc. that recorded through the new lenses of information technology, or data with new level of granularity. This provides the opportunity to study complexity at a new level. These types of data sets have been widely studied in the computer science field. Computer scientists have developed machine learning and deep learning algorithms to fulfill related objectives by utilising data sets. They developed tools and techniques for prediction, classification and clustering. These algorithms can help pattern recognition, recommendation system, web searching, data mining, understanding social network, etc.

Economic studies can apply these existing technology to study data sets beyond traditional types of economic data sets in order to progress in the understanding of economic behaviours. Athey [7] pointed out these tools can be effective as "intermediate step" in data-driven economic studies. For example, this thesis collected local newspaper articles in Sydney to create the intermediate step to learn the sentiment index for people's opinion about future housing price movement. Then we used this index to match the actual housing price movement in order to learn if people's opinion have prediction power of the future movement of housing price. Therefore, I think this "intermediate step" concept can contribute in forming the new methodology in economic empirical studies and fostering new knowledge

discovery in economic variable studies.

Second, this complexity exists in the situation that manually crafting the models becomes not achievable. In the traditional econometric economic studies, economists prefer to manually set the equations and models. The model selection is based on the theory and principles. When the scale of econometric study becomes large, it takes significant efforts and time to set such models manually and to run statistical testing. This doesn't include the work of updating new data sets and including new variables. On the contrary, machine learning has three advantages compared to the traditional econometric manually crafting models:

- The first is that instead of choosing one model based on the economic theories, a machine learning approach normally picks a few relevant algorithms, and sets the baseline model, to select the best performed results among these algorithms.
- The second is that it has the engineering pipeline concept in data processing, which is more organised in data storage, deploying, calculating and visualisation. Sometimes it can achieve automation in updating new data sets generated online and provide real-time output for decision making.
- The third is that machine learning algorithms have the ability to "discover a complex structure that is not specified in advance" [105]. Contrary to the econometric approach, machine learning does not focus on the parameter estimation, and the parameter estimation is rarely consistent, therefore not achievable by manual operation. This is why it has the advantage of dealing with complexity.

Third, the complexity also is reflected in the nature of non-linearity and network structure. Machine learning techniques are designed to deal with non-linear data patterns. A tree classification method can handle non-linear modeling by

growing a decision tree gradually when feeding the data set. It can reveal some structure that not easily discovered by a traditional linear regression model [148]. Other effective algorithms to deal with non-linearity include boosting, random forests, stochastic gradient boosting etc. Graph analytics is also applied in machine learning and deep learning to understand non-linearity and network structure, which has a great potential to tackle socioeconomic problems.

Economic Problems with Real Data

Computational socioeconomics has become a new research area in economics, which focuses on exploring socioeconomic data sets to understand how economic and social entities behave. The key element is to leverage the exponentially growing new social and economic data collected by applying the newly developed information technology. Like Buyalskaya et al. [20] said, the golden age of social science is coming.

Traditionally, economists have explored empirical study for more than half century with real economic data. Here comes a natural question, is there any difference between the current computational socioeconomic study and the traditional empirical study?

This question is answered partially by previous sections. New forms of data, such as text and image, are used for the computational socioeconomic study. New modelling, especially leveraged from machine learning algorithms, is explored with the challenges of new data, big data and complexity of the nature of the problems.

In this section, the focus is to add more elements to answer this question besides the aforementioned aspects.

First, the real data collected with current information technology is by mature multidisciplinary. Online products reviews, Tweets, videos contain people's opinions, purchasing behaviours, temporal and geographic information, social network

relations, demographics, etc. The nature of data requires a mind of collaboration of different disciplines to find research problems and solutions. The nature of data not only challenges traditional research methods, but also fosters fruitful discovery and creative approaches. Lots of examples in the current literature have proved this point.

The current COVID-19 pandemic forecasting is a good example to show how geo-spatial and temporal information, social structure, online social networks, public heath, epidemics, virology, economic incentive behaviours, policy study, big data, machine learning, physics, etc. work together to build the growth forecasting models [50, 63, 79, 86, 89]. It is self-evident that economics along cannot solve this forecasting problem, but multidisciplinary study can generate synergy and understanding closer to the real challenge.

Second, the division of micro and macro economics is getting blurry. For instance, Google search index contains individual's searching behaviour, which has a strong prediction power of macro-level trends. Such connection was not traceable by traditional demographic survey methods.

Third, real time policy making can be supported by the real time data collection to solve current problems. This will significantly improve data processing for policy makers. In addition, this will provide more opportunities for data economists to work with private sectors and fill the gap of providing strategic decision making for a real problem with controlled experimental environment.

People may argue that this type of research is not easily transferable to other cases with different environment and conditions. Varian [148] pointed out that "a good predictive model can be better than a randomly chosen control group, which is usually thought to be the gold standard". Due to the complex nature of socioeconomic problems, it is valuable to study a real problem with real data.

By doing this, we may discover some key indicators or interesting social patterns and behaviours that would be left in the noise with a more generalised model. In addition, the idea and methodology are transferable. It is worth duplicating the idea and methodology with a different social context to verify the transfer ability. This is a common practice in the traditional empirical economic study and logical to be extended in the current study.

Fourth, we can create a fast adaptive learning environment by opening up real data sets and code sharing. This is a common practice in the computer science research environment. This can stimulate the improvement of modelling by creating a standardised comparative environment. This also stimulate the formation of economics computing community and rapid development of computing empowered economics and social science.

The Role of Economists in Machine Learning

The role of economists to apply machine learning in empirical study is crucial.

First, economists have deep understanding of existing research problems and challenges in economics. If equipped with machine learning tools, they can find the joint opportunities to solve existing problems and even to establish new economic fields, as their interests and focus are within the development of economic domain. In comparison, computer scientists are keen to develop new machine learning algorithms and tools to solve emerging issues in order to push the edge of computer science.

Second, there are different focuses between current machine learning applications and economic problems [7]. The majority applications of machine learning are prediction and classification. Supervised machine learning is to predict an outcome of (Y) by calculating a set of features or covariates (X). Here the prediction doesn't mean forecasting. It means by learning from the training data sets, to

uncover generalised patterns. And test the estimation power by using independent testing data sets, in order to achieve high accuracy in predicting Y, by using independent testing sample X. Machine learning has the advantage to discover complex data structure during this data process using algorithm. The focus is on the generation of $\hat{y}$.

In most economic context, the prediction involves time-series data input and aims for a future forecast. In economic estimation applications, the focus is on *parameter estimation* [105]. It normally involves one estimation model, not as machine learning to test a few algorithms, and to get the best estimation of parameters β, to achieve $\hat{\beta}$, not $\hat{y}$. As pointed out by Athey [7] and Varian [148], the causal modelling hasn't been widely incorporated in the traditional machine learning practice, but highly used and discussed in the econometric field. Economists have great opportunity to develop the joint application of causal relationship study and machine learning to blossom the wide application in economic problems and policy studies.

Third, as the current and future challenges are becoming multi-disciplinary, there is a growing opportunity for economists to collaborate with experts from other domains and achieve outcomes that can't be done by people from one field. Therefore, economists need to adapt more into data science and familiar with data engineering concepts and workflows, in order to communicate and work effectively with other scientists and contribute its unique value in the group for much bigger and exciting challenges.

1.2.4 Validation Methods

The validation process is to test if a model can describe data consistently and with minimum error. The major difference of the validation process between machine

learning and econometrics is to choose in-sample or out-of-sample performance measures [148]. The reason of machine learning to use out-of-sample prediction is to avoid overfitting problem, which has a perfect performance in-sample, but fails badly when tested with out-of-sample data sets. Varian [148] pointed out that economists traditionally use in-sample measures for model validation is due to small data sets. Since the big data age has come, it is necessary to adopt the common practice in machine learning, to separate training and testing data sets. By keeping testing data sets out of model building, this can provide a good indication of prediction measurement in real data environment. There is a practical advantage in applying to social science study, such as, policy study.

1.3 The Research Problem—Understanding Housing Price with Traditional and Non-Traditional Data

In this section, the importance of the housing price research is introduced, followed by introduction of the existing literature. The main focus is the discussion of two specific research directions—housing price trend and appraisal prediction, and related detailed research problems.

1.3.1 The Importance of Housing Price Research

Understanding housing market behaviour is important and challenging.

Globally, more than half of the world's population lives in urban area in 2018. In 1950, only 30% was urbanised, and by 2050, 68% of the world's population is forecast to live in urban area [107]. Oceania in 2018 has reached 68% urban population. As the world is increasingly urbanised, the demand and supply of urban housing will become increasingly challenging. How to create a livable, affordable,

Table 1.1: US Median Home Values[a]

Type	2000	1990	1980	1970	1960	1950	1940
Adjusted to 2000 US dollars	$119,600	$101,100	$93,400	$65,300	$58,600	$44,600	$30,600
Unadjusted	$119,600	$79,100	$47,200	$17,000	$11,900	$7,354	$2,938

[a]Source: `https://www.census.gov/data/tables/time-series/dec/coh-values.html`

greener and smarter living space is a huge challenge for the current and future urban planning, housing, and economic policy. The study of housing market should only become more and more important as the urban population grows.

Firstly, the importance of understanding housing market can be attributed to the significant economic value of the housing sector. It plays an important role at household level to store and generate wealth. Table 1.1 shows the growth of median home values in the US from 1940 to 2000. It is clear home value doubled in every 30 years. This could secure the living standard for retirement if the household can finance and manage the property ownership since middle age. Figure 1.1 shows the similar trend in Australia. From 1980 to 2015, housing prices continuously overtook the growth of Headline CPI. All these figures could show the stable growth of home value in the long run. To make the comparison more bluntly, statistics shows the median net wealth of elder people who owns a property is 22 times more than people in the renting condition in 2017[2,3].

Secondly, the importance of understanding housing market comes from the connection with macroeconomic movement. The housing market has a strong impact on the economy worldwide. Housing-related industries represent a significant amount of the Gross Domestic Product (GDP): 15-18% of GDP in the United States, nearly 15% of the GDP in the European Union, and 13% of the GDP in

[2]`https://www.abs.gov.au/statistics/people/housing`

[3]The median wealth of property-owning households with at least one of the occupants over 65 years old was over $934,900, whereas renting households under similar conditions only had $40,800.

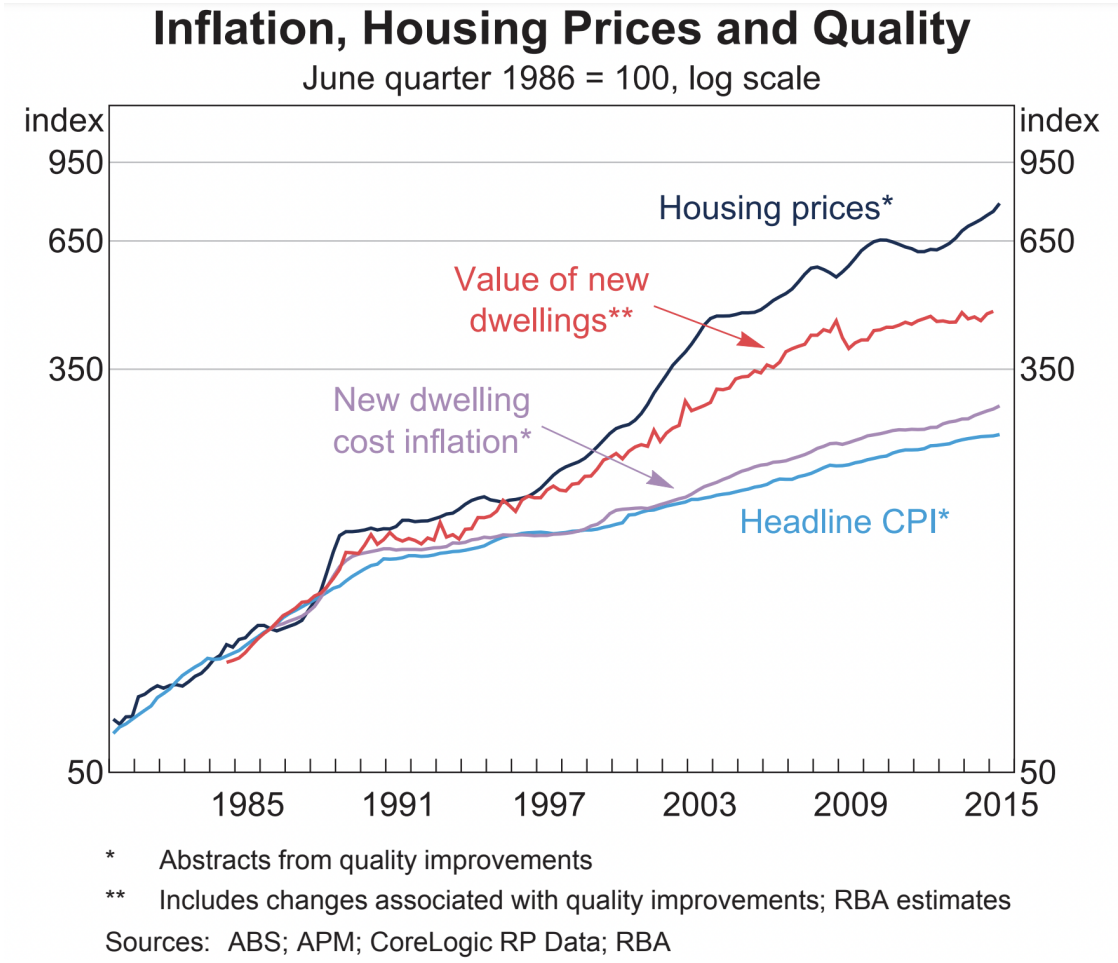


Figure 1.1: Australian Inflation and Housing Prices[a]

[a]Source: `https://www.rba.gov.au/publications/bulletin/2015/sep/3.html`

Australia[4]. This shows that housing industry is a significant leading indicator of a nation's economy [88] and important for economic and financial stability and growth. Figure 1.2 shows the significant amount of new loan commitments generated since 2004. The loan amount as a direct connection with the economic cycles. Housing market and macro-level economy are interconnected. In 2008, a downturn in the US housing market triggered a financial crisis that spread to the whole

[4]https://www.nahb.org/News-and-Economics/Housing-Economics/Housings-Economic-Impact/Housings-Contribution-to-Gross-Domestic-Product

Figure 1.2: New loan commitments, total housing (seasonally adjusted), values, Australia[a]

[a]Source: `https://www.abs.gov.au/statistics/economy/finance/lending-indicators/latest-release`

United States and to the rest of the world[5]. This further triggered bankruptcy of global financial institutes, millions of job loss, and eventually the deepest recessions since the Great Depression in the 1930s. The 2008 global recession showed strongly how the housing market and global economy are interconnected. Prediction and prevention of such event in the future is the reason for housing market research.

1.3.2 Housing Research Challenges

Despite the importance of housing market research, our understanding in this field is still very limited, due to challenges from different aspects. We provide an

[5]`https://www.rba.gov.au/education/resources/explainers/pdf/the-global-financial-crisis.pdf?v=2022-03-07-14-51-42`

overview of several such challenges in this subsection.

First, housing market prediction is one of the biggest challenges in the economic cycle research. The 2008 global recession has strongly showed that even the most sophisticated existing economic models failed to foresee the big crisis, and this crisis has been developing for at least two years before reaching the transition phase of crisis. One of the reason explains the difficulty of housing market prediction is that the measure of human behaviour, especially irrational behaviour has not been developed and tracked. How can we understand human behaviour at the individual level and join the connection with macro-level of economic cycles? How can we understand the mechanism of social network behaviour which contributes to the market movement?

Second, the data collected for economic analysis is always delayed and involved huge costs, such as most countries run a Census every five years. This can have a huge impact on the understanding of economic problems, such as understanding housing market. Other popular traditional methods for data collection include interviews, case studies, questionnaire. All these methods are slow and costly for data collection. It limits the quantity and detail level for data collection, therefore, economists focused on the sampling methods and econometric methods for reaching robust results with limited samples. Nowadays, information technology collects data on all kinds of devices, at any granular level, and at any time. We could tackle the challenge of limited data sets by leveraging new information technology with novelty.

Third, housing market provides a heterogeneous condition since each house has a unique location relative to its surrounding environment and other characteristics. This brings a challenge for housing appraisal. Traditionally, economists established a clever solution to target the same house sold twice or more times to

evaluate the appreciation or depreciation of the housing price [24, 25, 27]. However, this method has a few limitations. It is biased to calculate the housing price movement market-wide, and it fails to achieve individual housing appraisal due to the lack of sufficient information for non-sold houses or houses sold for once. This is the limitation for most housing price models due to the granularity level of the information. The second solution given by economists is called hedonic pricing model. The challenge is to find new ways of understanding housing price movement, by leveraging machine learning and artificial intelligence algorithms.

Fourth, to understand factors that influence home value is a multi-disciplinary task. The challenge is to understand housing value from different angles, such as urban planning, socioeconomic impact, transport, health, environmental sustainability. How can we look into different fields and capture those factors with high impact? How should we design the data and information flow to construct a complex model to understand factors of home value? These challenges need to apply technology from the recent development of data science.

Fifth, how can we leverage big data, new types of data to analyse the housing price problem? These are the cutting-edge research direction. It is challenging in the way to find the usable data and create a novel method to link with the housing price problem. It is also challenging in mastering the new technology involved in solving these problems. The current discovery of new types of data include geospatial data sets, satellite image data, textual data, travel records, mobile phone records, etc. Big data includes a combination of new types of data in a large scale.

1.4 Contributions

The author has made a series of research attempts to tackle the five challenges discussed in the previous section, from the angle of introducing new types of data sets and studying them with machine learning tools. The contributions are summarised as follows.

1.4.1 The Effect of Regional Economic Clusters on Housing Price

Housing appraisal is a fundamental topic in the housing market research. An accurate appraisal leads to rational decisions for home buyers that could benefit them in the long run, at the same time preventing housing bubbles if such irrational behaviour accumulates at a macro-level in a longer period. Therefore, an accurate housing appraisal is crucial for banking industry and economic stability as a whole. However, the current practice in mortgage industry is to hire a professional property valuer for housing appraisal which is expensive, time consuming and inconvenient. In the research field, there are limited investigation to understand the value of a good location and what factors contribute to a high quality location. This research challenge can be dealt with analysing new type of data sets that collected by the current information technology.

The key contribution of this study is that this is a pioneering study in Australia in the housing appraisal field. To our knowledge, this is the first study in Australia to use geospatial information, large variety of data sources, very large social data sets, and a high dimensional feature space, to implement a supervised machine learning pipeline for housing appraisal. The author conducts an investigation of housing factors from different aspects, mainly focusing on the economic impact

of geospatial nodes at different level. Firstly, investigate the neighbourhood characteristics by utilising free open data from OpenStreetMap, which goes beyond the traditional economic data sets. Secondly, use data scraping method to collect ranking information of point of interests (POI) through online location ranking system, such as TripAdvisor. Thirdly, explore regional impact by identifying regional economic clusters, which is the novel contribution of this study. Fourthly, deploy rich demographic characteristics to understand the impact of dwellers to create the suburb characteristics.

By consolidating with influencing factors from each aspect, we build a housing appraisal model, named HNED, including housing features, neighbourhood factors, regional economic clusters and demographic characteristics. Specifically, we introduce regional economic clusters within the metropolitan range into the housing appraisal model, such as the connection to CBD, workplace, or the convenience and quality of big shopping malls and university clusters.

By building the baseline of linear regression, we construct a series of algorithm applications, such as support vector machines, multi-layer perceptron, k-Nearest neighbour, XGBoost. When used with the gradient boosting algorithm XGBoost to perform housing price appraisal, HNED reached 0.88 in R^2. In addition, we found that the feature vector from Regional Economic Clusters alone reached 0.63 in R^2, significantly higher than all traditional features.

Other contributions include the research output contains valuable implications for different stakeholders, such as home buyers, councils and urban planners, housing developers.

1.4.2 Understanding Housing Market Behaviour from a Microscopic Perspective

To study the boom and bust of business cycles, otherwise known as economic cycles, is one of the greatest challenges in Macroeconomics. It has great impact on the fiscal policy, and the economic performance as a whole. The challenge shows as the great uncertainties about factors for economic growth and business cycle. This can't be solved by the existing analysis within the neoclassical framework. Current disputes in theory are concentrated in the assumptions departed from perfect rationality under uncertainty. As Simon [128, p. 35] asserted, there is only one settlement for such disputes, which is to focus on the empirical study of how human make decisions and solve problems, at the microeconomic level.

This is the first attempt to use newspaper articles for sentiment analysis for understanding housing price movements in Australia, to our knowledge. The boom and bust of housing market is a sub question in the economic cycle research in Macroeconomics. Therefore, the idea is to study housing decisions at the micro-level, in order to find solutions for this decades-long challenge. How to capture the micro-level human decision-making behaviour is the key to bring new ingredient to this old problem. Information technology sheds light on the method to capture human's thoughts and behaviour. This paper attempts to bring new understanding of housing market behaviours by leveraging the big data and new types of data recorded in the information age. More specifically, the aim of this paper is to study the macro-level housing market cycle problem from a microscopic perspective, with the support of newly available micro-level data sets, especially through text and sentiment analysis.

The micro-level behavioural data sets are collected through the scope of online local newspaper articles. The idea to use local newspaper instead of national or

city level newspaper is to increase the granularity, in order to study the behaviour at a more detailed level. This could help to understand the interrelationship and dependency among different locations.

Sentiment analysis is applied to study local newspaper articles discussing real estate at a suburb level in inner-west Sydney, Australia. We calculate the media sentiment index by using two different methods, and compare them with each other and the housing price index. The use of media sentiment index can serve as a finer-grained guiding tool to facilitate decision making for home buyers, investors, researchers and policy makers.

In this problem, the macro-level housing market cycle problem has been transferred into a suburb-level housing market behaviour problem. With the possibility of discovering new indicators, relationships, patterns, and behaviours at a suburb-level, the goal is to find the link between suburb-level and national-level movements, in order to reach a better understanding of the housing market behaviour. The novelty of this study is to compare two indexes, one is a human behaviour index of media sentiment about housing market; the other is the housing price index calculated through a periodic investigation of the recent sold property records. The other contribution is the development of a more general sentiment index compared to the existing literature.

1.4.3 New Behavioural Big Data Methods for Predicting Housing Price

This is a pioneering study in Australia to establish the theoretical framework for behavioural study of applying big data method for housing market forecasting. By discussing the gap between macro-level economic cycles and microeconomic behaviour theories, the author argued the possibility and necessity of using web gen-

erated information as data sets, in the purpose of filling this gap. In the theoretical framework, three approaches are raised and discussed to find online behavioural data sets for housing market forecasting problem. The information revolution and big data methods have served as a new lens to measure economic factors with novelty, apart from traditional methodologies. In addition, this research provides the theoretical discussion to use big data methods to study behavioural economics.

The three approaches discussed in this study include: Google Trend Index, text stream from online news articles, and individual opinions from Twitter and Facebook. Empirically, big data methods can be used in forecasting the housing market cycle in Australia. Specifically, this study made contribution to bring one new variable in the time series auto-regression model to forecast housing market cycles, which is auction clearance rate. This is a strong indicator for housing market trend in Australia due to its popular on-street auction choices from home sellers.

1.5 Thesis Structure

The rest of this thesis is structured as follows. Chapter 2 introduces background knowledge of Australian housing market in detail. Chapter 3 to Chapter 5 present my main research output and contributions. Specifically, Chapter 3 proposes a point level housing appraisal model named HNED, including housing features, neighbourhood factors, regional economic clusters and demographic characteristics. Chapter 4 proposes a sentiment analysis for Australian local newspapers to understand the housing price movement. Chapter 5 inspects the recent advances in big data technology, and proposes theoretical framework to leverage this technology to study housing market behaviour. The entire thesis is summarised with the achievements and touched on the future work in Chapter 6.

CHAPTER 2

LITERATURE REVIEW

The study of housing price modelling is not new, although historically housing price studies have been somewhat non-mainstream until recent years. In this chapter, the focus is on the reviews and comments on the research development for housing price models. Section 2.1 discusses the literature reviews on the development of traditional housing price theories and empirical studies. In Section 2.2, an overview of data science approaches on housing market problems is discussed. End of each section, the gap in the current literature which leads to some of the work in the following chapters is discussed.

2.1 Traditional Housing Price Models

Traditional housing price modelling is discussed in this section. The literature development in the financial market domain is discussed. Two major housing price models are compared. In the end, the research problems and challenges extend to the current information age are illustrated, which leads to the discussion of Section 2.2.

2.1.1 Existing Knowledge of Macroeconomic Housing Market Forecasting

Inheritance from Capital Market Theory

The study of housing market is within the field of financial economics, specifically within the theoretical framework of efficient-market hypothesis (EMH) [48]. The primary function of capital market is to allocate capital resources through ownership. Capital ownership decisions are based on the signals contained in the prices

of capital stocks. Ideally, all the information about the stock should be fully reflected in the market price. Then the market is considered fully efficient. In the capital market theory, there are significant amount of deep discussions of the definition of "fully reflect", which relates to the work of defining the level of efficiency of the stock market both theoretically and empirically [99]. In the empirical study of testing the stock market efficiency, researchers define three level of efficiency based on the amount of information reflected in the stock price. For example, the weak form of efficiency means the stock price only reflects historical prices. The semi-strong form of efficiency means the stock price also reflects public available information released by the company, such as earnings announcement, splits, etc. The strong form tests are set based on if any information relevant to the pricing formation is reflected in the stock price, even if the information is only accessed by a small amount of investors, prior to any company announcement. This detailed definition brings guidance for empirical testing. This is a main stream research area since 1970s. The related study dominated the capital market research for more than two decades.

The information efficiency is essential for institutional and individual investors to make high quality trading decisions. Therefore, the pricing theory is important in explaining the decision making behaviour in the capital market. Since 1990s, more research outputs started to question the efficiency level of stock market based on the detected trading behaviours and trading prices that cannot match the EMH. Researchers did some pioneering work in discovering overreactions, anomalies in the stock market behaviour [31,32,72]. This raised a key question that leads to the birth and blossom of a new economic field, called behavioural economics, which is still a very alive and exciting multidisciplinary research field nowadays. This field studies the rationality and decision making mechanism of human being by creating

novel psychological experiments and econometric empirical measurement, at the micro-level.

Adoption in Housing Market Study

Interestingly, not many studies have been done on the housing market efficiency behaviour in the literature, compared with the enormous work been done in the stock market. One of the earliest empirical tests of housing market efficiency for single-family home suggested the housing market is not efficient [25, 26], which means the current housing market performance contains valuable information for the future prediction. Pollakowski and Ray [114] continued the study of housing market efficiency in depth of the interaction of different geographic locations. Less housing market efficiency research is done due to a few reasons. One major reason is that relevant housing information is not clear to define, compared to the companies in the stock market, especially it is hard to trace the clear time frame before and after information release. Stock market has the advantage of both clear public information releasing system and fast transfer of ownership, to help the empirical information testing easy to conduct. Therefore, to test housing market efficiency needs breakthrough to match its own market characteristics.

As housing market is different from stock market in many ways, one of the challenges is to set up the empirical testing methodology for housing market efficiency. Case and Shiller [25, 26] were one of the first to enquire the efficiency of housing market. They established the methodology to test the efficiency level. Their method can be concluded as repeated sales to calculate housing price index, which will be discussed in the following part.

Traditional Housing Price Index Models

The first housing price index model to be discussed is called weighted repeat sales. In the housing market efficiency study [25], point level data of 40,000 sold homes were collected during the period of 1970-1986, in four metropolitan areas in US. They performed weak form test of housing market efficiency. In detail, they found the change of log real price index tends to be followed on the same direction in the next year. The importance of this literature is that it established the enquiry of housing market efficiency and started the study of housing price. It is included gradually into the main stream capital market efficiency research. Unfortunately, the housing market efficiency has not been given sufficient attention until the global financial crisis in 2008. Since 2008, researchers have continued contributing in this direction [45, 143].

The other housing price index developed is called hedonic housing price modelling. Hedonic modelling is trying to measure the market value of housing based on the service quality bundled [78]. The specific factors involved include aspects of "the dwelling units, structures, parcels, and micro-neighbourhoods". This model attempts to include housing qualities, conditions of neighbouring houses, the quality of the streets, road condition, block landscaping, etc. This set up the hedonic approach of housing price modelling.

Compared to repeat sales model, hedonic model has the advantage to include more housing related information in the model. Though researchers found it is difficult to deal with the statistical problems when both the computation power and statistical methods were limited, it provides great value of setting the theoretical framework for the housing price model in nowadays.

Coherent with the stock market efficiency study, one of the essence is to test the level of information being represented in the housing price. Therefore, the level

of information is the direction to push the research into the new frontier. This is the opportunity and challenge for the information age housing research.

Based on the hedonic modelling framework, the gap can be identified is the location information that includes the quality and quantity of economic and social activity leveraging such location. The relationship between the location and the housing price can be explored. For example, is the housing price movement due to the price increase, or improved neighbourhood, or the frequency of the trading behaviour, or the scarcity of the land [102]?

2.1.2 The Gap Identified in the Current Literature

In this section, two aspects from new inputs of economic theories are discussed as identified gaps in the current literature that formulating ideas of housing market research in this thesis.

First, in Porter's view [115], historically, geography was considered central in the economic theory. This includes Adam Smith and Marshall's *Principles of Economics*. However, the role of location was not given enough attention until Porter's paper to address location, clusters back to microeconomics. Relatively, the factors of geography and location in the context of economic growth and policy analysis has not been naturally integrated and discussed in the economic study. One of the practical reasons is the lack of both point level big data and models, tools, computing power to analyse such data. Now is the time shifting our attention, to include economic geography in the mainstream economic analysis, since more data can be collected with geography and location information, plus the tools are available to facilitate such integrated analysis. Housing market analysis can not ignore the factor of location, the power of economic clusters in a specific geographic environment. The productivity and prosperity of a location impact both the decisions

of locations for where to establish business and homes.

Second, the impact of behavioural economics also cannot be ignored in the housing market study. Kahneman and Tversky demonstrated in [74] that the decision making has been proved systematic apart from the rational behaviours. This brings the needs to study related behaviour in the housing market. The study can be guided by the development of behavioural theory. The lack of empirical study of the behavioural decision making of home buyers, investors, mortgage industries, construction industries, etc., can be addressed with novel design of data collection and analysis. This could potentially bring value to understand the eco-system and behaviours of housing market. The challenge and also the gap is to connect both micro-level and macro-level for the behavioural analysis.

2.2 Machine Learning Based Housing Study

Due to advances in information and communication technology in recent decades, the availability of much larger and higher granularity data sets became possible [2, 17, 118, 120, 138]. This, coupled with significantly improved compute power – including advanced computing architectures such as cloud and high performance computing – resulted in large machine learning models with thousands of machine-tuned parameters being practically deployable for research purposes. As a result, machine learning research has proliferated in the 21st century, and machine learning based predictive modelling has achieved phenomenal success in a variety of applications [55, 66, 111]. Of the models used, tree based models and deep neural network based models have been particularly noteworthy in a variety of applications [3, 119, 121]. In this section, we provide a review of machine learning technology and applications for housing price appraisal literature.

Much of the recent key machine learning based literature methodologically

belong to one of more of the following categories: usage of "user side big data", such as the usage of Google Trends dataset or analysis of twitter data [11, 95]; natural language processing [53, 94, 124], such as using sentiment analysis on newspaper archives; image processing [129, 134], such as the usage satellite image data; and combining diverse types of features in the machine learning model [54, 80, 157].

2.2.1 Housing Appraisal Study with Geographic Data and Information

In this section, we discuss the development of housing appraisal methods in both traditional housing market research and computer science research fields. Both fields involved recent development of methods in dealing with spatial data. Spatial autocorrelation and spatial heterogeneity recognised in both fields are the two main challenges in models involving the spatial data. In addition, high-dimensionality is recognised in the computer science field as more features involved in the model compared to the domain model. Both fields realise that the better prediction relies on introducing geographic characteristics.

Recent housing research developed spatial autoregressive model to incorporate the dependency between nearby houses [91, 109] based on the traditional hedonic framework. Bency et al. [12] asserted that the limitation of such method are two folds: first it requires domain knowledge to set number of geographical nearest neighbours included in the pricing model, second the model assumes linear relationship.

In the computer science field, more focus is about how to incorporate newly available data into the house price prediction model. Most of these new data helps to expand the understanding the geographical characteristics towards housing valuation. The nature of these newly available data is often beyond the angle of

traditional economic variables, therefore, new methods also introduced into the real estate valuation field.

Image as Economic Data

Some researchers have used satellites image as an indicator for economic outcomes prediction, such as Google Street View to predict income level [59], the illumination of night to interact with economic output [64]. An overview of relevant literature has been provided by work [39].

For example, satellite image [12], street-view image, house image from real estate websites [96, 116, 158] are used for house price prediction [33, 82, 87]. Other paper investigated how the satellite image and street view can help to understand the neighbourhood [106], demography [56] and commercial activities. These results are also highly relevant to housing appraisal, though they didn't investigate this question directly.

The other possibility is to use the photo images in the housing market advertisements to estimate the housing prices.

Location-Based Social Network

Besides satellite, the location-based social network provides us the flow of locations that patterns of human traffic behaviours can be analysed [161].

Besides image data, other types of data are used for housing appraisal beyond the traditional scope of economic modeling. OpenStreetMap data was used for collecting Point of Interests in the neighbourhood [33, 52, 131]. Mobile phone data was used to empirically study the human activity and urban vitality [16, 34]. Taxi-cab trajectory data was mined to estimate neighbourhood popularity, in order to understand the geographic factors for housing appraisal [52]. Google search index was used for housing prediction model to understand how people's attention of real

estate would influence the future housing price [153]. Text data from real estate related news was studied to learn how sentiment is related to housing price [84,130].

These data applications are innovative and inspiring for improving housing price modeling. However, most of the newly applied spatial data are focused on the discovery of neighbourhood characteristics, such as crime perception [18], walkability [30], cultural influence [65]. These applications neglect the understanding beyond the neighbourhood. And this is the main focus of our work.

2.2.2 Behavioural Housing Study

Instead of using traditional sampling techniques which may subject to bias, we can enjoy the availability of larger data or even nearly complete data sets (big data), which typically leads to higher accuracy in behavior prediction. Besides this advantage, we emphasize on the nature of the newly available data.

The growing online data has gradually enlarged the variety of economic variables. Pioneering researchers started adventure in private sector companies who have first-hand real-time online data, such as Google, MasterCard, Federal Express, UPS [29]. Also, credit card transactions and eBay online auctions nowadays become part of the economic variables [41,100]. Furthermore, consumer behaviours can be analysed with mobile internet data sets from major telecom providers [68].

Google Trend Index

Google Trend Index can be regarded as a new type of data. Choi and Varian [29] built prediction models using Google search engine trend index as an indicator and found it effective to do near-term economic prediction. Many researchers followed this idea to implement in different areas such as automobile market, entertainment industry, tourism industry, labour market [4, 23, 60, 101], and found significant

improvements in prediction accuracy.

Particularly, in the real estate area, Wu and Brynjolfsson [153] did a pioneer study of housing market cycle prediction in the US using Google Trend Index and showed a significant improvement in prediction accuracy, compared with the state-of-the-art prediction results. McLaren and Shanbhogue [101] also showed the power of using Google trend index as an indicator in predicting house price index in UK. They also provided a comprehensive discussion of the searching key words selection.

Unfortunately, the aforementioned studies [101, 153] do not go into the suburb-level, due to the current Google trend's degree of releasing. The housing market study may be extended to the postcode-level when Google is ready to release the data to the public. Another concern is that it may not be accurate due to the reason that people may move to different suburbs and reduce the location precision.

Recently, researchers started to exploit image and language information to study economic behaviour. The nature of image and language information data is beyond the traditional econometric linear regression processing method.

2.2.3 The Use of Text Analysis for Housing Market Studies

Analyzing texts in the social media big data content has a rich literature, such as work [113, 135, 149, 156]. In economics, the early application of natural language processing is to study the feasibility of predicting stock market prices. The sentiment analysis is used in processing the data collected from Blogs, News, Twitter feeds, etc.

Literature about the relationship between sentiment analysis and stock returns is quite rich. The sentiment analysis in stock market started before the big data era. The arguments started with the challenge in behavioural finance about the

efficient market hypothesis (*EMH*). *EMH* assumes all the publicly known information about a company has been reflected in its stock price, which means there is no opportunity to arbitrage the difference between the current stock price and future stock price based on the available information. On the contrary to *EMH*, researchers started to discover overreactions, anomalies in the stock market behaviour (see [32,72]). Historical stock prices showed predictive power for the future prices, which means current stock prices of companies don't fully reflect intrinsic value of companies. There exists opportunities to arbitrage based on publicly known information.

Following the investigation of behavioural finance, Baker and Wurgler [8] used financial indicators to form the sentiment proxies, but these indicators are not available in the housing market [130]. With the study of the predictive power of online chat, the nature of data for sentiment study has changed into textual analysis. Relevant literature includes [5, 15, 58, 61, 123, 160]. Particularly, Schumaker and Chen [123] predicted the stock price with analysis of textual financial news articles. Bollen, Mao and Zeng [15] applied Granger causality analysis and Self-Organizing Fuzzy Neural Network to analyse Twitter feeds. Based on the public mood status generated from Twitter feeds, they predicted the daily changes with 86.7% accuracy.

Housing Sentiment Study

In [130], Soo presented an early study on applying textual sentiment analysis on housing market area. The study found that newspaper sentimental analysis can be a good predictor for the future housing prices in the US. Research work in this thesis follows the data collection method used in [130], and goes beyond [130], in that I use local suburb-level – instead of city- or country-level – newspaper text as the source to define sentiment on housing market. There are limited studies [62, 137]

which use Twitter feeds as a sentimental indicator, which is however widely used in stock market. This may become an inspiration for the future research work to investigate the feasibility of predicting housing prices using local Twitter feeds.

Forecasting model with online news

The key element to predict housing market cycle with the information from online news is to build the relationship through "sentiment".

Barberis et al. [10] studied stock market investors sentiment on the stock performance news, such as earnings announcements. And he generalized the sentiment as "overreaction" and "underreaction" with empirical evidence. Tetlock [139] also quantified the interactions between Wall Street Journal column and the stock market performance. A pessimistic view can drive down the market prices as an overreaction, which will follow a reversion to fundamentals.

With the development in the stock market sentiment study, the housing market study follows the trend. Soo [130] did a pioneer study of American housing market sentiment study by analysing the newspaper tones as "positive" or "negative".

The textual analysis is used to quantify the tone of financial documents. The standard dictionary-based method is used to count the raw frequency of positive and negative words in a text. Soo [130] prepared a housing dictionary and presented the calculation of the overall tone of housing market news sentiment by:

$$S = \frac{\#pos - \#neg}{\#totalwords} \tag{2.1}$$

Kou et al. [84] followed this study and calculated the media sentiment index with two different dictionary methods at a suburb-level in Australia. This approach can be continued by extending the suburb study to a macro-level.

Forecasting with text streams from Facebook and Twitter

Articles directly using Facebook or Twitter text streams to predict housing market cycles are very limited in the literature. But there are quite a few research papers using the idea of mood or sentiment analysis to predict stock market behaviour [15, 104, 160]. We could reasonably assume that the nature of the prediction algorithm for housing market will be similar to the stock market.

How to find emerging topics is a challenge in social media analysis. Early detection can improve the understanding of people's behaviour towards market movement. Novel tracking method can be found in work [69]. Big data analysis could be another challenge. Some technique and solutions are described in [92, 159].

It is reasonable to assume that the selection of Twitter feeds need to address the location difference, unlike the stock market prediction, because the housing market cycle shows a strong trend difference world widely. But Shiller [126] has shown the big glamorous cities experienced massive boom within similar timeframe (1999-2014). Australian cities, Sydney and Melbourne are recognised as the glamorous cities and following the global boom trends. Therefore, it may be a good direction to test the Tweets in Australia and world-wide for the location hypothesis.

2.2.4 The Gap Identified in the Current Literature

In this section, the focus is to illustrate how information technology can fill the gap of bringing the role of location and behavioural economics in the discussion of housing market.

First, data, tools in machine learning can be used for the housing market study. Especially, geographic information can be easily collected and integrated in the housing study. The location information not only includes the geography point information, such as longitude and latitude of the housing, but more importantly,

layers of neighbourhood information can be included into the analysis as different level of factors. These layers can be defined as the immediate neighbourhood which can be reached by walking, and middle level of neighbourhood which can be easily traveled by car, and the metropolitan layer, the high impact economic clusters, such as CBD, which can not be ignored when establishing the relationship between the house and its embedded environment. Other type of layers can also be added into the housing study, such as the satellite image and housing detailed image from real estate websites. This may improve the accuracy of the characteristics of the housing attributes, and geographic factors.

Second, information about opinions, comments, thoughts and behaviours can be collected and analysed through the current development of data scraping and natural language processing technology. This can be studied in the housing market analysis, to have empirical evidence about how at the macro-level, people's behaviours and opinions interact with the trend of the housing market.

CHAPTER 3

THE EFFECT OF REGIONAL ECONOMIC CLUSTERS ON HOUSING PRICE

3.1 Chapter Abstract

A good location goes beyond the direct benefits from its neighbourhood. Unlike most previous statistical and machine learning based housing appraisal research, which limit their investigations to neighbourhoods within 1km radius of the house, we expand the investigation beyond the local neighbourhood and to the whole metropolitan area, by introducing the connection to significant influential economic nodes, which we term *Regional Economic Clusters*. By consolidating with other influencing factors, we build a housing appraisal model, named HNED, including housing features, neighbourhood factors, regional economic clusters and demographic characteristics. Specifically, we introduce regional economic clusters within the metropolitan range into the housing appraisal model, such as the connection to CBD, workplace, or the convenience and quality of big shopping malls and university clusters. When used with the gradient boosting algorithm XGBoost to perform housing price appraisal, HNED reached 0.88 in R^2. In addition, we found that the feature vector from Regional Economic Clusters alone reached 0.63 in R^2, significantly higher than all traditional features.[1]

[1]This chapter is based on the following article:
Jiaying Kou, Jiahua Du, Xiaoming Fu, Geordie Z. Zhang, Hua Wang, and Yanchun Zhang. "The Effect of Regional Economic Clusters on Housing Price." In Australasian Database Conference, pp. 180-191. Springer, Cham, 2021.

3.2 Introduction

The housing market has a strong impact on the economy worldwide. At the national level, housing-related industries contribute to 15-18% of the Gross Domestic Product (GDP) in the United States, nearly 15% of the GDP in the European Union, and 13% of the GDP in Australia[2]. The housing value also represents a major part of household wealth: the total value of residential dwellings in the U.S.[3] was $33.6 trillion at the end of 2019 and that in Australia[4] was $7.1 trillion in June, 2020. These numbers show that housing is a leading indicator of a nation's economic cycle [88] and important for economic and financial stability and growth [14].

Housing is crucial not only at national level, but also at each individual or household level, as it is deeply involved in each individual's economic and social life. Housing is one of the most popular topics among people's casual conversations [36]. Real estate is the biggest assets for most households and also one of the strongest factors as a financial assurance to afford a comfortable retirement life. In 2017, the median net wealth[5] of property-owning households with at least one of the occupants over 65 years old was over $934,900, whereas renting households under similar conditions only had $40,800, showing almost 23 times difference.

Housing appraisal is crucial for the housing market. An accurate appraisal leads to rational negotiation and decision making and thus helps preventing home buyers from buying over-valued homes. Housing appraisal is also highly relevant to financial stability as most banks require a specific house valuation process to decide a healthy mortgage amount. In practice, however, it is difficult for home buyers

[2] https://www.nahb.org/News-and-Economics/Housing-Economics/Housings-Economic-Impact/Housings-Contribution-to-Gross-Domestic-Product

[3] https://www.zillow.com/research/us-total-housing-value-2019-26369/

[4] https://www.abs.gov.au/statistics/economy/price-indexes-and-inflation/residential-property-price-indexes-eight-capital-cities/latest-release

[5] https://www.abs.gov.au/statistics/people/housing

to access information during house hunting and price negotiation stage because hiring a professional property valuer is expensive, time consuming and inconvenient. These difficulties pose a strong need for a timely, accurate, automatic, and affordable housing appraisal system.

Despite that housing appraisal is important in both macro and micro economics, "housing price remains as much art as science" [87]. The understanding of housing price is still very limited. Existing studies focus on housing attributes (e.g., size of land, number of bedrooms) but largely ignore the relationship between a house and its surroundings. Traditional econometric models have revealed a strong spatial correlation to housing price, which can be explained with two theories [91]: (1) the spillover effect between regions–when physical and human capital, or technological improvement concentrates in one region, it will naturally have a positive impact on its neighbouring regions [46]; (2) unobserved or latent geolocational factors.

Recent availability and appreciation of social, economic and geographic data have enabled researchers to trace the spillover effect on human capital or technological breakthroughs, and to discover unobserved or latent factors. Housing appraisal becomes more viable thanks to granulated geo-spatial rich information available through multiple online resources, such as satellite, street views and housing images, and map data. With the support of real world data, we can empirically investigate how people make decisions. These socioeconomic data sets are multi-source, heterogeneous, and high-dimensional.

Current economic theories for housing appraisal came in the era before massive data could have ever been collected and evaluated. As mentioned in [88], insufficient data led to the lack of understanding the strong impact of housing sector to the whole business cycle. Traditionally, economists adopt hedonic approaches to evaluate a property based on its housing attributes and neighbourhood related

characteristics [21, 22, 40, 142]. Although proved to be useful, these methods failed to explain substantial portions of housing price variability.

Most of the recent development focuses on finding new spatial factors by leveraging the new available online data and machine learning methods to reveal the unexplained elements [73, 93, 113]. These new findings have shown that housing price is correlated with safer environment [33, 35], intangible assets from its neighbourhood, and associations with neighbouring houses [87], design [116], culture [65], Point of Interests measure [52], etc. The new development is mostly from the perspective of neighbourhood characteristics, and is normally within the range of 1 to 2km from the house. However, only exploring the near neighbourhood has a few limitations. First, certain living functions can't be fulfilled in the near neighbourhood and these functions are not captured in the previous housing price models. For example, shopping malls, hospitals, universities are strategically located to service at regional level, or national level, not at suburb level. But these functions do influence the demographic distribution in the nearby suburbs and hence influence the housing value. Second, these regional services are economically highly concentrated clusters. For example, a shopping mall can contain 500 shops. A university campus can service 20,000 students. Therefore, economic value and service activities are highly concentrated in these nodes, and would influence the price of houses that are beyond their immediate neighbourhoods.

How can we expand the investigation and identify key factors beyond the immediate neighbourhood?

A few studies [12, 82] have shown that enlarging the neighbourhood area can improve housing price prediction. However, these studies were based on satellite image without further investigation of influential visual features, or area beyond the neighbourhood. Therefore, we can't identify whether the improvement is due

to merely enlarging the neighbourhood area, or due to the inclusion of new features in the calculation.

To expand the investigation beyond neighbourhood, we can either merely increase the neighbourhood area, or identify and add new key features at the metropolitan level. Increasing the neighbourhood area is not an ideal approach, as it can increase computation significantly without providing additional insights. Therefore, we take the second approach as finding the new key features. It is more challenging, but with the reward of less computation and potentially bringing implication values to home buyers, investors and urban planners.

Our approach extends the relational closeness by investigating the economic proximity. This establishes an economic closeness between the household and the place–economic cluster. This paper aims to study the intangible value of a house beyond neighbourhood value by evaluating the relationship between house and existing regional economic clusters. Specifically, we identify economic clusters by some significant categories, such as CBD, shopping malls, universities. By consolidating with other influencing factors, we build a housing appraisal framework including *Housing* features, *Neighbourhood* characteristics, regional *Economic* clusters and *Demographic* characteristics, called the *HNED* model. This approach may potentially help decision-making for home buyers, property investors and urban planner. It may also indicate solutions for affordable living without compromising the essential needs.

The rest of this chapter is structured as follows. Related work is discussed in Section 3.3. Section 3.4 explains the conceptual framework and main factors in detail. Section 3.5 explains the methodology and experimental settings. Section 3.6 discusses the results. Section 3.7 deals with discussion and implications. Finally, concluding remarks are offered in Section 3.8.

3.3 Related Work

In this section, we discuss the development of housing appraisal methods in both traditional housing market research and computer science research fields. Both fields involved recent development of methods in dealing with the spatial data. Spatial autocorrelation and spatial heterogeneity recognised in both fields are the two main challenges in models involving the spatial data.

Recent housing research developed spatial autoregressive model to incorporate the dependency between nearby houses [91, 109] based on the traditional hedonic framework. Bency et al. [12] pointed out the limitation of such method are two folds: first it requires domain knowledge to set number of geographical nearest neighbours included in the pricing model, second the model assumes linear relationship.

In the computer science field, more focus is about how to incorporate newly available data into the house price prediction model. Most of these new data helps to expand the understanding the geographical characteristics towards housing valuation. The nature of these newly available data is often beyond the angle of traditional economic variables, therefore, new methods also introduced into the real estate valuation field. For example, satellite image [12], street-view image, house image from real estate websites [96, 116, 158] are used for house price prediction [33, 82, 87]. Other paper investigated how the satellite image and street view can help to understand the neighbourhood [106], demography [56] and commercial activities. These results are also highly relevant to housing appraisal, though they didn't investigate this question directly.

Besides image data, other types of data are used for housing appraisal beyond the traditional scope of economic modeling. Open Street Map data was used for collecting Point of Interests in the neighbourhood [33, 52]. Mobile phone data

was used to empirically study the human activity and urban vitality [34]. Taxi-cab trajectory data was mined to estimate neighbourhood popularity, in order to understand the geographic factors for housing appraisal [52]. Google search index was used for housing prediction model to understand how people's attention of real estate would influence the future housing price [153]. Text data from real estate related news was studied to learn how sentiment is related to housing price [84,130].

These data applications are innovative and inspiring for improving housing price modeling. However, most of the newly applied spatial data are focused on the discovery of neighbourhood characteristics, such as crime perception [18], walk-ability [30], cultural influence [65]. These applications neglect the understanding beyond the neighbourhood. And this is the main focus of our work.

3.4 Conceptual Framework

In this section, we introduce the conceptual framework of our housing price estimation model.

3.4.1 Overview of the HNED model

As stated in the introduction, the purpose of our model is housing price appraisal. In this work, we introduce a supervised machine learning model with which we experimented housing appraisal. This model may be considered to be an extension of traditional hedonic economic housing appraisal models – an attempt at bringing advances in data science and machine learning into the housing appraisal domain. Accordingly, the target variable of the model is the price of the house. As described in later subsections, we formulated the target variable in two different ways. In the first formulation, the target variable is treated as a numerical dollar value, and

the model attempts to predict the target variable value from the features. This is analogous to linear regression in traditional econometrics, where the regression attempts to predict the independent/outcome variable from the dependent/input variables. In the second formulation, the target variable is treated as a value range, and the research problem becomes a classification problem – the model uses the features to predict into which price range the appraised housing price falls. This is analogous to logistic regression in traditional econometrics, which also attempts to perform classification of the independent/outcome variable from the dependent/input variables. In the rest of the chapter, these two formulations are called (out of convenience) the *housing appraisal regression task* and the *housing appraisal classification task*.

Next, we describe the feature space of the HNED model. The combined feature space of the HNED model is large for this research domain – almost 50 dimensions – and comprises data sets from numerous data suppliers (e.g. the AURIN APM data set), as well as those collated manually from official sources on the web (e.g. the universities' revenues data set for Melbourne for Feature Vector 3, which were collated from the universities' annual reports downloaded from their respective websites). One of the main contributions of the model is exploring the problem of housing appraisal by using a feature space that is novel, using features/data previously not used for housing appraisal. The features of the HNED model are logically grouped into four feature vectors, corresponding to four different types of attributes, which together influence the price of the house. The types of attributes are: housing attributes, neighbourhood characteristics, regional economic clusters, and demographic characteristics. A schematic of the model architecture is shown in Figure 3.1.

We hypothesise that all four feature vectors affect the price of houses, and the

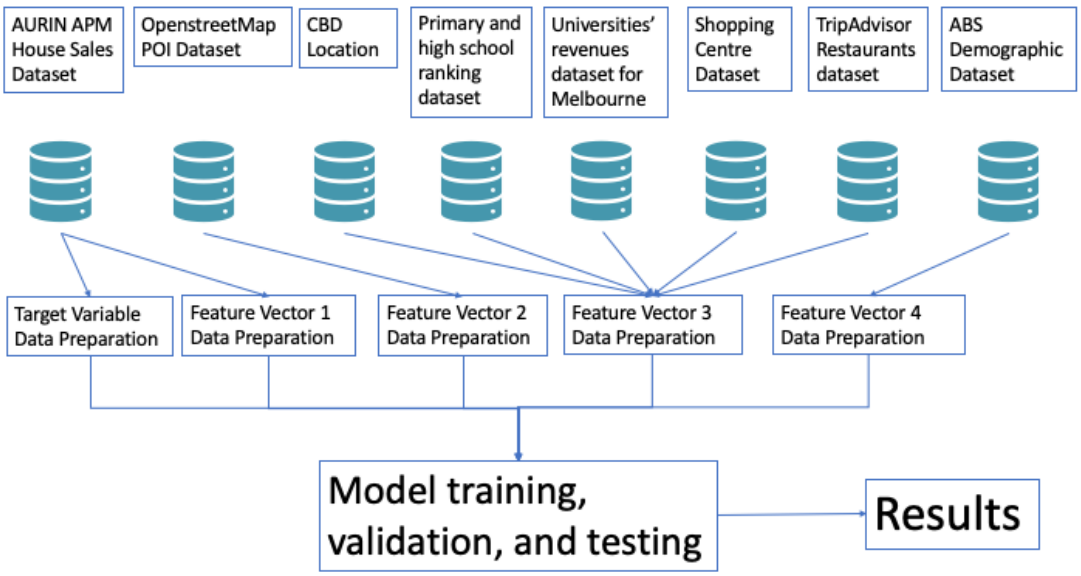


Figure 3.1: HNED Supervised Machine Learning Pipeline

purpose of this study is to investigate (a) How well a supervised machine learning model/pipeline can appraise housing price, based on the four feature vectors of the model and (b) The relative importance of each feature vector in the model's ability to predict the target variable. In the following subsections, we describe in detail the definition of each feature vector.

3.4.2 Feature Vector 1: Housing Attributes

There are basic attributes about the house itself, which people can easily acquire information through advertisement or inspection. These attributes are primary functions that fulfil people's needs in dwellings. The first feature vector corresponds to the influence of these housing attributes, and follows the conceptual framework of the traditional hedonic model for housing appraisal. For our model, 18 property attributes are selected (corresponding to an 18 dimensional feature vector), for which data is obtained through the AURIN APM data set described in Section 3.5.

Below is the full list of the features in Feature Vector 1:

- Property type (unit, house, townhouse)
- Area size (total square metres)
- Number of bedrooms
- Number of bathrooms
- Number of parking spaces
- Separate study? (Yes/No)
- Separate dinning? (Yes/No)
- Separate family room? (Yes/No)
- Rumpus room? (Yes/No)
- Fireplace? (Yes/No)
- Walk-in-wardrobe? (Yes/No)
- Air conditioning? (Yes/No)
- Balcony? (Yes/No)
- En suite? (Yes/No)
- Garage? (Yes/No)
- Lockup garage? (Yes/No)
- Polished timber floor? (Yes/No)
- Barbeque? (Yes/No)

3.4.3 Feature Vectors 2 and 3: The Housing Location

Before describing feature vector 2 and feature vector 3 in detail, we would like to discuss first some common background considerations that motivated the definition of both feature vectors. Feature vectors 2 and 3 both relate to the location of the house. In the context of this paper, we interpret location as the accessibility to certain local amenities from the house, as well as the accessibility to important economic clusters from the house, such as the CBD, large shopping centres, universities, etc. Our model captures the quality, quantity and accessibility of these amenities.

What is location?

There is a cliché used by real estate agents, that the three most important factors when buying a home are location, location, and location. Intuitively, feature vectors 2 and 3 aim to capture the effects of this cliché. Interestingly, there is limited literature within urban economics and computational urban informatics to discuss the relationship between location and housing value, especially how location influences the housing value.

We define location as the position and relation of an individual property associated with its relevant urban context. This context includes both the immediate social network around the property, and relative association with remote strategically and economically significant clusters. In the HNED model, we conceptually use both aspects of urban context as features. Feature Vector 2 is used to represent the immediate social network around the property. Feature Vector 3 is used to represent the property's relative association with remote strategically and economically significant clusters.

There exists rich literature with the investigation of relationship between hous-

ing value and its neighbourhood [30, 33, 34, 51, 52]. Differing from the current research, we emphasise the value of location rather than neighbourhood. Neighbourhood is the direct geographical and social influence around the individual property. The existing literature neglects the investigation of the relationship between individual property and the regional economic clusters within the metropolitan area. Urban sociologist Burgess [19] emphasised that the urban growth radially expanses from its CBD and physically attractive neighbourhoods. This provides us the theoretical guidance to investigate how location and social networks influence housing prices.

We assume that people choose a particular location, because they believe such location provides the convenience for them to build their social network, for activities such as providing or consuming local goods and services, outdoor activities, family and friend gathering, etc. In other words, we assume people are willing to pay higher house prices for the benefit of close to high quality educations (public and private schools), nice shops and restaurants, convenient public transport options, parks, work places, etc. More importantly, we also assume that people choose to bear the cost of travelling beyond their local neighbourhood when the remote connections outperform the nearby connections or remote connection is the only suited option for a particular activity to happen. Practically, people have to travel for work, education, shopping, recreation or gathering purpose. With the budget limit, people need to make a trade-off among functions of house, quality of nearby neighbourhood and distance to regional economic clusters.

Quantifying location

In order to use location-related feature vectors in the HNED model, we need to be able to express the concepts described by these feature vectors in quantitative terms. As discussed in the previous subsubsection, we have the following concep-

tual definition for feature vectors 2 and 3. Feature Vector 2 is used to represent the immediate social network around the property. Feature Vector 3 is used to represent the property's relative association with remote strategically and economically significant clusters.

The intuition behind this part of the HNED model is that places of interest (POIs), whether they are small like a bus stop, or large like a shopping centre, exert a degree of influence upon the housing price of a nearby property. Thus, we are quantifying is this degree of influence, to be expressible in terms of features for which we have data sets. We first discuss this in the general, with respect to a variable for which we have the data for many kinds of POIs: distance between the POI and the property.

In urban research and geography, there is a famous statement known as Tobler's law, which was cited earlier in this thesis: *"Everything is usually related to all else, but those which are near to each other are more related when compared to those that are further away"*. We attempt here to construct a quantitative version of Tobler's law, which will be used to describe how the influence of a POI upon a property's price is affected by the distance between the POI and the property.

The inspiration of our quantitative statement comes from Astronomy. In Astronomy, Newton's law of universal gravitation is given by the formula

$$\vec{F} = -\frac{GMm}{r^2} \tag{3.1}$$

In this famous equation, the gravitational force diminishes as the inverse square of the distance between the two masses ("inverse square law"). This can be attributed to a geometric fact – that space is three dimensional, and is more clearly expressed by the Gauss' law form of gravitation:

$$\Phi_G = \oiint_{\partial V} \vec{g} \cdot d\vec{A} = \oiint_{\partial V} \frac{\vec{F}}{m} \cdot d\vec{A} = -4\pi GM \tag{3.2}$$

which says that the total gravitational flux around mass M is proportional to M. Since in simple cases, the gravitational field extends radially from the mass M, the flux integral evaluates to be simply the field strength times the surface area of the sphere ∂V of radius r:

$$\Phi_G = \oiint_{\partial V} \frac{\vec{F}}{m} \cdot d\vec{A} = \frac{\vec{F}}{m}(4\pi r^2) \tag{3.3}$$

Therefore in 3-dimensions, we recover Newton's law of gravitation from Gauss' law by combining Equations 3.2 and 3.3:

$$\oiint_{\partial V} \frac{\vec{F}}{m} \cdot d\vec{A} = \frac{\vec{F}}{m}(4\pi r^2) = -4\pi GM \tag{3.4}$$

$$\frac{\vec{F}}{m} = -\frac{GM}{r^2} \tag{3.5}$$

$$\vec{F} = -\frac{GMm}{r^2} \tag{3.6}$$

In our case of a quantitative version of Tobler's law, instead of being in 3-dimensional space, the POIs reside on the surface of the earth which, for the purposes of our study, is contained within the boundary of a disk of approximately 30km (the radius of Greater Melbourne), which we can approximate as a flat 2-dimensional disk. Thus, we propose to adapt the 2-dimensional version of Newton's law of gravity as the quantitative Tobler's law. We adopt the mathematical relationship between the 2-dimensional version of the gravitational force and the distance between the two masses, as the mathematical relationship between the influence of a POI on a house's price, and the distance between the POI and the property (assuming that all other variables/features are held constant, or in economics terminology, *ceteris paribus*). Whilst clearly the influence of a POI on the price of a house is not the same as the gravitational force, we consider the two-dimensional version of Newton's law of gravity to be an excellent starting point

and approximation to the true relationship between the influence of a POI on a property's price, and the distance between the POI and the property.

The derivation of the 2-dimensional form of Newton's law of gravity is as follows. We begin with the 2-dimensional form of Gauss' law of gravity:

$$\Phi_G = \oint_{\partial A} \vec{g} \cdot d\vec{l} = \oint_{\partial A} \frac{\vec{F}}{m} \cdot d\vec{l} = -4\pi GM \tag{3.7}$$

At this point, we only need the two rightmost expressions of the Equation 3.7:

$$\oint_{\partial A} \frac{\vec{F}}{m} \cdot d\vec{l} = -4\pi GM \tag{3.8}$$

In simple cases, the gravitational field extends radially from the larger mass M, so in 2-dimensions, the flux integral evaluates to be the field strength times the perimeter distance of the disk of radius r:

$$\oint_{\partial A} \frac{\vec{F}}{m} \cdot d\vec{l} = \frac{\vec{F}}{m}(2\pi r) \tag{3.9}$$

Hence

$$\frac{\vec{F}}{m}(2\pi r) = -4\pi GM \tag{3.10}$$

$$\vec{F} = -\frac{2GMm}{r} \tag{3.11}$$

Equation 3.10 shows that in 2-dimensions, the gravitational force diminishes as the inverse (and not the inverse squared) of the distance between the two masses. Therefore, for the POI features in both feature vectors 2 and 3, the mathematical relationship between the influence of a POI on a house's price, and the distance between the POI and the property, is given as the same inverse law relationship (assuming that all other variables/features are held constant, or in economics terminology, *ceteris paribus*). This mathematical relationship will be used in the next sections, where we discuss how the features of feature vector 2 and feature vector 3 are defined in detail.

3.4.4 Feature Vector 2: POI-based Neighbourhood Characteristics

As defined in the previous subsection, feature vector 2 is used to represent the immediate social network around the property. It focuses on the small points of interest (POIs) within walking distance (1 kilometre), such as shops, bus stops, parks, places of worship, etc. For convenience, we refer to these POIs as *local POIs*. Conceptually, feature vector 2 forms a substantial part of the property's location based social network (LBSN). In order not to duplicate features, we exclude any POIs that form a part of feature vector 3, such as schools, shopping centres, and universities.

Considering both conceptually what feature vector 2 ought to capture, as well as what substantial data sets we are practically able to obtain, we used 13 categories of local POIs from OpenStreetMap, where the data set comes from. We did not model this as a 13-dimensional feature vector, but rather transformed the data into a single value (1-dimensional vector) as follows.

In the previous subsection, we explained our gravitation inspired treatment of a quantitative version of Tobler's law, whereby for the POI features in both feature vectors 2 and 3, the mathematical relationship between the influence of a POI on a house's price, and the distance between the POI and the property, is given as the inverse law relationship (assuming that all other variables/features are held constant, or in economics terminology, *ceteris paribus*).

We have not thus far discussed how the influence of a POI on a house's price is related to the other attributes of the POI. For the local POIs described by feature vector 2, as these are smaller POIs than those described by feature vector 3, whose influence would be short-ranged, we decided to construct the model to weigh the other aspects of the local POIs equally. That is, in the HNED model:

1. The influence of each local POIs on a house's price is dependent only on the inverse law of physical distance between the POIs and the house
2. The combined influence of all local POIs on the house's price is the sum of the inverse physical distances between the POIs and the house (this is analogous to the case of gravitational force, with the only difference being that gravitational force is the vector sum of all of the forces from masses exerting gravity, whereas for housing price, the influence is a scalar sum)
3. As discussed earlier, feature vector 2 only uses local POIs within 1km of the house

As a worked example, if a house is 0.2km from a bus stop, 0.4km from a tennis court, and 0.5km from a place of worship, and these are the only local POIs within 1km of the house, the total value of the local POI influence (i.e. feature vector 2) for this house given by:

$$\frac{1}{0.2} + \frac{1}{0.4} + \frac{1}{0.5} = 9.5 \tag{3.12}$$

Below is the full list of the 13 categories that are combined to create feature vector 2:

- Transport
- Food
- Shop
- Accommodation
- Health
- Land use
- Amenity

- Place of worship
- Sport
- Tourist
- Money
- Barrier
- Water

3.4.5 Feature Vector 3: Regional-level Economic Clusters

One of the major novelties of our work is to bring in regional-level economic clusters as features in the HNED model, for the purpose of housing appraisal. These regional-level economic clusters capture economic activities not happen at neighbourhood level, but are nevertheless influential to the price of the house. As discussed in the introduction, capturing all the features in the regional or metropolitan level is time consuming and not realistic. The challenge is to find the most important link between a location and the house.

In the preliminary considerations for the HNED model, we reasoned that as much daily exigencies of life surround commonplace economic activities such as buying food, clothes, schooling, banking, post, etc., one may conclude that economic/value exchange is a very important link between a house and its location. Therefore, we try to capture the economic clusters that provide huge economic values.

In the HNED model, we represent regional-level economic clusters as feature vector 3 in the following way. We used the following clusters in the model: CBD, Primary Schools, High Schools, Universities, Shopping Centres, and Restaurants. In order to capture their respective influences on the housing price, we used both

their physical distances from the house (in the same way as for feature vector 2; the inverse law), as well as other factors depending on the type of the regional-level economic clusters. Due to the different degrees that the regional-level economic clusters may exert influence on housing price, and that these degrees could vary a lot (e.g. the CBD's influence would be considerably higher than a High School, even if they are the same distance from the house), we did not attempt to combine the influences into a single dimensional vector as for feature vector 2. Instead, feature vector 3 is basically a Cartesian product its components, each of which correspond to a type of regional-level economic cluster.

For the CBD, as we could not find a good proxy for how we should weigh the influence of the CBD in addition to inverse distance, we ended up just using the distance from the CBD for this feature vector component. We used the GPS coordinates of Melbourne Central Railway Station, which is approximately in the middle of the Melbourne CBD (which we define to be the region from Parkville to Southbank), as the point to measure the physical distance between the CBD and the house (the "centre of mass" of the CBD).

For Primary Schools, we used a 2-dimensional feature vector component, comprising the inverse distance and the ranking of the school (see Section 3.5 for detailed descriptions of the data).

For High Schools, in the same way as for Primary Schools, we used a 2-dimensional feature vector component, comprising the inverse distance and the ranking of the school (see Section 3.5 for detailed descriptions of the data).

For Universities, we used the universities' revenue as a factor of influence in addition to distance between the universities and the house. So the feature vector component for Universities is 1-dimensional, and defined as $(\log_{10} r)/d$, where r is the revenue of the university in Australian Dollars, and d is the distance between

the university and the house in kilometres. The reason why a logarithm is taken on the revenue is because we expect that organisational revenue increase exponential with respect to size, so the logarithm of revenue would provide a more accurate indication of the university's size and influence on housing price.

For Shopping Centres, restricted by the availability of data, we ended up using the number of shops of the shopping centre as a proxy as the "gravitational mass" of the shopping centre in exerting influence on housing price. The feature vector component for Shopping Centres,is 1-dimensional, and defined as $(\log_{10} s)/d$, where s is the number of shops in the shopping centre, and d is the distance between the Shopping Centre and the house in kilometres. The reason why a logarithm is taken on number of shops in the shopping centre is the same reason as for the universities' revenue. We expect that a shopping centre's number of shops to increase exponential with respect to size, so the logarithm of revenue would provide a more accurate indication of the shopping centre's size and influence on housing price.

For restaurants, we followed the same reasoning as for Primary and High Schools. We used a 2-dimensional feature vector component, comprising the inverse distance and the ranking of the restaurant (see Section 3.5 for detailed descriptions of the data).

In summary, all of these 6 different types of clusters appear as feature components of feature vector 3. Thus, feature vector 3 has a total of 9 dimensions; being the Cartesian Product of the 6 feature vector components described above. For clarify, we reiterate that the components for CBD, Universities, and Shopping Centres are 1-dimension vectors, and the components for Primary Schools, High Schools, and Restaurants are 2-dimensional vectors. The feature vector components bring into the HNED model the influence on the housing price from the

different regional-level economic clusters, and are formulated to be attenuated by the inverse of the physical distance between the cluster and the house, as well as weighted by some other attribute of the cluster such as ranking or revenue (except for the CBD, whose feature has only the distance relationship). Finally, we emphasise that unlike with feature vector 2, the regional-level economic clusters do not have to be within 1km of the house to exert influence on the housing price. As we shall describe in Section 3.5, every single regional-level economic cluster (e.g. every primary school and high school) is used in the model for feature vector 3, and not just those within 1km of the house.

Below is the full list of features in feature vector 3:

- CBD (used the Melbourne Central Railway Station GPS coordinates as proxy "centre of mass" for the whole CBD)
- Primary Schools (Distance and ranking)
- High Schools (Distance and ranking)
- Universities (weighted by $\log_{10}$ of revenue ($))
- Shopping Centres (weighted by $\log_{10}$ of number of shops)
- Restaurants (Distance and ranking)

3.4.6 Feature Vector 4: Socio-demographic Attributes

Feature Vector 4 involves the socio-demographic attributes of the suburb to which the house belongs, which characterise the social community that is physically closest to the house.

Previous urban economic research has explored the relationships between housing attributes and demographic characteristics of population and found that socio-demographic profiles determines the demand segregation and forms different trends

city-wide [142]. We include social-demographic profiles into our modelling for a few reasons. First, we consider social-demographic characteristics can create long-term effect to shape the local economy and community. Housing is not easily transferrable as investment in stock market because of its physical moving difficulty and potential extra capital gain tax for short-term penalty. Residents would stay in the same suburb for a long period and co-create the taste, economy and culture of its local community. The suburb would grow with its residents. Second, human capital can generate externalities and have spill over effect in the neighbourhood regions [46]. Related businesses are more likely to be adjacent and form cluster effect.

Feature vector 4 includes this aspect by adding into the model the relationship between property price and its social-demographic profile. The features/components of this feature vector are carefully selected from four sections of the 2016 Australian census data. This include features about income, people and population, education and employment, family and community. We use these features to understand people and their life in each suburb. Below is the full list of features in feature vector 4:

- Median Age - Persons (years)
- Working Age Population (aged 15-64 years) (%)
- Persons - Total (no.)
- Population density (persons/km2)
- Speaks a Language Other Than English at Home (%)
- Median employee income ($)
- Mean employee income ($)
- Median investment income ($)

- Mean investment income ($)
- Total income earners (excl. Government pensions and allowances) (no.)
- Total income earners (excl. Government pensions and allowances) - median age (years)
- Median equivalised total household income (weekly) ($)
- Completed Year 12 or equivalent (%)
- Bachelor Degree (%)
- Unemployment rate (%)
- Average household size (no. of persons)
- Households with mortgage repayments greater than or equal to 30% of household income (%)
- Households with rent payments greater than or equal to 30% of household income (%)
- Homeless rate per 10,000 persons (rate)
- Median commuting distance to place of work (kms)

3.5 Methodology and Experimental Settings

This section describes the methodology and experimental implementation of the HNED model described in Section 3.4.

3.5.1 Data Description

Target Variable

We use metropolitan Melbourne, Australia as our experiment city. We obtained a comprehensive property rental and sales data set from the Australian Urban Re-

search Infrastructure Network (AURIN) – the AURIN APM data set, comprising all property sales and rental transactions in Melbourne from 2014 to 2019. For this project, we mainly focus on the sold housing price data in 2018, which includes a total of $161,179$ recorded sold properties.

As the goal of this research project is housing appraisal, the target variable is the housing price. The data for the target variable is contained within the AURIN APM data set. As a part of the pre-processing, we removed all properties whose sales price were less than $10,000$ (as the price is too low to be "normal" house sales). We also removed properties whose geographical locations were missing, as they are unusable for the HNED pipeline, whose feature vectors as described in Section 3.4 require the distances between the features and the house as inputs. The remaining data set contains $158,588$ properties. A sample screenshot of the data set can be found in Figure 3.2. Each row of the data set corresponds to one property sale in Melbourne in 2018.

Figure 3.2 shows some key columns that are related to the target variable in the pre-processed data set. The column *eventprice* (the 10th column counted from the left in Figure 3.2) contains the sales prices. Columns *property_latitude* and *property_longitude* (the 11th and 12th columns counted from the left in Figure 3.2) contain latitude and longitudes data of the property sales.

Feature Vector 1

The housing attributes for each property sale also come from the also come from the AURIN APM data set. Figure 3.3 shows some key columns of the data set that are related to Feature Vector 1. Columns from *property_type* to *has_familyroom* (from the 7th column to the 13th column, counted from the left) contain housing attribute features described in Section 3.4.2.

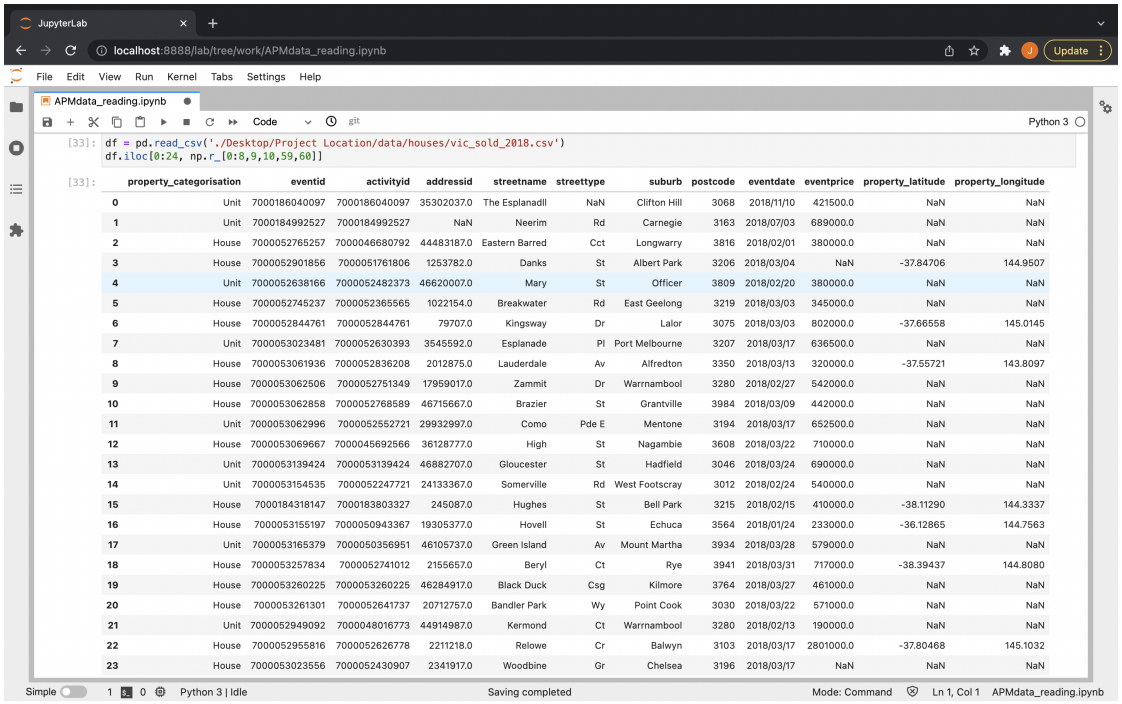


Figure 3.2: Sample housing sales price data from AURIN APM data set (Melbourne 2018)

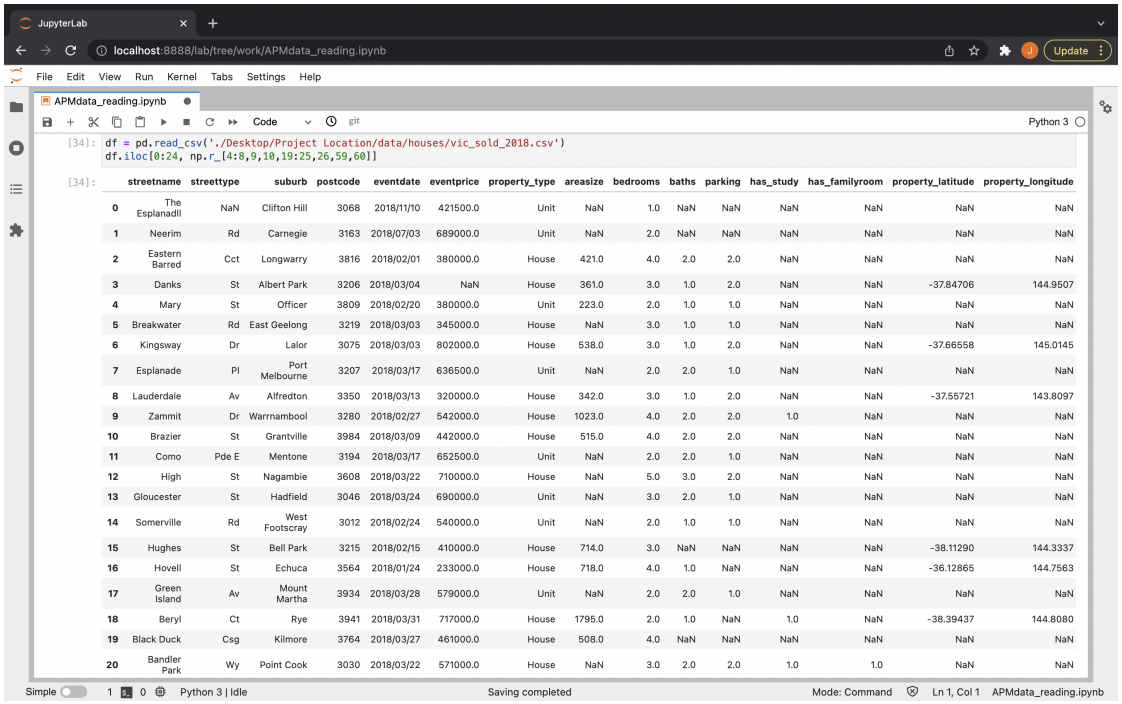


Figure 3.3: Feature Vector 1 housing attributes sample data set, from the AURIN APM housing sales data set (Melbourne 2018)

Feature Vector 2

We collect local POIs data from OpenStreetMap, for the 13 types of POIs as described in Section 3.4.4. Figure 3.4 provide a sample of the collected feature vector 2 data. The key columns/attributes of the local POI data, for the purposes of input into the HNED model, are their POI unique identifiers (Column 2) and latitudes/longitudes (Columns 3/4), which enable the local POIs to be included into the model according to the framework described in Section 3.4.4.

	poi_type_id	poi_id	lat	lng	poi_name	poi_type_name	dist
55466	165	N2229432386	-37.782249	144.891206	Stop 55: Lyric Street	TRANSPORT_TRAMSTOP	0.001886
55471	165	N2229436304	-37.782084	144.891274	Stop 55: Lyric Street	TRANSPORT_TRAMSTOP	0.001927
118834	161	N4427569768	-37.781619	144.891349	Lyric Street	TRANSPORT_BUSSTOP	0.002000
118835	161	N4427569771	-37.781446	144.891383	Macedon Street	TRANSPORT_BUSSTOP	0.002062
118846	165	N4427569789	-37.779944	144.890877	Stop 54: Gordon Street	TRANSPORT_TRAMSTOP	0.002439
118847	165	N4427569790	-37.779949	144.891161	Stop 54: Gordon Street	TRANSPORT_TRAMSTOP	0.002621
118845	165	N4427569788	-37.779508	144.888027	Stop 53: Maribyrnong Secondary College	TRANSPORT_TRAMSTOP	0.002658
118844	165	N4427569787	-37.779561	144.887737	Stop 53: Maribyrnong Secondary College	TRANSPORT_TRAMSTOP	0.002820
118836	161	N4427569772	-37.784299	144.890822	Monash Street	TRANSPORT_BUSSTOP	0.002846
55469	165	N2229435815	-37.784391	144.890823	Stop 56: Edgewater Boulevard	TRANSPORT_TRAMSTOP	0.002926
55470	165	N2229435816	-37.784639	144.890698	Stop 56: Edgewater Boulevard	TRANSPORT_TRAMSTOP	0.003268
118821	161	N4427569743	-37.785279	144.890632	Birdwood Street	TRANSPORT_BUSSTOP	0.003651
148540	105	W73346754	-37.784284	144.892316	ALDI	SHOP_SUPERMARKET	0.003823
145083	21	W48045999	-37.778009	144.889093	Maribyrnong College	EDUCATION_SCHOOL	0.003856
71349	77	N2493373385	-37.784686	144.892171	The Edgewater Meats	SHOP_BUTCHER	0.003867
71348	25	N2493373116	-37.785133	144.891830	Desserts by Night	FOOD_CAFE	0.004102
148541	40	W73346526	-37.786099	144.888838	Harmony Park	LANDUSE_GRASS	0.004276
144212	40	W46942708	-37.778062	144.891382	Thompson Reserve	LANDUSE_GRASS	0.004298
196012	40	W491535008	-37.778367	144.886767	Thomas Barrett Reserve	LANDUSE_GRASS	0.004337
118843	165	N4427569786	-37.779298	144.885712	Stop 52: Rosamond Road	TRANSPORT_TRAMSTOP	0.004458

Figure 3.4: Feature Vector 2 sample data that shows different types of local-level POIs

Feature Vector 3

For the 6 types of regional-level economic clusters described in Section 3.4.5, we obtained all required data for Melbourne for all of the components of feature vector 3. Specifically:

1. The latitude/longitude of Melbourne Central Railway Station
2. All primary school latitudes and longitudes, as well as their ranking based on their standardised exam results
3. All high school latitudes and longitudes, as well as their ranking based on their standardised exam results. Figure 3.5 provides a sample of this data set
4. The 8 major universities based in Melbourne with their latitudes and longitudes, as well as their revenues. If a university has multiple major campuses, each major campus is treated as a single university, and the revenue is split across the major campuses. Figure 3.6 provides a sample of this data set
5. 42 major shopping centres in Melbourne, their latitudes and longitudes, and the total number of shops in each shoppoing centre. Figure 3.7 provides a sample of this data set
6. We collected from TripAdvisor restaurant latitudes and longitudes, as well as their ranking on TripAdvisor. Figure 3.8 provides a sample of this data set

Feature Vector 4

For the socio-demographic data, we use Australian Bureau of Statistics census data 2016. A screenshot of a sample of the data set for feature vector 4 is shown in Figure 3.9, for the Melbourne suburb of Brunswick.

3.5.2 Experimental Settings and Algorithms

We showed in Figure 3.1 the high-level logic of the HNED model pipeline. As the model itself is novel and the proposed study is new in the Australian housing

	A	B	C	D	E	F	G	H	I
1	rank	name	Locality	geojson	auto_lat	auto_lng	lat	lng	
2	1	Bialik College	HAWTHORN EAST	{'type': 'Feat	-37.84151077	145.0415192	-37.84151077	145.0415192	
3	2	Mac.Robertson Girls' High Schl	MELBOURNE	{'type': 'Feat	-37.81739044	144.967514	-37.81739044	144.967514	
4	3	Yeshivah College	ST KILDA EAST	{'type': 'Feat	-37.894445	145.00834	-37.894445	145.00834	
5	4	Ballarat Clarendon College	BALLARAT	{'type': 'Feat	-37.55889893	143.8509827	-37.55889893	143.8509827	
6	5	Melbourne High School	SOUTH YARRA	{'type': 'Feat	-37.83649063	144.9962311	-37.83649063	144.9962311	
7	6	Mount Scopus Memorial College	BURWOOD	{'type': 'Feat	-33.87768936	151.105484	-37.8486843	145.1201707	
8	7	Ruyton Girls' School	KEW	{'type': 'Feat	-37.81233978	145.0385284	-37.81233978	145.0385284	
9	8	Shelford Girls' Grammar	CAULFIELD	{'type': 'Feat	-37.87995148	145.0226135	-37.87995148	145.0226135	
10	9	Beth Rivkah Ladies College	ST KILDA EAST	{'type': 'Feat	-37.894445	145.00834	-37.894445	145.00834	
11	10	Haileybury Girls College	KEYSBOROUGH	{'type': 'Feat	-37.99234009	145.1754456	-37.99234009	145.1754456	
12	11	Huntingtower School	MOUNT WAVERLEY	{'type': 'Feat	-37.87648	145.14029	-37.87648	145.14029	
13	12	Korowa Anglican Girls' School	GLEN IRIS	{'type': 'Feat	-37.86104965	145.0553131	-37.86104965	145.0553131	
14	13	Lauriston Girls School	ARMADALE	{'type': 'Feat	-37.8508606	145.0241547	-37.8508606	145.0241547	
15	14	Leibler Yavneh College	ELSTERNWICK	{'type': 'Feat	31.87029457	34.74778748	-37.8928496	145.006882	
16	15	Loreto Mandeville Hall	TOORAK	{'type': 'Feat	-37.84865952	145.0143433	-37.84865952	145.0143433	
17	16	Melbourne Girls Grammar	SOUTH YARRA	{'type': 'Feat	-37.83140183	144.9842834	-37.83140182	144.9842834	
18	17	Penleigh & Essendon Grammar	KEILOR EAST	{'type': 'Feat	-37.73173904	144.8717041	-37.73173904	144.8717041	
19	18	Sacre Coeur	GLEN IRIS	{'type': 'Feat	-37.86206055	145.0508423	-37.86206055	145.0508423	
20	19	Strathcona Baptist Girls GS	CANTERBURY	{'type': 'Feat	51.2759819	1.075600028	-37.8299791	145.0801375	
21	20	Trinity Grammar School	KEW	{'type': 'Feat	-37.81050873	145.0348358	-37.81050873	145.0348358	
22	21	Camberwell Anglican Girls GS	CANTERBURY	{'type': 'Feat	-37.8362999	145.0625153	-37.8362999	145.0625153	
23	22	Camberwell Grammar School	CANTERBURY	{'type': 'Feat	-37.81687927	145.0683289	-37.81687927	145.0683289	
24	23	Caulfield Grammar School	WHEELERS HILL	{'type': 'Feat	-37.90559006	145.1893158	-37.90559006	145.1893158	
25	24	Caulfield Grammar School	ST KILDA EAST	{'type': 'Feat	-37.87757111	145.0036469	-37.87757111	145.0036469	
26	25	Fintona Girls School	BALWYN	{'type': 'Feat	-37.81481934	145.0809326	-37.81481934	145.0809326	
27	26	Firbank Grammar School	BRIGHTON	{'type': 'Feat	-37.9057312	144.9964294	-37.9057312	144.9964294	
28	27	Genazzano F.C.J. College	KEW	{'type': 'Feat	-37.80981064	145.0562744	-37.80981064	145.0562744	
29	28	Goulburn Valley Grammar Schl	SHEPPARTON	{'type': 'Feat	-36.32749176	145.415741	-36.32749176	145.415741	
30	29	Haileybury College	KEYSBOROUGH	{'type': 'Feat	-37.99156	145.16554	-37.99156	145.16554	
31	30	Kilvington Grammar School	ORMOND	{'type': 'Feat	-37.89926147	145.0420837	-37.89926147	145.0420837	
32	31	Lowther Hall Anglican GS	ESSENDON	{'type': 'Feat	-37.75152969	144.9095154	-37.75152969	144.9095154	
33	32	Mentone Girls' Grammar School	MENTONE	{'type': 'Feat	-37.98936081	145.0647125	-37.98936081	145.0647125	
34	33	Mortlake College	MORTLAKE	{'type': 'Feat	-33.83742	151.10572	-38.0864938	141.6976787	
35	34	Nossal High School	BERWICK	{'type': 'Feat	-38.03900146	145.3367157	-38.03900146	145.3367157	
36	35	Presbyterian Ladies' College	BURWOOD	{'type': 'Feat	-37.84972	145.1062775	-37.84972	145.1062775	
37	36	Scotch College	HAWTHORN	{'type': 'Feat	-37.83401871	145.0294189	-37.83401871	145.0294189	
38	37	St Catherine's School	TOORAK	{'type': 'Feat	-37.83778	145.0219574	-37.83778	145.0219574	
39	38	St Kevin's College	TOORAK	{'type': 'Feat	-37.83889008	145.0277252	-37.83889008	145.0277252	
40	39	The King David School	ARMADALE	{'type': 'Feat	55.89889145	-3.695770025	-37.8581622	145.0106846	
41	40	Woodleigh School	LANGWARRIN SOUTH	{'type': 'Feat	-38.18259048	145.1849976	-38.18259048	145.1849976	
42	41	Yesodei HaTorah College	BRIGHTON	{'type': 'Feat	50.82806015	-0.136790007	-37.873084	144.9833242	
43	42	Balwyn High School	BALWYN NORTH	{'type': 'Feat	-37.79888916	145.0768433	-37.79888916	145.0768433	
44	43	Brighton Grammar School	BRIGHTON	{'type': 'Feat	-37.90832901	144.9936829	-37.90832901	144.9936829	
45	44	Carey Baptist Grammar School	KEW	{'type': 'Feat	-37.81597137	145.0481262	-37.81597137	145.0481262	
46	45	Girton Grammar School	BENDIGO	{'type': 'Feat	-36.76147079	144.2706146	-36.76147079	144.2706146	

high_schools

Figure 3.5: Feature Vector 3 sample data: High Schools location and ranking

appraisal setting, we approached the experimental design by comparing the performance of several well-known supervised machine learning models, when applied to the problem described in this chapter. These models include linear regression, logistic regression, k-nearest neighbour, and XGBoost, which we shall soon describe in detail.

In the practical implementation of the pipeline for our experiment, we set up the pipeline in two difference forms. Although the target variable is housing price, this can be interpreted in two different forms. The first form is appraising the exact

	A	B	C	D	E	F
1	name	campus	lat	lng	revenue	
2	The University of Melbourne	Parkville	-37.7963	144.9614	2927232	
3	Monash University	Clayton	-37.909604	145.1364	1787951.4	
4	Monash University	Caulfield	-37.876891	145.044834	893975.7	
5	Monash University	Peninsula	-38.151779	145.13636	297991.9	
6	Latrobe University	Bundoora	-37.718962	145.048272	867426	
7	RMIT University	Melbourne	-37.80234	144.963371	1063458.9	
8	RMIT University	Bundoora	-37.678729	145.069381	455768.1	
9	Victoria University	Footscray	-37.799778	144.899361	192336.8	
10	Victoria University	Melbourne	-37.81802	144.963997	192336.8	
11	Victoria University	St Albans	-37.750345	144.79798	96168.4	
12	Swinburne University	Hawthorne	-37.821476	145.039	783659	
13	Deakin University	Burwood	-37.847334	145.114932	676387	
14	Australian Catholic University	Fitzroy	-37.807222	144.977889	164910.9	
15						

Figure 3.6: Feature Vector 3 sample data: Universities location and revenue

dollar value of the housing price, so the target variable is a continuous variable. The second form is estimating the price range of the house, so the target variable is a categorical variable, representing the estimated price range. We decided to investigate both forms of the experiment, and report the results in Section 3.6. As stated in Section 3.4, for convenience we refer to the first form of the experiment as the *Housing Appraisal Regression Task*, and the second form of the experiment as the *Housing Appraisal Classification Task*.

Housing Appraisal Regression Task

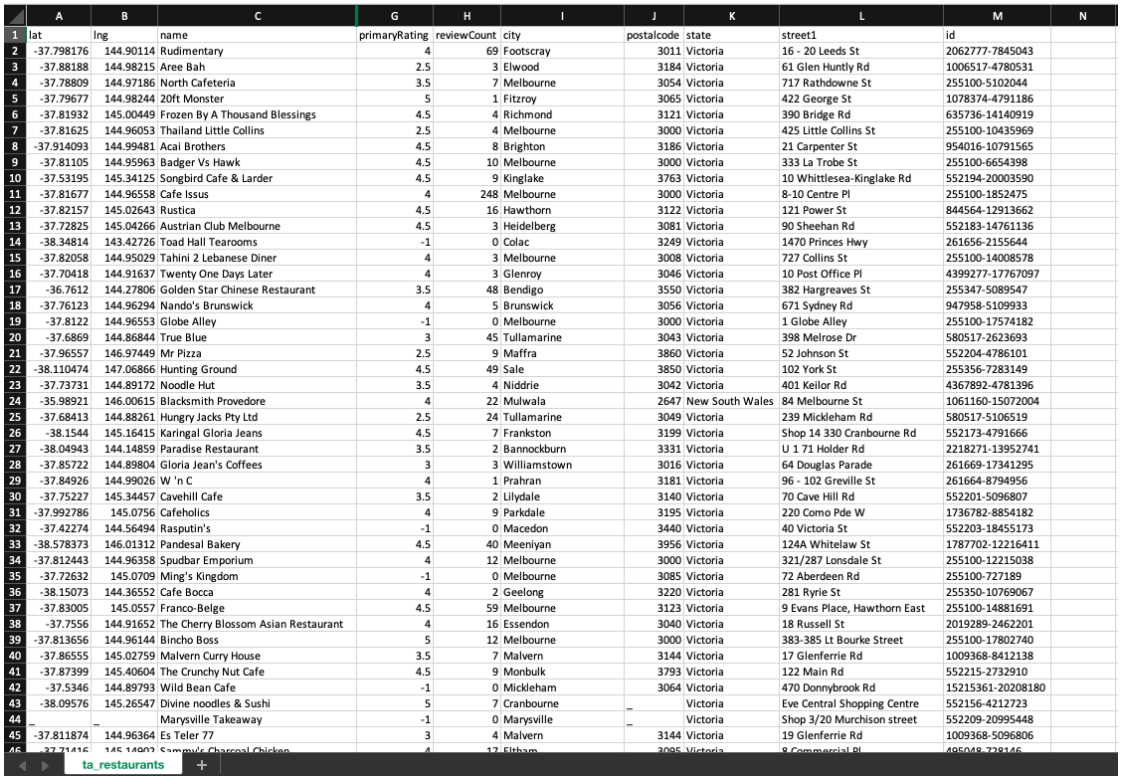
For the *Housing Appraisal Regression Task*, we use Linear Regression as the baseline algorithm/model, as this is what is commonly used for such problems in Classical Economics. The performance of three other model (Support Vector Machine, Multilayer Perceptron, XGBoost) were compared against the performance of the baseline. A total of 4 models were experimented upon.

For clarity, we shall describe here in detail the set up of the supervised ma-

	A	B	C	D	E
1	name	lat	lng	num_stores	
2	Armada Dandenong Plaza	-37.990833	145.220278	165	
3	Bayside Shopping Centre	-38.141138	145.124648	208	
4	Box Hill Central Shopping Centre	-37.819444	145.123333	187	
5	Broadmeadows Shopping Centre	-37.680278	144.919444	171	
6	Casey Central	-37.876389	145.165	90	
7	Chadstone Shopping Centre	-37.885833	145.0825	487	
8	Chirnside Park Shopping Centre	-37.757222	145.3125	119	
9	Cranbourne Park Shopping Centre	-38.1	145.283333	137	
10	Eastland Shopping Centre	-37.813056	145.229167	385	
11	Forest Hill Chase Shopping Centre	-37.8423	145.1654	166	
12	The Glen Shopping Centre	-37.876389	145.165	247	
13	Highpoint Shopping Centre	-37.773333	144.885833	500	
14	Northcote Shopping Plaza	-37.768889	145.001667	71	
15	Northland Shopping Centre	-37.738333	145.029722	293	
16	Pacific Werribee	-37.875213	144.679659	288	
17	Point Cook Town Centre	-37.883889	144.735833	119	
18	Westfield Doncaster	-37.783333	145.125	393	
19	Westfield Fountain Gate	-38.018611	145.304167	420	
20	Westfield Knox	-37.86888	145.241348	317	
21	Westfield Plenty Valley	-37.650937	145.06897	183	
22	Westfield Southland	-37.958	145.05	347	
23	Emporium	-37.8124	144.9638	180	
24					

Figure 3.7: Feature Vector 3 sample data: Shopping Centre location and total number of shops

chine learning pipeline for this task. The pipeline software was written in Python 3.8. We followed the high-level architecture of Figure 3.1. The data sets used for the training, validation, and testing of the 4 models described in the previous paragraph are from the data sets described in Section 3.5.1. Indeed, these are the target variable data and the data of Feature Vectors 1, 2, 3, 4. The feature vector data sets are mapped into the model by following the mathematical relationships described in Section 3.4 and the practical implementations described in Section 3.5.1. The total combined feature space of the model (i.e. the dimension of the vector space formed by taking the Cartesian Product of all 4 feature vectors)



	A	B	C	G	H	I	J	K	L	M	N
1	lat	lng	name	primaryRating	reviewCount	city	postalcode	state	street1	id	
2	-37.798176	144.90114	Rudimentary	4	69	Footscray	3011	Victoria	16 - 20 Leeds St	2062777-7845043	
3	-37.88188	144.98215	Aree Bah	2.5	3	Elwood	3184	Victoria	61 Glen Huntly Rd	1006517-4780531	
4	-37.78809	144.97186	North Cafeteria	3.5	7	Melbourne	3054	Victoria	717 Rathdowne St	255100-5102044	
5	-37.79677	144.98244	20ft Monster	5	1	Fitzroy	3065	Victoria	422 George St	1078374-4791186	
6	-37.81932	145.00449	Frozen By A Thousand Blessings	4.5	4	Richmond	3121	Victoria	390 Bridge Rd	635736-14140919	
7	-37.81625	144.96053	Thailand Little Collins	2.5	4	Melbourne	3000	Victoria	425 Little Collins St	255100-10435969	
8	-37.914093	144.99481	Acai Brothers	4.5	8	Brighton	3186	Victoria	21 Carpenter St	954016-10791565	
9	-37.81105	144.95963	Badger Vs Hawk	4.5	10	Melbourne	3000	Victoria	333 La Trobe St	255100-6654398	
10	-37.53195	145.34125	Songbird Cafe & Larder	4.5	9	Kinglake	3763	Victoria	10 Whittlesea-Kinglake Rd	552194-20003590	
11	-37.81677	144.96558	Cafe Issus	4	248	Melbourne	3000	Victoria	8-10 Centre Pl	255100-1852475	
12	-37.82157	145.02643	Rustica	4.5	16	Hawthorn	3122	Victoria	121 Power St	844564-12913662	
13	-37.72825	145.04266	Austrian Club Melbourne	4.5	3	Heidelberg	3081	Victoria	90 Sheehan Rd	552183-14761136	
14	-38.34814	143.42726	Toad Hall Tearooms	-1	0	Colac	3249	Victoria	1470 Princes Hwy	261656-2155644	
15	-37.82058	144.95029	Tahini 2 Lebanese Diner	4	3	Melbourne	3008	Victoria	727 Collins St	255100-14008578	
16	-37.70418	144.91637	Twenty One Days Later	4	3	Glenroy	3046	Victoria	10 Post Office Pl	4399277-17767097	
17	-36.7612	144.27806	Golden Star Chinese Restaurant	3.5	48	Bendigo	3550	Victoria	382 Hargreaves St	255347-5089547	
18	-37.76123	144.96294	Nando's Brunswick	4	5	Brunswick	3056	Victoria	671 Sydney Rd	947958-5109933	
19	-37.8122	144.96553	Globe Alley	-1	0	Melbourne	3000	Victoria	1 Globe Alley	255100-17574182	
20	-37.6869	144.86844	True Blue	3	45	Tullamarine	3043	Victoria	398 Melrose Dr	580517-2623693	
21	-37.96557	146.97449	Mr Pizza	2.5	9	Maffra	3860	Victoria	52 Johnson St	552204-4786101	
22	-38.110474	147.06866	Hunting Ground	4.5	49	Sale	3850	Victoria	102 York St	255356-7283149	
23	-37.73731	144.89172	Noodle Hut	3.5	4	Niddrie	3042	Victoria	401 Keilor Rd	4367892-4781396	
24	-35.98921	146.00615	Blacksmith Provedore	4	22	Mulwala	2647	New South Wales	84 Melbourne St	1061160-15072004	
25	-37.68413	144.88261	Hungry Jacks Pty Ltd	2.5	24	Tullamarine	3049	Victoria	239 Mickleham Rd	580517-5106519	
26	-38.1544	145.16415	Karingal Gloria Jeans	4.5	7	Frankston	3199	Victoria	Shop 14 330 Cranbourne Rd	552173-4791666	
27	-38.04943	144.14859	Paradise Restaurant	3.5	2	Bannockburn	3331	Victoria	U 1 71 Holder Rd	2218271-13952741	
28	-37.85722	144.89804	Gloria Jean's Coffees	3	3	Williamstown	3016	Victoria	64 Douglas Parade	261669-17341295	
29	-37.84926	144.99026	W 'n C	4	1	Prahran	3181	Victoria	96 - 102 Greville St	261664-8794956	
30	-37.75227	145.34457	Cavehill Cafe	3.5	2	Lilydale	3140	Victoria	70 Cave Hill Rd	552201-5096807	
31	-37.992786	145.0756	Cafeholics	4	9	Parkdale	3195	Victoria	220 Como Pde W	1736782-8854182	
32	-37.42274	144.56494	Rasputin's	-1	0	Macedon	3440	Victoria	40 Victoria St	552203-18455173	
33	-38.578373	146.01312	Pandesal Bakery	4.5	40	Meeniyan	3956	Victoria	124A Whitelaw St	1787702-12216411	
34	-37.812443	144.96358	Spudbar Emporium	4	12	Melbourne	3000	Victoria	321/287 Lonsdale St	255100-12215038	
35	-37.72632	145.0709	Ming's Kingdom	-1	0	Melbourne	3085	Victoria	72 Aberdeen Rd	255100-727189	
36	-38.15073	144.36552	Cafe Bocca	4	2	Geelong	3220	Victoria	281 Ryrie St	255350-10769067	
37	-37.83005	145.0557	Franco-Belge	4.5	59	Melbourne	3123	Victoria	9 Evans Place, Hawthorn East	255100-14881691	
38	-37.7556	144.91652	The Cherry Blossom Asian Restaurant	4	16	Essendon	3040	Victoria	18 Russell St	2019289-2462201	
39	-37.813656	144.96144	Bincho Boss	5	12	Melbourne	3000	Victoria	383-385 Lt Bourke Street	255100-17802740	
40	-37.86555	145.02759	Malvern Curry House	3.5	7	Malvern	3144	Victoria	17 Glenferrie Rd	1009368-8412138	
41	-37.87399	145.40604	The Crunchy Nut Cafe	4.5	9	Monbulk	3793	Victoria	122 Main Rd	552215-2732910	
42	-37.5346	144.89793	Wild Bean Cafe	-1	0	Mickleham	3064	Victoria	470 Donnybrook Rd	15215361-20208180	
43	-38.09576	145.26547	Divine noodles & Sushi	5	7	Cranbourne	_	Victoria	Eve Central Shopping Centre	552156-4212723	
44	_	_	Marysville Takeaway	-1	0	Marysville	_	Victoria	Shop 3/20 Murchison street	552209-20995448	
45	-37.811874	144.96364	Es Teler 77	3	4	Malvern	3144	Victoria	19 Glenferrie Rd	1009368-5096806	
46	-37.71416	145.14902	Sammy's Charcoal Chicken	4	17	Eltham	3095	Victoria	8 Commercial Pl	495048-738146	

ta_restaurants +

Figure 3.8: Feature Vector 3 sample data: Restaurant location and ranking on TripAdvisor

is $18 + 1 + 9 + 20 = 48$.

Finally, it is worth noting that multiple features in feature vectors 2 and 3 require the inverse of the distance between the POI and the house (this is the "inverse law" described on multiple occasions in previous sections). To calculate the distance between the POIs and the houses, we used *haversine_distances* from the *scikit-learn* Python library to calculate the Haversine/Great Circle distance along the Earth's surface between the latitude/longitude coordinates of the POIs and the houses[6].

Mathematically, the calculation of the Haversine distance is based on the *haversine* function:

[6] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.haversine_distances.html`

Code	Label		
206011105	Brunswick	Median Age - Persons (years)	32.8
		Working Age Population (aged 15-64 years) (%)	79.2
		Persons - Total (no.)	27,435
		Population density (persons/km2)	5,335.0
		Speaks a Language Other Than English at Home (%)	27.9
		Median employee income ($)	50,573
		Mean employee income ($)	59,254
		Median investment income ($)	200
		Mean investment income ($)	4,926
		Total income earners (excl. Government pensions and allowances) (no.)	16,555
		Total income earners (excl. Government pensions and allowances) - median age (years)	34
		Median equivalised total household income (weekly) ($)	1,156
		Completed Year 12 or equivalent (%)	74.8
		Bachelor Degree (%)	29.4
		Unemployment rate (%)	5.9
		Average household size (no. of persons)	2
		Households with mortgage repayments greater than or equal to 30% of household income (%)	4.2
		Households with rent payments greater than or equal to 30% of household income (%)	16.6
		Homeless rate per 10,000 persons (rate)	67.7
		Median commuting distance to place of work (kms)	8.5

Figure 3.9: Components of Feature Vector 4 and Sample data set

$$\mathrm{hav}(\theta) = \sin^2\left(\frac{\theta}{2}\right) = \frac{1-\cos(\theta)}{2} \tag{3.13}$$

The Haversine/Great Circle distance between two place x and y on the surface of the Earth, $D(x, y)$, is then given by the following formula:

$$D(x, y) = R\,\mathrm{archav}(\mathrm{hav}(x_1 - y_1) + \cos(x_1)\cos(y_1)\,\mathrm{hav}(x_2 - y_2)) \tag{3.14}$$

where R is the radius of the Earth, archav is the inverse haversine function (arc-haversine function), (x_1, x_2) are the latitude and longitude of x in radians, and (y_1, y_2) are the latitude and longitude of y in radians. For the purposes of practical computation, there is a form of Equation 3.14 written only using sin, cos,

and arcsin, which can be found on sources such as the documentation page of the scikit-learn *haversine_distances* function[7].

Housing Appraisal Classification Task

For the *Housing Appraisal Classification Task*, we use Logistic Regression as the baseline algorithm/model, as this is what is commonly used for classification problems in Classical Economics. The performance of four other algorithms were compared against the performance of the baseline: Support Vector Machine, Multilayer Perceptron, k-Nearest Neighbour, and XGBoost.

The practical implementation of the pipeline is very similar to that of the *Housing Appraisal Regression Task*. The data sets used for the training, validation, and testing of the 4 models described in the previous paragraph are from the data sets described in Section 3.5.1. The one major difference for the classification pipeline, is that with the data for the target variable, we had to transform the *eventprice* column data from the actual dollar price to the price range, in order to formulate housing appraisal as a classification task.

3.5.3 Performance Evaluation

We use standard performance metrics for both the *Housing Appraisal Regression Task* and the *Housing Appraisal Classification Task*, which are described below. It should be noted that since the target variable is a directly measurable quantity in the regression task, and a directly measurable class in the classification task, no human expert annotation is required for the target variable data, unlike other artificial intelligence fields such as Natural Language Processing, where human expert annotation is commonly required. Thus, in the problem of this chapter, we

[7] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.haversine_distances.html`

do not need to establish what are gold/silver/bronze standards of annotation, nor do we need to calculate statistics of inter-annotator agreement (such as Cohen's Kappa, κ) in order to measure the inherent variability of the target class variable.

For the Housing Appraisal Regression Task, performance evaluation of the dollar value estimation uses the standard metrics for regression problems: Mean Absolute Error (MAE), Root-Mean-Squared Error (RMSE), and the Coefficient of Determination (R^2).

For the Housing Appraisal Classification Task, performance evaluation of the price range estimation uses the standard metrics for machine learning classification problems: Accuracy, Precision, Recall, and $F1$, which are defined as follows [117]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.15}$$

$$Precision\,(P) = \frac{TP}{TP + FP} \tag{3.16}$$

$$Recall\,(R) = \frac{TP}{TP + FN} \tag{3.17}$$

$$F1 = 2\frac{PR}{P + R} \tag{3.18}$$

where TP and TN stand for True Positive and True Negative, which measures the number of prices being classified within the correct range. FP and FN stand for False Positive and False Negative, which measures the number of prices being classified within the incorrect range.

3.6 Results and Analysis

3.6.1 Overall Performance

Table 3.1 shows the performance of the different algorithms in the task of dollar value estimation. The performance of XGBoost considerably better than the other algorithms, achieving an R^2 value of 0.8779, compared to the baseline performance of 0.6422.

Table 3.1: Housing Appraisal Regression Task

Model	MAE	RMSE	R^2
Linear Regression	0.2450	0.3525	0.6422
Support Vector Machine	0.4119	0.7620	-0.6723
Multilayer Perceptron	0.1994	0.2904	0.7572
XGBoost	**0.1428**	**0.2059**	**0.8779**

Table 3.2 shows the performance of the different algorithms in the task of price range classification. The performance of XGBoost again is considerably better than the other algorithms.

Table 3.2: Housing Appraisal Classification Task

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.6219	0.5432	0.4452	0.4662
Support Vector Machine	0.5172	0.5162	0.3533	0.3697
k-Nearest Neighbour	0.8044	0.7639	0.7631	0.7633
Multilayer Perceptron	0.7219	0.6778	0.6145	0.6328
XGBoost	**0.8601**	**0.8590**	**0.8226**	**0.8395**

3.6.2 The Importance of the Regional Cluster Variable

In this section, we discuss a very important finding that regional cluster variables played a significant role in prediction [69,155]. Table 3.3 used different combination of subsets of feature vectors. Regional economic cluster as feature vector 3 reached

0.6290 in R^2 individually, and also consistently reached the highest performance when combined with other feature vectors. It is noticeable that feature vectors 1 and 3 combined could reach 0.1552 in MAE, 0.8585 in R^2 which means housing attributes with regional economic cluster variables can give a good prediction. This shows that in the HNED model, feature vector 1 and feature vector 3 are the two most important feature vectors affecting the HNED model's performance in the housing appraisal tasks.

Table 3.3: Housing Price Estimation Performance Using a Subset of the Feature Vectors

Model	Feature Vectors Used	MAE	RMSE	R^2
XGBoost	1	0.3614	0.4778	0.3427
XGBoost	2	0.3412	0.4578	0.3965
XGBoost	3	**0.2582**	**0.3589**	**0.6290**
XGBoost	4	0.3214	0.4399	0.4427
XGBoost	1, 2	0.2568	0.3482	0.6508
XGBoost	1, 3	**0.1552**	**0.2217**	**0.8585**
XGBoost	1, 4	0.1875	0.2728	0.7856
XGBoost	2, 3	0.2492	0.3466	0.6540
XGBoost	2, 4	0.2664	0.3680	0.6100
XGBoost	3, 4	0.2507	0.3486	0.6500
XGBoost	1, 2, 3	0.1530	0.2180	0.8632
XGBoost	1, 2, 4	0.1699	0.2447	0.8275
XGBoost	1, 3, 4	**0.1446**	**0.2086**	**0.8747**
XGBoost	2, 3, 4	0.2404	0.3345	0.6778
XGBoost	1, 2, 3, 4	**0.1428**	**0.2059**	**0.8779**

3.6.3 Analysis using the Shapley Additive Explanations library (SHAP)

As a final part of the analysis, we used the Python SHAP (SHapley Additive exPlanations) library to look at which features "dominate" the estimation of the target variable[8]. As discussed in Sections 3.4 and 3.5.2, the input feature space of the NHED model is almost 50 dimensions. Using the Python SHAP library, we are able to check experimentally which specific features of the overall feature space "dominate" the model's performance.

Figures 3.10 and 3.11 show two different plots of the SHAP analysis of relative dominance of each of the features on the model's performance. From the figures, we can see that the most dominant features are number of bedrooms (feature vector 1), distance to CBD (feature vector 3), and areasize (feature vector 1); followed by demo_Median_investment_income (feature vector 4), and demo_Population_Density_(persons_km2) (feature vector 4).

3.7 Discussions and Implications

Our results show how housing price is related to housing attributes, location and socioeconomic characteristics. Each element contributes different implications for different social agents, such as home buyers, investors, local and regional councils, urban planners.

3.7.1 Implications for Home buyers

Generally, home buyers consider both current living functions and investment value of a property. Firstly, housing attributes are the primary focus to meet the daily

[8] `https://shap.readthedocs.io/en/latest/index.html`

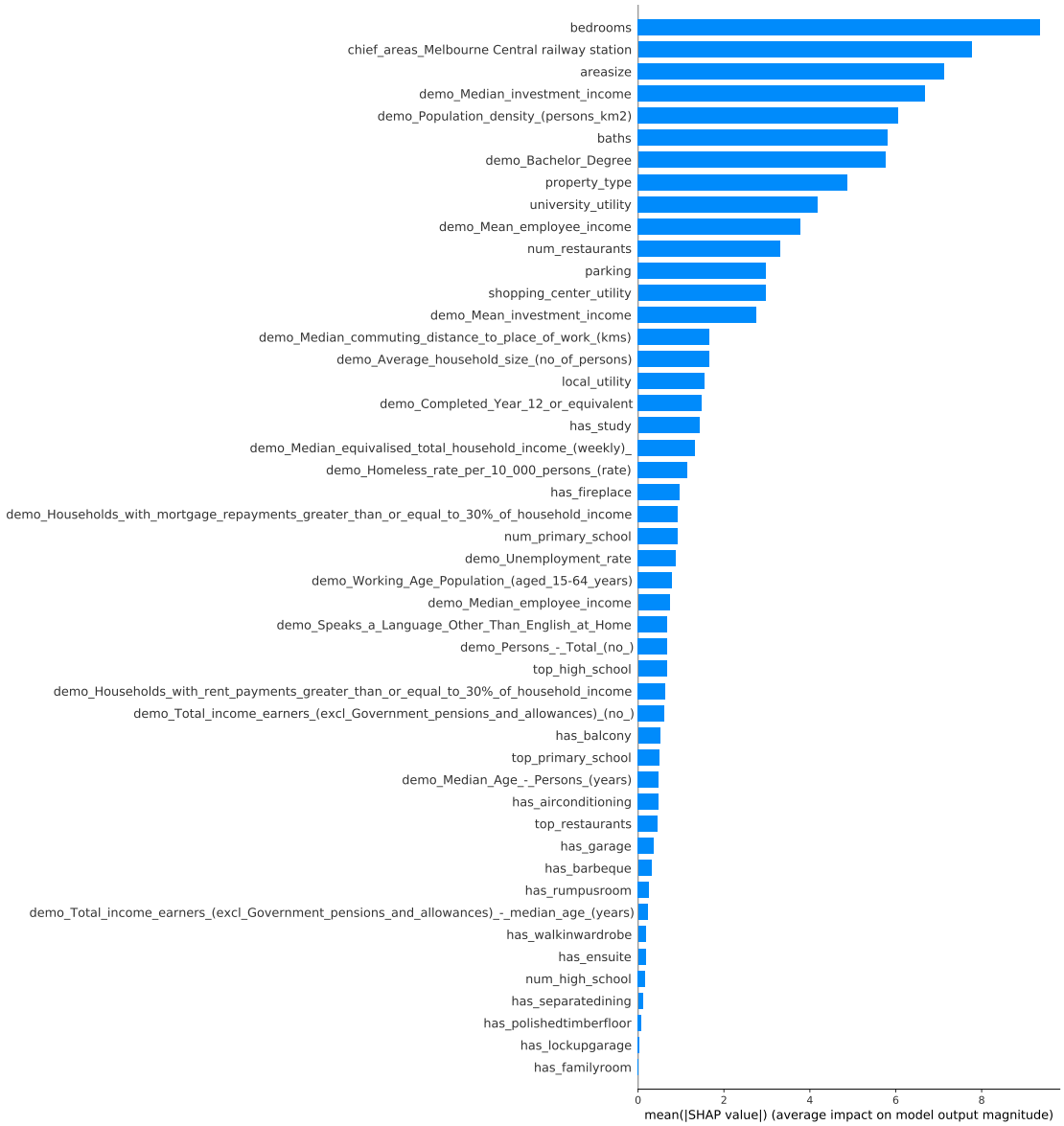


Figure 3.10: Mean SHAP Value Impact on Model Output Magnitude (All Features)

needs of dwelling. Extra bedroom or bathroom can drive the property value up with better functionality. Secondly, people value more for being in a highly ranked school zone. Walking distance to schools, supermarkets, public transport are preferable by most home buyers. However, the power of strong connection to regional economic clusters may be neglected in the decision making process. To capture a long-term investment return, home buyers need to identify a location

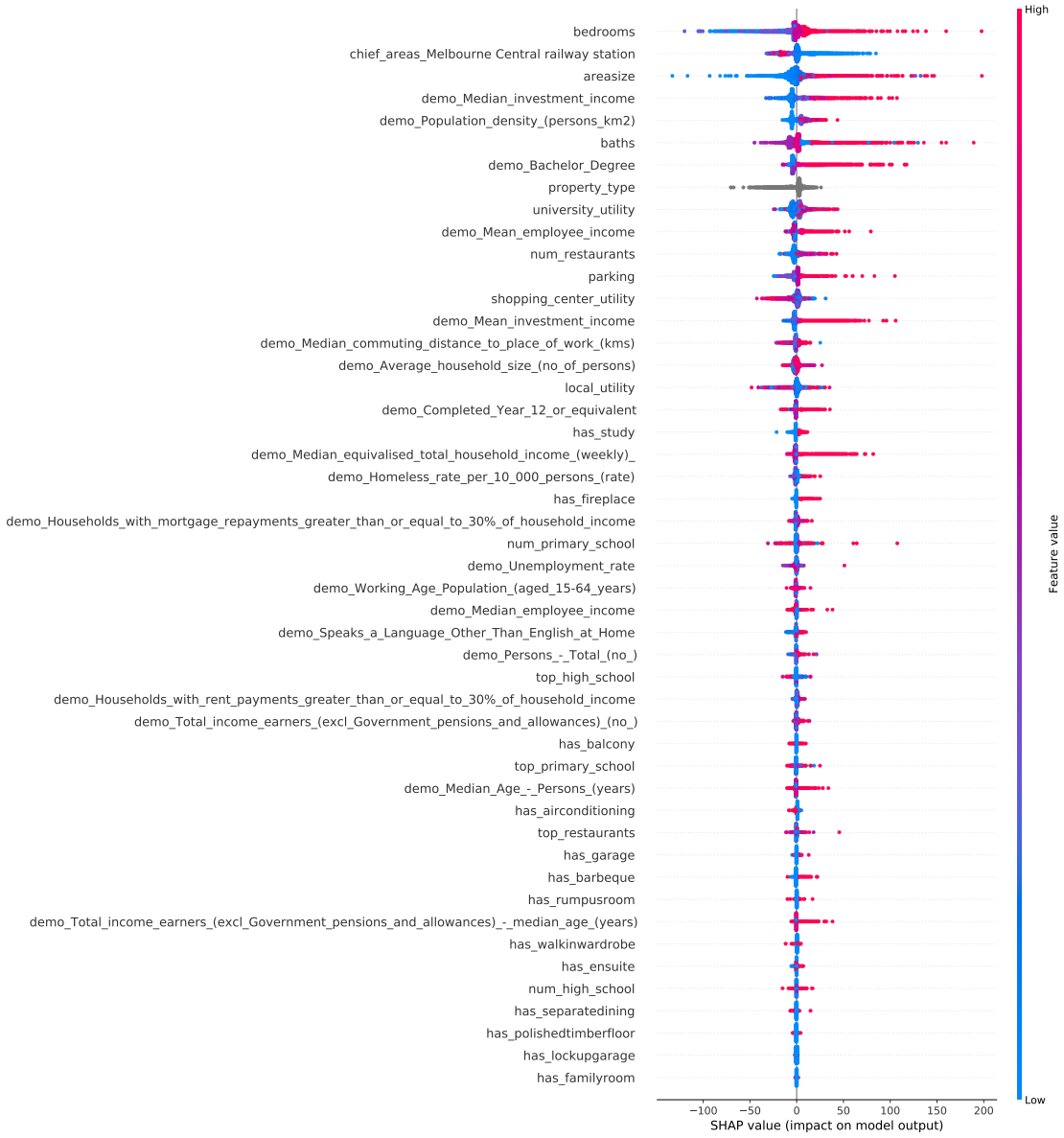


Figure 3.11: SHAP Value Impact on Model Output (All Features)

with growing highly educated population, with a strong connection to regional economic clusters.

3.7.2 Implications for Councils and Urban Planning

Based on our results, income and education are strong indicators to drive the housing value, especially investment income is more relevant. High investment in-

come indicates people with multiple source of income or they are business owners. Firstly, councils can attract these group of people by stimulating business district, business park development, or building strong industrial clusters. Secondly, councils can attract young highly educated or highly skilled people to settle down by providing high quality infrastructures, such as high quality public schools, welcoming new campus for highly ranked private schools, or planning shopping centres and sports facilities. Thirdly, councils can make strategic long-term planning of fostering regional economic clusters in a prominent industry, such as forming an IT cluster, education cluster, medical cluster or warehouse cluster, etc. By forming a super cluster, local areas can concentrate human capital in one expertise direction and achieve high economic growth rate.

3.7.3 Implications for Real-estate Investors and Developers

Both investors and developers need to identify high demand of housing. Investors focus on both future return and sustainable rental demand. Good location is essential for both rental income and long-term return. Rich POI in the neighbourhood and close distance to regional economic cluster would guarantee a good location. The growth of high educated, skilled population in one area will contribute for future demand of such location and hence drive the future property return. Developers also need to consider maximise the housing needs for potential buyers, extra bedroom and bathroom can significantly increase house value.

3.8 Conclusions and Future Work

We have studied how factors beyond neighbourhood impact housing values. Specifically, we established regional economic clusters as the significant source of impact beyond neighbourhood. We presented our housing price appraisal model that combined housing attributes, neighbourhood characteristics, and demographic factors. Our model using the XGBoost algorithm has reached 0.88 in R^2, showing the significant impact of regional economic clusters.

Our work enlightens two related research questions worthy of future investigation. (1) Building a customised recommendation system for home buyers. This system aims to tailor and optimise personal needs of affordable living, and provide smart suggestions for trade-offs between different needs and opening up opportunities for locations. (2) Methods to systematically identify regional economic clusters and appropriately weight these clusters in our model. Currently, we used revenue of the entity, the total number of floor space, etc. to weight shopping centres (which is one type of a regional economic cluster). Is there a better way of doing the weighting in the model?

With better understanding of this behavioural mechanism, we could improve our community, facilitate sustainable gentrification, and lead to location diffusion, population growth, and new regional economic clusters emerging.

Acknowledgement

This work has been partly funded by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 824019 and the DAAD-PPP Australia project "Big Data Security".

CHAPTER 4

UNDERSTANDING HOUSING MARKET BEHAVIOUR FROM A MICROSCOPIC PERSPECTIVE

4.1 Chapter Abstract

This paper discusses the usage of micro-level behavioural data for understanding macro-level housing market behavior. We use sentiment analysis to examine local newspaper articles discussing real estate at a suburb level in inner-west Sydney, Australia. We calculate the media sentiment index by using two different methods, and compare them with each other and with the housing price index. The use of media sentiment index can serve as a finer-grained guiding tool to facilitate decision making for home buyers, investors, researchers and policy makers.[1]

4.2 Introduction

The world was plunged into one of the greatest global economic recessions since the Great Depression, in the aftermath of the 2008 global financial and economic crisis (GFEC). The cause of the GFEC was rooted in serious financial issues in the US housing market. Whilst some economists warned that there may be a bubble in the US housing market a few years before the crisis (e.g. [27]), this was far from the popular view amongst economists at the time. Indeed, according to the Economics Nobel Laureate Paul Krugman, most of the economic world did not see the GFEC coming before it happened [85].

Understanding business cycles, otherwise known as economic cycles, is one of

[1]This chapter is based on the following article:
Jiaying Kou, Xiaoming Fu, Jiahua Du, Hua Wang, and Geordie Z. Zhang. "Understanding housing market behaviour from a microscopic perspective." In 2018 27th International Conference on Computer Communication and Networks (ICCCN), pp. 1-9. IEEE, 2018.

the greatest challenges in Macroeconomics. Many traditional lines of development in the past have resulted in more and more top-down, sophisticated mathematical models for dealing with this complex problem of prediction. However, there has been another lesser-known approach, suggested by researchers such as esteemed economists Wassily Leontief and Herbert Simon, which argues that finding the solution of forecasting business cycle lies not in developing more and more technically sophisticated models, but investigating newly available data sets, which add greater dimensionality to the existing data sets and in particular, provide empirical data about human behaviour and decision making (see [90, 128]). Simon asserted:

> Existing uncertainties about the correct explanations for economic growth and business cycles cannot be settled by aggregative analysis within the neoclassical framework. Current disputes in theory rest largely on ad hoc, casually empirical, assumptions about departures from perfect rationality under uncertainty. Such disputes can only be settled by painstaking microeconomic empirical study of human decision making and problem solving [128, p. 35].

This latter approach has blossomed in recent times, supported by recent theoretical breakthroughs in behavioural finance, especially in the area of irrationality, which focuses on the multi-disciplinary discovery of human economic and financial decision-making behaviour with psychological experimental methods. Furthermore, recent developments in Internet technologies provide a big-data-based method for microeconomic empirical study. These advancements have made more practicable Simon's vision of "painstaking microeconomic empirical study of human decision making and problem solving".

A key element to build a functional forecasting model for economic systems, where humans are the main actors, is to find valuable data sets about micro-

level human decision-making behaviours, especially irrational behaviours. This paper attempts to understand housing market behaviours from such a data-driven approach. More specifically, the aim of this paper is to study the macro-level housing market cycle problem from a microscopic perspective, with the support of newly available micro-level data sets and available methods of big data analysis, especially text and sentiment analysis.

We propose a big-data-based method for understanding housing market behaviour, with a preliminary case study of Australian housing market. Instead of following the traditional macroeconomic perspective to build models top-down from a national-level, we study the housing market behaviour bottom-up, starting from a suburb-level, but not ending at the suburb-level. There are several reasons for studying housing market behaviour at the suburb-level:

Firstly, the availability of big data nowadays provides us with richer micro-level data sets [43]. These data sets were not readily available previously, due to the difficulty of gathering and measuring activities, such as social network, Tweets on a social event, geolocation data, commercial preferences, job search, etc. These data sets contain information relevant to decision making behaviour at a micro-level.

Secondly, new variables and patterns can be extracted and studied at finer granularity. These studies can expand our understanding of real estate market behaviour at the macro-level. The rich micro-level data sets can provide the opportunity to look inside the "black box" of suburb-level economic status, and to understand how people make investment decisions. The new findings may improve housing market cycle modelling, and may generate a better predictive results at the macro-level.

Thirdly, houses are, by nature, not fungible, which brings challenges for property valuation. The availability of micro-level data sets provides new methods

for property valuation. Based on the anchoring effect [146], people tend to use an anchor to adjust their perceived valuation of an item. Property valuation can serve as an anchor to influence people's decision-making behaviour. We could assume that widely available online information of selling prices and other relevant information of housing in the neighbourhood can shape the housing price level in people's minds.

In this paper, we granulate the macro-level housing market cycle problem into a suburb-level housing market behaviour problem. With the possibility of discovering new indicators, relationships, patterns, and behaviours at a suburb-level, we are able to find the link between suburb-level and national-level, and reach a better understanding of the housing market behaviour. Two challenges exist though: one is to collect and interpret new data which is more fine-grained than traditional macroeconomic data; the other is to exploit big data analytic methods in dealing with multiple variables and outcomes. This paper will address the challenges with the aid of sentiment analysis of housing price index (HPI) and media sentiment index (MSI).

The rest of this chapter is organised as follows. Section 4.3 motivates the discussion with a real-world problem, and presents the literature review. Section 4.4 explains the overarching theoretical framework. Section 4.5 describes the methodology of sentiment analysis using local media data source. Section 4.6 shows the results and further explores the relations between HPI and MSI. In Section 4.7, conclusions and future work are discussed.

4.3 Related Work

In this section, we will use a real-world problem to motivate our work, and review the relevant literature on big data in economics, especially in housing market study. Machine learning techniques and data collections are discussed for housing market research in suburb-level study.

4.3.1 A Real-World Problem

As discussed in the previous section, one of the biggest challenge for business cycle is to understand the market behaviour and to generate a better forecasting. Here the challenge is how can we find the pathway to use the advantage of new development in the Internet and new understanding of behavioural economic and finance to understand the housing market. The problem is how do we apply the new data sets and analysis methods to understand the housing market behaviour. Our preliminary research example is: do home buyers make decisions based on the sentiment? Can we use a sentiment analysis to facilitate our understanding of the housing market behaviour?

To such end, one of the core ideas we need to use is that of the Housing Price Index (*HPI*). The Standard and Poor's Case-Shiller home price indices are one of the standardised *HPI* in the US, which has been studied in a finer granularity of the postcode level. Australian Bureau of Statistics (*ABS*) also publishes a seasonal *HPI* in national level and 8 major cities, but not in the suburb-level. In order for a more detailed behavioural study on Australian housing market, one may resort to establishing indicators at a more microscopic level, such as a suburb-level *HPI*.

Three distinct *HPI* methods are commonly used (see [24]): the hedonic approach (which uses the characteristics of the property itself and its surrounding environment as main factors for house pricing), the repeated sales approach (which

calculates changes in the sales price of the same piece of real estate over specific periods of time), and the stratification approach (which decomposes the market into separate types of property and estimates prices for different properties with a real estate price index).

4.3.2 New Sources of Data for the Housing Market

Instead of using traditional sampling techniques which may subject to bias, we can enjoy the availability of larger data or even nearly complete data sets (big data), which typically leads to higher accuracy in behavior prediction. Besides this advantage, we emphasize on the nature of the newly available data.

The growing online data has gradually enlarged the variety of economic variables. Pioneering researchers started adventure in private sector companies who have first-hand real-time online data, such as Google, MasterCard, Federal Express, UPS [29]. Also, credit card transactions and eBay online auctions nowadays become part of the economic variables [41,100]. Furthermore, consumer behaviours can be analysed with mobile internet data sets from major telecom providers [68].

Google Trend Index

Google Trend Index can be regarded as a new kind of data. [29] built prediction models using Google search engine trend index as an indicator and found it effective to do near-term economic prediction. Many researchers followed this idea to implement in different areas such as automobile market, entertainment industry, tourism industry, labour market [4,23,60,101], and found significant improvements in prediction accuracy.

Particularly, in the real estate area, [153] did a pioneer study of housing market cycle prediction in the US using Google Trend Index and showed a significant

improvement in prediction accuracy, compared with the state-of-the-art prediction results. [101] also showed the power of using Google trend index as an indicator in predicting house price index in UK. They also provided a comprehensive discussion of the searching key words selection.

Unfortunately, studies in [101, 153] do not go into the suburb-level, due to the current Google trend's degree of releasing. Another possible concern is that it may not be accurate due to the reason that people may move to different suburbs and reduce the location precision.

Recently, researchers started to exploit image and language information to study economic behaviour. The nature of image and language information data is beyond the traditional econometric linear regression processing method.

The Use of Text Analysis for Housing Market Studies

Analyzing texts in the social media big data content has a rich literature, such as [69, 113]. In economics, the early application of natural language processing is to study the feasibility of predicting stock market prices. The sentiment analysis is used in processing the data collected from Blogs, News, Twitter feeds, etc.

Literature about the relationship between sentiment analysis and stock returns is quite rich. The sentiment analysis in stock market started before the big data era. The arguments started with the challenge in behavioural finance about the efficient market hypothesis (*EMH*). *EMH* assumes all the publicly known information about a company has been reflected in its stock price, which means there is no opportunity to arbitrage the difference between the current stock price and future stock price based on the available information. On the contrary to *EMH*, researchers started to discover overreactions, anomalies in the stock market behaviour (see [32,72]). Historical stock prices showed predictive power for the future prices, which means current stock prices of companies don't fully reflect intrinsic

value of companies. There exists opportunities to arbitrage based on publicly known information.

Following the investigation of behavioural finance, [8] used financial indicators to form the sentiment proxies, but these indicators are not available in the housing market [130]. With the study of the predictive power of online chat, the nature of data for sentiment study has changed into textual analysis. Relevant literature includes [5, 15, 58, 61, 123, 160]. Particularly, [123] predicted the stock price with analysis of textual financial news articles. [15] applied Granger causality analysis and Self-Organizing Fuzzy Neural Network to analyse Twitter feeds. Based on the public mood status generated from Twitter feeds, they predicted the daily changes with 86.7% accuracy.

In [130], the authors presented an early study on applying textual sentiment analysis on housing market area. The authors found that newspaper sentimental analysis can be a good predictor for the future housing prices in the US. Our paper follows the data collection method used in [130], and goes beyond [130] in that we use local suburb-level – instead of city- or country-level – newspaper text as the source to define sentiment on housing market. Interestingly, to the best of our knowledge, there are no prior studies which use Twitter feeds as a sentimental indicator, which is however widely used in stock market. This may become an inspiration for the future research work to investigate the feasibility of predicting housing prices using local Twitter feeds.

Image as Economic Data

Some researchers have used satellites image as an indicator for economic outcomes prediction, such as Google Street View to predict income level [59], the illumination of night to interact with economic output [64]. The overview of relevant literature has been provided by [39]. Besides satellite, the location-based social network

provides us the flow of locations that patterns of human traffic behaviours can be analysed [161].

4.3.3 Big Data Methods for Housing Market Studies

Traditional econometrics approach employs regression modelling to detect and summarise relationships among variables expressed in limited sizes of survey or statistics data. We apply text analysis and sentiment analysis, to process data sets newly collected in this paper.

Recently, there are some studies to discuss the perspective of applying machine learning algorithms in econometrics, such as [6, 41, 105, 148]. Coherent with our thinking of "bottom-up" way of approach of real estate housing price study, machine learning is good at revealing data's feature by "letting the data speak", instead of a "top-down" theory-based deductive investigation. The traditional econometric modelling typically relies on linear regression; when the data is non-linear or involves intensive interactions, machine learning approach like decision trees may perform much better in extracting related features and predicting housing prices than the traditional econometric approaches.

Particularly, in real estate study, there is a new opportunity to get access to more integrated big data, since many governments nowadays encourage research on economic data mining, meanwhile making more available real estate data online freely. For example, [105] used a few machine learning methods to test 150 covariates for house valuation. They demonstrated how machine learning methods, such as LASSO and Random forest, could work better in solving heterogeneous problem than ordinary least squares.

4.4 Theoretical Framework

As discussed in the introduction, in this study we explore bottom-up and behavioural approaches for understanding the Australian housing market, instead of traditional top-down and non-behavioural methods. Practically speaking, we use a two-step approach. In the first step, we explore how we can gauge people's attitude and behaviour, at a grass-roots level, towards the housing market, and how this attitude/behaviour changes over time. This leads to the news media sentiment index work that will be discussed shortly. The news media sentiment index work is our attempt at constructing a proxy index for gauging grass-roots attitude/behaviour towards the housing market in Australia. In the second step, we compare the media sentiment index constructed in step one with a traditional economic measure of the housing market, the *Housing Price Index* as published by the Australian Bureau of Statistics. From this comparison, we analyse to what extend the media sentiment index may be used as a leading/lagging indicator for the housing price index. These two steps together provide an attempt at analysing the macroscopic housing market by using grass-roots and behavioural data, and to the authors' knowledge, the first attempt of such study in Australia.

4.4.1 News Media Sentiment Analysis in Housing Market

The term "animal spirits" has attracted substantial attention in recent years since behavioural economics and finance gradually became mainstream economic topics [1]. The wild movements of prices in stock market cannot be explained by *EMH*, which believes that rational investors will drive stock prices close to its intrinsic value based on the company released news. This leads to the discussion that "animal spirits" also plays a role in human decision-making which can drive stock prices away from the real value.

Economists contributed to our understanding of irrationality by designing psychological experiments to show evidence that decision-making under uncertainty situation doesn't follow the rational assumption. People tend to make decisions based on an anchor, even if the anchor has no casual relationship with the target [146].

This leads to a natural inquiry: what can be the anchor that people use for decision-making?

Robert Shiller [126] discussed the limitation of human information transmission and processing. This provides the ground for anchors. Shiller summarised that people's decision-making are highly influenced by stories, word-of-mouth and face-to-face communications. The conventional media, such as print media, television, radio is good at speaking ideas, but limited with empowering actions. The effectiveness of new media, such as e-mail, Twitter, Weibo, Skype, FaceTime, is still unknown. This provides us a theoretical ground that the anchoring effect for such media of information transmission can be tested.

We cannot conclude that newspaper is the best proxy for sentiment analysis, but it can be a starting point for testing. Thus, we use new media sentiment analysis as the first step of our inquiry. The method of constructing the data and calculating the media sentiment index is discussed in detail in Section 4.5.

4.4.2 Using Media Sentiment Index as an indicator for Housing Price Index

In the previous subsection we explained why we are using news media and media sentiment index as a starting point for our inquiry of gauging grass-roots attitude/behaviour. As we shall demonstrate in later sections, the media sentiment index may be calculated for different years, and thus deriving a time series MSI

for a period of time.

How do we use this to establish a relationship with movements of the housing market? Traditionally, the housing price index is calculated as a measure of movements in the housing market. Once we establish a time series MSI, by comparing this time series with the time series of the housing price index, this would allow us to look for patterns between the two indices, and investigate to what extent the MSI could be used as a proxy leading/lagging indicator for the HPI.

4.5 Methodology

In the previous section, we provided a theoretical framework of how to approach the problem of understanding the housing market from microeconomic and behavioural information. This section discusses the methodological issues in detail. Specifically: how to calculate the media sentiment index, and how to investigate how well the media sentiment index could be used as an indicator for the housing price index.

For the media sentiment analysis, we used two related but different methods. As this research is new to Australia, we used a previous media sentiment analysis based in the US by Soo et al [130], to create a baseline media sentiment index for the Australian media dataset we used. We then created a more generalised modification of Soo's method using the VADER sentiment classifier [70], and compared the results from both media sentiment indices in Section 4.6.

4.5.1 Data and Preprocessing for Media Sentiment Analysis

The media data used for sentiment analysis in this paper is sourced from Australia and New Zealand News Stream, an online database that offers newspapers from Australia and New Zealand's major publishers. Specifically, we choose Inner West Courier[2] as our case study due to its uniqueness in publishing real property information of Glebe and its vicinity among all Sydney local newspapers. To retrieve most relevant news records, we initiate a query using the keyword "real estate" in the database, and collect all returned items via web crawling. As a result, we manage to obtain news materials published between 21 June, 2007 to 27 February, 2018. Figure 4.1 demonstrates the number of news publications over years.

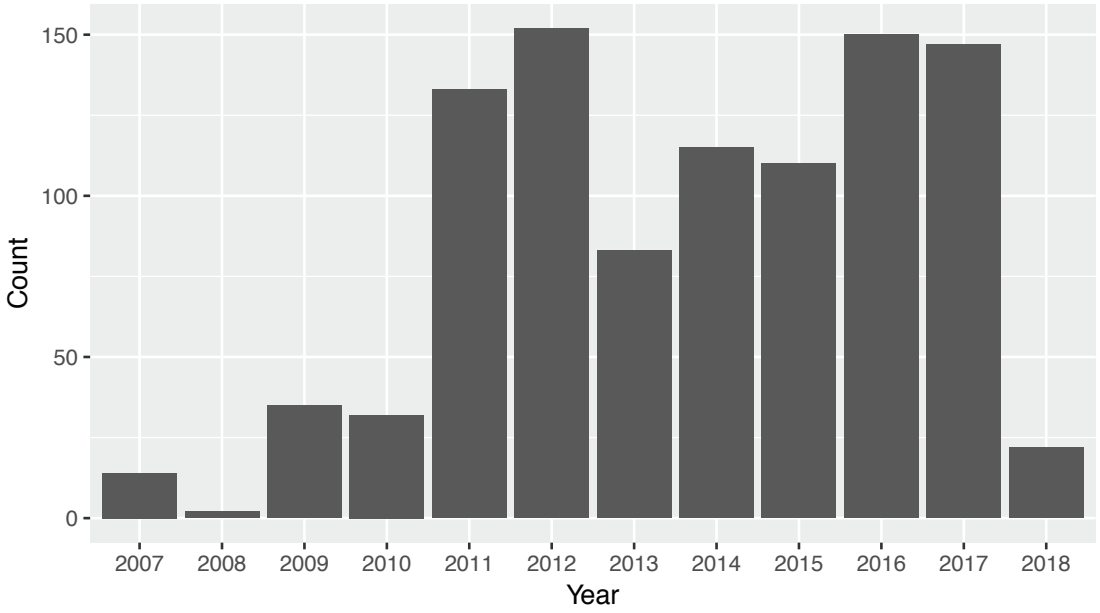


Figure 4.1: Distribution of Annual Publication Numbers

The collected materials are raw web pages written in HTML syntax. Such markup tags and attributes, while useful in locating specific elements in a DOM

[2] https://www.dailytelegraph.com.au/newslocal/inner-west

tree, benefit nothing in sentiment analysis. Therefore, we only extract necessary content from individual documents. As Table 4.1 shows, one extracted example includes the title, abstract, content, author and published date of a news article. For fine quality, stopwords, redundant lines and white spaces are removed during word tokenization. Exceptionally short articles indicates advertising otherwise conveys little meaning, and thus those whose contents contain less than 50 words are also discarded.

Table 4.1: News Article Sample

Attribute	Value
ID	921154426
Title	Hot spots tick boxes for best-buy homes of year
Author	Craw, Victoria
Date	14 Feb 2012
Abstract	Marrickville's median house price is $750,000, reflecting capital growth of 42.9 per cent in last five years, according to RP Data.
Content	Marrickville and Abbotsford are top spots to buy in for 2012, according to PRD Nationwide research. (141 words)

Finally, the whole dataset consisting of 995 news articles is indexed in chronological order and stored in a database system for fast querying and retrieval. Each news article has 370.19±189.97 words.

4.5.2 Dictionary Preparation

We leverage two dictionary-based approaches to quantify the sentiment of news papers. The first method follows the methodology proposed in [130], for creating a decent dictionary for housing sentiment analysis, but differs in that we employ lexical knowledge learned by machine learning algorithms to automate the generation process. The second method adopts a more general classifier trained on social media sources to calculated the news sentiments. Subsequently, all estimated scores

of individual news articles are concatenated into a time series on a daily basis, also known as the sentiment index over the collected news articles. In both cases, if multiple news articles are found published in one day, their scores are averaged.

Housing-specific Sentiment Analysis

Soo [130] applies double expansion to a small list of words to generate the dictionary of positive and negative word candidates. In practice, this requires tremendous feature engineering and thus is time-consuming and laborious.

Here, we discuss our optimizations for each step to keep all possible manual effort to a minimum while retaining the usefulness of the generated dictionary. In essence, lexical knowledge learned by algorithms from large-scale linguistic resources is utilized to supervise the generation process.

Step 1. The dictionary starts with a couple of seed words strongly indicating an up/down trend. Following the convention, we initialize the positive word list with "Increase" and "Rise", and the negative one with "Decrease" and "Fall", provided by *Harvard IV-4 Psychological Dictionary* [3].

Step 2. The first expansion focuses on synonyms for the seed words. One of Harvard IV-4's functions is to return entities similar to a given term, which is used to query results close in meaning to each seed word. Consequently, we collect 111 and 25 candidates for "Increase" and "Rise", and 82 and 42 ones for "Decrease" and "Fall", respectively. However, as argued by previous literature [44, 83, 140], the psychologically related words do not necessarily hold in a semantic sense. In this work, rather than scrutinizing the results, we employ the Google News embedding model [4] to estimate the semantic similarity between the candidates and seed words. Basically, the model maps words into the same vector space and eval-

[3] `http://www.wjh.harvard.edu/~inquirer/Rise.html`

[4] `https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing`

uates how one word is similar to another based on the cosine similarity between two vectors. Trained on three million words and phrases, the dense vector representations [103] can generally reflect semantics and syntax of common words. By sorting the candidates based on their similarity values in descending order, we provide an automatic yet more objective approach to separate irrelevant terms from the synonym candidates.

Step 3. The second expansion fetches additional synonyms for those found in Harvard IV-4 from external lexical resources. In [98, 130], the authors manually refer to *Rogets 21st Century Thesaurus* for dictionary synonyms. Such method has an obvious drawback since it suffers from huge cost of labor and fails to determine whether the retrieved terms are sufficiently relevant to the seed words in terms of semantic relatedness. Our solution towards this predicament is to utilize WordNet[5], a large lexical database that groups four common types of English (nouns, verbs, adjectives and adverbs) into sets of cognitive synonyms (synsets), each expressing a distinct concept. Specifically, we search in synsets of a given term for matching candidates, and test the semantic similarity between each candidate and the term via word embedding. As before, candidates of high similarity are merged into the synonym list.

Step 4. Finally, [130] diversifies the generated dictionary by introducing inflections such as tenses and singular-plural relations of each word into both the positive and negative list. This move surely enlarges the scope of sentiment detection, but the price can be painfully expensive as well due to the irregularity of some nouns and verbs, let alone other various modifications expressing different grammatical categories. To keep the idea effective and bypass the pain, we instead handle the inflection problem in the reverse way, by reducing all news article words back to their base forms prior to sentiment analysis.

[5] `https://wordnet.princeton.edu/download`

It is noteworthy that our method is generalizable and can be extended to a wide range of research fields to aid building sentiment dictionaries given a set of seed words.

Towards General Sentiment Classification

The method in [130], or more generally methods where dictionaries are constructed upon specialized sources, fall into two limitations. First, they can solely capture sentiment in certain aspects. For example, a set of words stemming from the four seed words in Harvard IV-4 can track no more than the up/down trend among housing media. It is arguable that many other implicit factors can also contribute to the overall sentiment, like the word "War" will decrease the outcome but rarely is a synonym for "Fall". Second, matching news articles to a look-up list is too simple to handle real world scenarios. In practice, sentences hugely vary in expression and might contain out-of-vocabulary words, posing a serious challenge to traditional dictionary-based analysis.

This paper employs the VADER sentiment classifier [70] as a supplement to our method that automatically generates housing-specific dictionaries. The algorithm combines lexical features attuned to microblog-like contexts with five simple heuristics that adjust sentiment intensity, and is empirically proven domain-agnostic and computationally efficient. The adoption of VADER offers a new perspective on understanding the tone of news articles by capturing implicit sentimental factors and improving robustness for sentence variety.

4.5.3 Sentiment Scoring

Let D_{pos} and D_{neg} respectively be the positive and negative word collection that form a complete dictionary D, $N = \{w_1, w_2, \ldots, w_m\}$ a news article composed of

m discrete words. The first dictionary quantifies the sentiment of N by counting the occurrence of words in the news paper belonging to D_{pos} and D_{neg}:

$$\text{Sentiment}_1(N) = \frac{\sum_{w \in N}[\mathbb{1}_{D_{pos}}(w) - \mathbb{1}_{D_{neg}}(w)]}{|N|}, \tag{4.1}$$

where $\mathbb{1}$ is the indicator function, i.e., $\mathbb{1}_D(w)$ returns 1 if a word w exists in the dictionary D otherwise 0; $|N|$ is the total number of words in the news article.

The second dictionary modifies the equation by assigning a weight $t_d \in \mathbb{Z}$ to each entity $d \in D$, where positive words have $t_d > 0$ and $t_d < 0$ stands for negative words. The value of $|t_d|$ indicates the emotion strength d conveys; in fact, one can think of the intensity t_d in Equation (4.1) is constantly set to 1. The calculation of sentiment formally follows:

$$\text{Sentiment}_2(N) = \sum_{w \in N} \sum_{d \in D} t_d \times \mathbb{1}(w = d). \tag{4.2}$$

To normalize scores resulting from Equation (4.2), a score s is further transformed into $\hat{s} \in [1, -1]$ via the formula $\frac{s}{\sqrt{s^2+\alpha}}$, with $\alpha = 15$ as an experience value. With the dictionaries and scoring methods introduced above, we are able to perform sentiment analysis over all Glebe news articles collected earlier.

4.5.4 Comparing the MSI to the HPI using cross correlation

The housing price index is a traditional measure of price movement in the housing market. Therefore, we consider this to be an "objective measure" of housing price, and compare it with the MSI, a "subjective measure" of housing price.

We approach the technical problem of comparing the MSI and the HPI by using cross correlation, a standard technique from Signal Processing and Econometrics. The cross correlation is defined as a sliding inner product between signals/time

series data, and measures the similarity of two signals/time series as a function of the displacement of one relative to the other. Mathematically, this is formalised as follows.

Given two signals/time series data in (continuous) time, $u(t)$ and $v(t)$, the cross correlation with time offset τ is define as

$$\rho_\tau = \int_{-\infty}^{\infty} u(t)v(t+\tau)dt \tag{4.3}$$

If the signals/time series data are in discrete time, as in our specific case, the signals may be written as $u[m]$ and $v[m]$, and the cross correlation with discrete time lat n is defined as

$$\rho[n] = \sum_{m=-\infty}^{\infty} u[m]v[m+n] \tag{4.4}$$

If we set the time offset as 0, the cross correlation formula is reduced to the standard signal correlation formula (i.e. inner product) between two signals/time series data:

$$\rho[0] = \rho = \sum_{m=-\infty}^{\infty} u[m]v[m] \tag{4.5}$$

If we are calculating cross correlation for the actual data corresponding to signals/time series, obviously the real datasets do not stretch from time negative infinity to positive infinity, and thus the integrals/series reduce to finite integrals/sums. We apply the aforementioned formulas by padding with 0's all of the times outside the time frame where the time series data are collected.

In practice, the cross correlation is used in the following way. If we vary the time offset n and calculate the cross correlation for each value of n, as one signal "slides" past the other signal, we obtain two pieces of valuable information. The first is what is the *maximum value* of $\rho[n]$, $\rho_{\max}$, for all possible offsets n. The second is

at what value of n does this maximum value of $\rho[n]$ occur (let's call this $n_{\max}$). These tell us the following: how strong the correlation between the two signals are, allowing for time offset; whether one signal is a leading/lagging indicator of the other signal; and assuming a strong enough $\rho_{\max}$, the actual lead/lag time of the indicator, which would be $n_{\max}$. It should be noted that in practice the calculation of $\rho_{\max}$ and $n_{\max}$ is feasible, since as real signals have finite time frames, if n becomes sufficiently large in magnitude, to the point where the two signals have "slid past each other completely", then $\rho[n]$ will be zero regardless of the values of the signals. Therefore, there is only a finite window of values of n for which we need to compute the cross correlation.

4.6 Results and discussion

In this section, we present the *MSI* results and discussion. Firstly, we compare *MSI* results using two methods discussed in Section 4.5. Secondly, we compare *MSI* results with *HPI* from the Australian Bureau of Statistics (ABS).

4.6.1 Comparison of General-score *MSI* and House-specific-score *MSI*

We choosing Sydney, Australia for our suburb-level microscopic perspective study has two strong reasons. This is in part based on Shiller's reasoning in [126]. Firstly, Shiller has claimed that the tradition view that housing market behaves differently in regions is not completely true. It has shown national behaviour since mid-2000. This shows a sign that housing prices move beyond fundamental factors. It provides a ground to test the assumption of animal spirits. Secondly, Sydney is chosen by Shiller as one of the big glamorous cities that tend to have housing

bubbles. This provides us a very unique opportunity to discover the behaviour with the microscopic perspective. This may reveal why there is an international housing market cycles in big glamorous cities around the world.

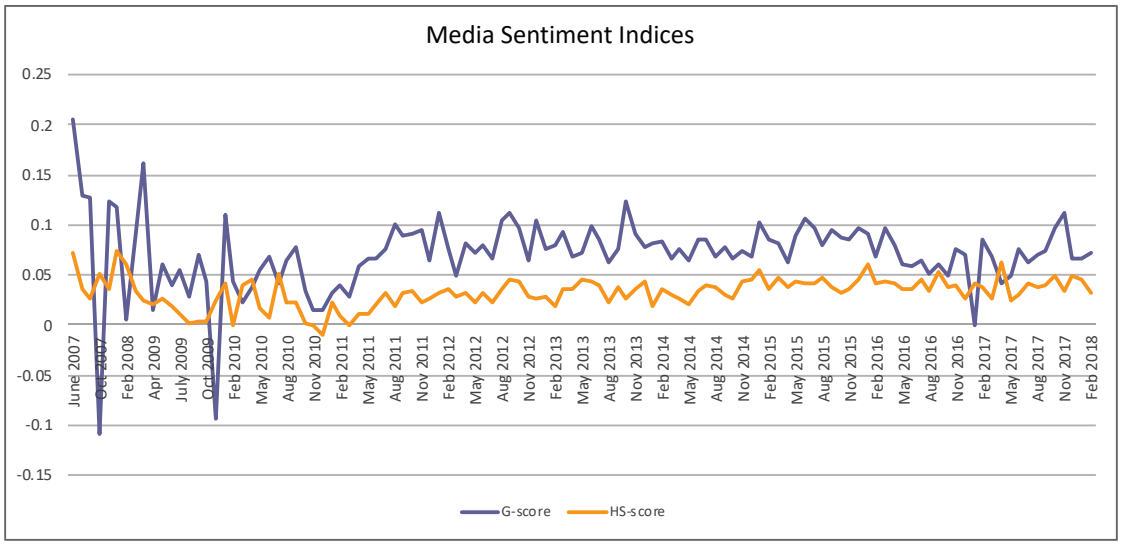


Figure 4.2: Media Sentiment Analysis with Two Methods

We calculate two *MSI* based on the methods described in Section 4.5. The score generated from Housing-specific sentiment calculation is called House-specific-score (HS-score); the score generated from more generalised sentiment classification is called General-score (G-score). To our knowledge, this is the first housing market sentiment analysis in Australia using local newspaper information.

Figure 4.2 compares the patterns of two sentiment scores using different text analysis, respectively HS-score and G-score. HS-score shows the sentiment score by calculating the positive and negative words used especially in housing market. G-score has a broader categories of positive and negative words for calculation.

In Figure 4.2, both HS-score and G-score have positive trends from 2007 to 2018. This is coherent with the continuous growth of *HPI* in Sydney. Compared with HS-score, G-score shows stronger positive sentiment. Both scores have a trough around 2010-2011, and follow a positive trend since 2011.

In detail, we pick a few representative points and read local news released around those time. First, there are two noticeable strong negative points of G-score (10/2007 and 11/2009). The local news is about gunshot crime happened in the inner west suburbs. This strong negative sentiment was not picked up by HS-score because of the specialised selection of sentiment words. Second, at 01/2017, HS-score showed a local max index (0.041), G-score just the opposite during the same period (-0.0006). Here we found a few articles had mixed views about the future housing market prediction, but more towards a negative view, especially expressed worries and concerns of the high percentage of people's income (nearly 50%) is used to pay the mortgage. Local news at these points support our argument that G-score shows stronger sentiments.

The reason we use two sentiment analysis methods is to find out if we could use a simpler method to calculate *MSI* without sacrificing the predictive power of local news for the housing market. HS-score has a painstaking preparation of the relevant positive and negative words in housing market. G-score is a more generalised sentiment analysis that doesn't require manual selection of words. If G-score can replace HS-score, *MSI* can be easily calculated in all suburbs. And in the future, the standardised sentiment scores will help researchers to compare the market behaviour analysed by different research institutes around the world. This will be discussed in the next section in more details.

The two lines in Figure 4.2 have similar trends, and the sentiment by G-score is more dramatic compared to HS-score, implying that the dictionary used by G-score is broader than HS-score. For example, G-score is consistently above 0.05 from 08/2011 to 05/2016. This is corresponding to the continuous significant growth of HPI in the same period (Sydney's HPI increased by 50% in five-years). On the contrary, HS-score peaked from early 2015 to early 2018. From 2008 to 2011,

G-score is more volatile compared to HS-score, reflecting the fluctuation of HPI.

In order to understand the difference between the two scores, we picked a few critical points, where either there is rapid change in the value of the sentiment indices or there is significant discrepancy between the values of the G-score and HS-score, and searched the original news and read them manually, in order to ascertain what may have caused those changes. We describe the close investigation of three critical points below.

The first case is the two sharp negative points (10/2007 and 11/2009) in G-score. These negative points were related to the local news of gunshot crime happened in the inner west suburbs. It seems that the crime brought negative impacts on the real estate market, especially in the west suburbs. G-score picked this negative news by a larger vocabulary calculation.

In the second case, we note a sharp maximum in the G-score value in March 2009, whereas HS-score is quite moderate. We searched the news and found out there were four real estate sales advertisements. The *ABS* housing price index turned from negative to positive 5 percent in the following three months. We could argue that the four advertisements shown up in the local news may be attributed to the increased confidence level of the sellers. It seems the housing market is more active in the booming period. To be conservative, we searched all the real estate local news, only found 55 out of 995 are real estate advertisements. We assume this is not significantly amplifying the G-score graph.

In the third case, we considered the time point at 01/2017. HS-score showed a local max index (0.041), G-score just the opposite during the same period (-0.0006). Here we found a few articles had mixed views about the future housing market prediction, but more towards a negative view, especially expressed worries and concerns of the high percentage of people's income (nearly 50%) is used to pay

the mortgage.

Finally, this result shows a strong evidence that using G-score to replace HS-score is highly doable. Following the discussion in the previous section, if we can use G-score for *MSI* calculation in the future, firstly it can save good effort on finding the relevant sentiment words; secondly, G-score can be used as a standard to allow comparison and discussion among different research results much easier. This will provide a great opportunity for us to underpin the behaviour of housing market from a microscopic perspective in the future.

4.6.2 Analysis of MSI vs HPI

General discussion

As discussed in Section 4.4.2, one of the questions investigated in this chapter is to what extent the *MSI* could be used as a proxy leading/lagging indicator for the *HPI*. Practically, this is done by comparing the time series data signals of the MSI and the HPI. The current literature tries to find the relationship between the sentiment analysis and the price level, in order to find out whether the sentiment analysis would be useful in predicting the stock market or housing market [130,140].

Regarding the choice of *HPI* data, since we have calculated the *MSI* for the Sydney suburb of Glebe, we should compare this to the *HPI* of Glebe. However, at the time when the authors were working on this paper, the AURIN dataset used in Chapter 3 was not yet available to the authors, and therefore it was not possible to calculate the localised *HPI* of Glebe. Instead, we used the *HPI* of Sydney available from the *Australian Bureau of Statistics (ABS)* as a proxy for the *HPI* of Glebe.

The *HPI* data from ABS used for this section is calculated at the date of exchange of contracts. The date is very important in the *HPI* calculation, because a normal sales contract has a 30, 60, or 90 days of settlement. As *ABS* uses date of

exchange of contracts as the counting date of selling a house, it reduces the delay for sales to be recognised, and therefore, it captures a more precise momentum of housing price changes. If *MSI* is leading in 3 months, we can confirm this is not due to the delay of dates of recognition of sales in *HPI*.

Figure 4.3 shows the Sydney HPI data from ABS plotted versus time, from 2007 to 2017.

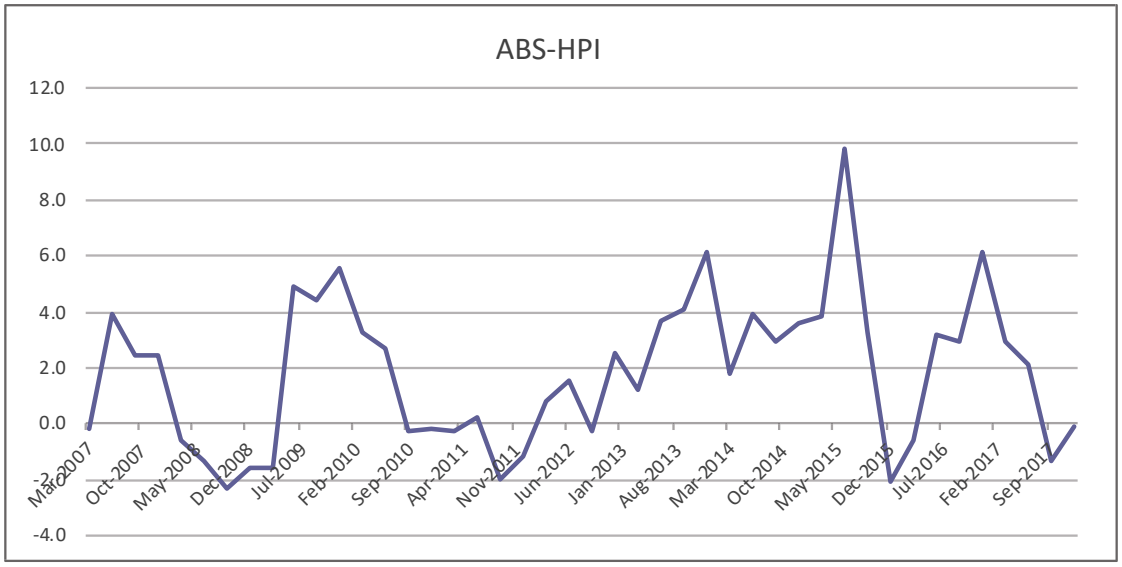


Figure 4.3: Sydney HPI from 2007 to 2017

Based on Figures 4.2 and 4.3, one cannot draw the conclusion that *MSI* would be a powerful predictor as shown from Soo's results in [130] . It is shown as more subdued, or random patterns. As mentioned in the previous section, we could match the turbulence period of *HPI* with up and downs in *MSI* before 2012. The constant positive *MSI* from 2012 to 2016 is parallel with the significant growth of *HPI*. But it is hard to prove *MSI* is a leading indicator for housing price at this stage.

The reasons for this difference are as follows. Firstly, Sydney's housing price had very few negative indices since 06/2002. The majority indices are above 2% seasonally, which means a compound growth rate is roughly 8% per annum. It

didn't experience a full boom and bust cycle in the last fifteen years. On the contrary, only 6 out of 34 major cities didn't have significant boom and burst cycles from 2000-Q1 to 2013-Q4 in Soo's study. Without a major turning point, it is hard to capture the cycle feature from the graph.

Soo's results showed that cities with milder changes in *HPI* had more subdued, or random patterns in *MSI*. On the contrary, Sydney's *HPI* rising was not mild, but significant. Logically, if *MSI* has a strong predictive power of *HPI*, the strong continuous increasing *HPI* in Sydney should have a companion of strong positive *MSI*.

Secondly, the difference may be due to the lack of sufficient data sets for testing. We may extend our experiments to combine larger areas with suburbs close by to collect more news articles, or do a city-level study with 8 major cities in Australia.

Thirdly, Soo [130] explained two competing reasons of the strong predictive power in the sentiment index. One is following the behavioural finance theory; people's over-exuberant beliefs drove the prices away from fundamentals. This idea has been described as "animal spirits" in [1]. The other reason suggests there may be local unobserved fundamentals which explained the synergy. As our results show no strong predictive power of sentiment index, the strong increasing power in Sydney's housing price in the recent 5 years needs to be explained in other aspects. We couldn't draw conclusions from Soo's suggested reasoning at this stage.

Using cross-correlation to gain additional insights into housing market trends

MSI and *HPI* offer indicators in the real estate market through very different data sets. In order to explore the relationship between *HPI* and *MSI* in a more precise way, we calculated cross-correlations between them, doing both HS-Score vs HPI and G-Score vs HPI. By computing the cross correlation between *MSI* and the *HPI*,

we obtain 1) the delay between the index signals at which the correlation coefficient is maximised and 2) the value of coefficient in this instance (refer to Section 4.5 for the details about this technique). This provides information on which of the two indicators is the leading/lagging indicator, as well as how well correlated the two indicator signals are. It turned out that the two cross-correlation results are very similar, so we only present one of the results in this paper: G-Score vs HPI.

From our calculations in the previous subsection, we obtained monthly G-Score *MSI* for a period between 2009 and 2015. As the *MSI* data is monthly and the *ABS-HPI* is calculated quarterly, we interpolated the *HPI* data by assigning the value of the *HPI* to all months of the quarter. For example, if the 2012 Q2 *HPI* has value x, we assigned x to April 2012, May 2012 and June 2012. We then performed the cross correlation, between the G-Score (March 2009 to Feb 2015, 72 data points), and the *ABS-HPI* data, spanning the same length of time, but offset by an amount from -83 months to +49 months (this choice was based on the *HPI* data we had available).

The result of the cross correlation is presented in Figure 4.4. A visual inspection of Figure 4.4 reveals that the maximum correlation coefficient occurs when the *HPI* signal is offset by 20 months after the Sentiment Analysis signal (the Sentiment Analysis signal starts in March 2009, the *HPI* signal starts in Nov 2010). This is coherent with Soo's results that housing media sentiment leads housing prices by nearly two years. We can see from Figure 4.4 that at the offset of maximum correlation, the correlation is 0.37. This shows a weak correlation between the indices.

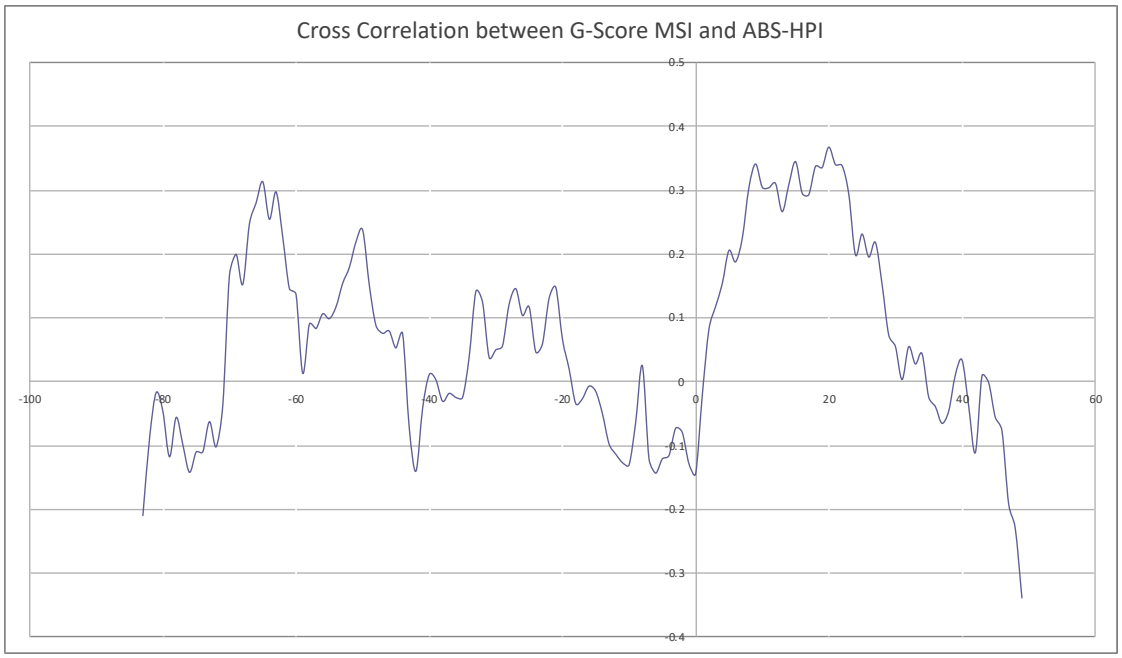


Figure 4.4: Cross correlation between G-Score and ABS-HPI

4.7 Conclusion and future work

As the nature of economic data has enlarged to textual, geographical, imaging, sentimental by the influence of big data revolution, the big challenge is how to organise, store, extract and analyse the high-dimensional data sets, and how to integrate with the traditional numerical data to work for the economic analysis. Housing market research is a good area to test this idea due to its nature of recent available high-dimensional data sets and unique difficulty of heterogeneous feature for pricing and evaluation.

The work in this paper has illustrated the housing market related data analysis that could be done in a microscopic level regarding real estate cycles. The method could be extended in several ways. For example, one may expand the variables and explore the deep learning methods for evaluation with high-dimensional data sets and continue explore new data sets (e.g., those with imaging and geographical feature). We will extend our investigation to other suburbs and do a comparison

study and generalise our findings. Furthermore, we would expand the variables and explore the deep learning methods for evaluation with high-dimensional data sets. At the suburb-level housing market study, we plan to design experiment or quasi-experiment for the subpopulation situation [41]. One of limitations of current work is that we only investigate Glebe of Sydney as our preliminary work, therefore, it is hard to run classification tree to discover the hidden patterns as [105]. This problem can be solved with a full study of suburbs in Sydney, or even cross the country. We may also use improved natural language processing algorithms which have a better *MSI* based on the sentence-level meaning processing.

CHAPTER 5

NEW BEHAVIOURAL BIG DATA METHODS FOR PREDICTING HOUSING PRICE

5.1 Chapter Abstract

Housing market price prediction is a big challenge. The 2008 global recession strongly showed that even the most sophisticated traditional economic models failed to foresee the crisis. New developments of behavioural economic theory indicate that the information from micro-level's decision making will bring new solution to the age-old problem of economic forecasting. Additionally, the information revolution and big data methods have provided a new lens to study economic problems apart from traditional methodologies.

This research provides the theoretical link between irrationality and big data methods. Empirically, big data methods will be used in forecasting the housing market cycle in Australia. Specifically, Google trends is included as a new variable in a time series auto-regression model to forecast housing market cycles.[1]

5.2 Introduction

Housing market prediction is one of the biggest challenges in the economic cycle research. The global financial and economic crisis (GFEC) that happened in 2008 is strong evidence to show that most economists failed to predict this real estate cycle. US $13 trillion of property wealth was wiped off in the US real estate market after the crisis. Noticeably Case and Shiller [27] raised the question of housing bubble before the crisis happened, but it didn't get sufficient attention

[1]This chapter is based on the following article:
Jiaying Kou and Yashar Gedik. "New Behavioural Big Data Methods for Predicting Housing Price." EAI Endorsed Transactions on Scalable Information Systems 6, no. 21 (2019).

both from the academic and political world. If the economic world had strong and reliable forecasting system to raise the alarm long before the crisis happened, the great loss to the world could have been prevented.

Why housing price prediction is such a difficult task? The assumption is that the key element to stop the forecasting model to be functional is the omission of the valuable information from the micro-level human decision-making behaviour, especially the irrational behaviours. The economic forecasting model cannot be improved by a technically more and more sophisticated models, without investigation of newly available high-dimensional data [90, 128]. Business cycle forecasting or economic growth forecasting is still one of the most difficult questions in the economic field. Just like the Nobel laureate Murray Gell-Mann once said, "Think how hard physics would be if particles could think." This described exactly the challenge of forecasting housing market performance.

The question is: could we capture other information besides the traditional economic data to reach a better prediction of housing price? Specifically, on-line information has become a new rich mining to understand human behaviour. Can we capture new information from on-line mining to increase the economic data dimension for a better and quicker prediction? This challenge involves finding useful behavioural information, identifying new variables, using appropriate big data methods to improve prediction power. In summary, the challenge is two-fold:

(i) will the adoption of on-line data methods improve the understanding of housing market forecasting from economic perspective? And how?

(ii) what kind of information processing methods would suit for economic forecasting problem? What kind of adaptation process needs to be done?

Luckily, there have been some fruitful results produced from both outside of economic world and within economic world in particular forecasting application.

For example, information embedded in the Twitter stream can track the H1N1 disease levels in a timely manner and this is extremely important as it shows the potential to control the flu spreading in a faster and more efficient way [127]. Another example of using Twitter stream sentiment information to forecast stock market movement shows how to apply the behavioural economics theory to capture the public mood to support the prediction [15]. This is a very good example to show the interdependence of economic theory and computer science methods to foster a multi-disciplinary fruit. In summary, these examples shed lights on the possibility of improving forecasting ability in housing market movement through new information from big data.

Stock market prediction has a longer research history compared to the housing market, due to the available stock price data and significant amount of wealth involved. Academic finance has passed the excitement of efficient markets theory based on rational expectation. Excess volatility has continuously been found which cannot be explained by the efficient market hypothesis. People tended to look at this problem from a different angle and introduced behavioural finance. Psychologists found people consistently making irrational decisions in experimental environment [125].

The newly development of behavioural economics can provide theoretical guidance of housing market prediction. The challenge is how to apply the theory into the empirical discovery stage. First we introduce some relevant behavioural economic concepts.

"Winner's curse" is a behavioural economic term to describe the situation that people tend to pay far above the expected price in a competitive pricing system. If people pay much higher than the expected value, this difference should gradually disappear because rational people will learn from their mistakes. If such anomaly

consistently appears in the human decision making empirically, this is considered against the traditional economic assumption of rationality [141]. It is natural to assume such behaviour would happen in the housing market. When people are in the street auction, the winner of the auction pays much higher than the intrinsic value of the house.

There are other theories can also be applied in the housing market environment, such as prospect theory [77] , the endowment effect [75], or anchoring theory [146].

All these theories study the human decision behaviour when they are apart from the rational status. These theories can show evidence of people's irrational decision-making behaviour in a micro experimental environment. If these individual behaviours aggregate in a social context, it is reasonable to assume the housing price also reflects people's overreact and underreact behaviour. These behaviours are hard to be captured from the traditional economic data collection and model building.

This shows the gap between the macro-level economic forecasting modelling and the available established micro-level behavioural theories. In other words, how to use the advantages of the findings in the micro behavioural theories in the macro behavioural predictions, is a research problem that needs to be solved. The current macro-economic cycle study is based on the rational assumption, which means that the irrational behaviour is not considered in the housing price forecasting models. Even though lots of experimental evidence can be found in the micro economics, this cannot be transferred directly into the macro economics context.

The recent development in the big data analysis in business and e-commerce shed lights on the aggregation power of millions and billions of individual on-line information for a macro-understanding. The hidden information in the on-line search clicks, business transactions, on-line news, chats and Tweets is studied

by researchers. For example, since Google has freely opened the Google Trend Index database in 2015, researchers have started to use these search engine data to improve economic indicator predictions [29]. Wu and Brynjolfsson [153] did a pioneering study of housing market cycle prediction in the US and demonstrated the forecasting had a significant improvement with the search engine indicator input.

This paper aims to improve the economic forecasting ability by using the new lens of online big data method, as a way to accumulate millions and billions of individual behavioural decision-making information. It is now possible to produce a more accurate, timely, low cost forecasting of housing market cycles with the support of new technology—big data. The detailed research aims are listed below:

(i) Investigate the forecasting ability of Google Trend Index in the Australian housing market.

(ii) Introduce a new variable—auction clearance rate in the performance model to simulate Australian housing market better. Test the interpretation ability of this variable.

(iii) Test how Google Trend Index interact with the auction clearance rate.

(iv) Develop the forecasting model based on other big data method, such as indicator generated from text stream analysis from online news, Twitter and Facebook.

This will be the pioneer study in Australia to use big data methods to forecast housing market cycles. Housing market performance indicators which have better interpretation for Australian housing market will be developed. Additionally, text stream data will be analysed in the housing market forecasting model. There is a gap to study housing market behaviour using text from Facebook or Twitter.

Finally, this research provides an original illustration of the theoretical connection between micro-level behavioural economic theories and macro-level business cycle models.

A more accurate forecasting model for the Australian housing market will benefit investors and policy makers to understand the market behaviour. The prediction of housing market cycle is extremely important as the recent global financial and economic crisis showed us the power of the housing market to the economy as a whole. This pioneering study will demonstrate how to implement the big data methods in the economic forecasting study. Hopefully more research can follow and improve the forecasting ability significantly.

The uniqueness of Australian housing market provides a special opportunity to study the human decision-making behaviour under a street auction environment and how such behaviour will influence the housing market as a whole. Australian housing market has been booming continuously since early 2000, only had short slow downs around 2010 and 2018. Financial institutions and government are closely monitoring the current downturn. It is a good timing to test the big data method forecasting ability. This improved prediction will serve as a watchdog of the potential risk or the notice of housing bubbles and help to provide more time for policy reaction for financial institutions and government policy makers. A potential crisis may be prevented with a timely solution.

5.3 Related Work

As this research is a cross-disciplinary research joining the economic business cycle forecasting with the big data forecasting methods, literature will be introduced in both areas.

5.3.1 Existing knowledge of macroeconomic housing market forecasting

The study of housing market is based on the market efficiency hypothesis (EMH). The concept of efficiency came from information theory. The level of efficiency of the stock market is defined on the degree of all available information being reflected in the current stock market prices. A highly efficient market absorbs all available information into its stock prices [99]. Millions and billions of micro-transactions in trading time form stock prices. Institutions and individual investors make trading decisions based on available information and analysis. The same idea works in the housing market. Each of the micro selling and buying decisions forms the housing market prices.

Researchers started to discover overreactions, anomalies in the stock market behaviour that questioning the rationality of stock investors [32, 72]. Interestingly, not many studies have been done on the housing market efficiency behaviour in the literature, comparing with the enormous work been done in the stock market. One of the earliest empirical tests of housing market efficiency for single-family home suggested the housing market is not efficient [25], which means the current housing market performance contains valuable information for the future prediction. Pollakowski and Ray [114] continued the study of housing market efficiency in depth of the interaction of different geographic locations.

5.3.2 Existing knowledge of microeconomic behavioral theories

Besides testing the market efficiency directly, many researchers turned to psychology and co-developed theories across economics and psychology disciplinary. New

theories developed under empirical psychological experiments. These theories are guiding current behavioural research.

A recent World Bank report [9] on human decision making has developed three principles to direct behavioural studies. They are named as "thinking automatically", "thinking socially" and "thinking with mental models". These models are fundamental to the understanding of this area, and literature relating to them are reviewed in the subsections below.

Thinking automatically

People tend to make their decisions automatically, without careful thinking. This is also called System 1 [47, 74]. Making decisions with careful thinking is called thinking deliberatively, also called System 2. People tend to use very limited information to make analysis and make decisions based on very few alternatives. These decisions are easily biased based on the discovery of behavioural theories, such as prospective theory [77], heuristics in judgment and decision-making [76], anchoring and adjustment [146], intertemporal choice [97], the endowment effect [75]. These are the theoretical assumptions of how people make decisions in housing market. Applications have been done in the housing market, such as [108].

Thinking socially

People are highly influenced by the society they live in. Their personal beliefs, prestige, desires, the sense of belongings, motivations are highly influenced by the social norms, social preferences [144]. This will naturally influence the big decisions such as buying a house. Are we buying what we like or are we buying what other people may think of us? Would buy house in a rich suburb be an important factor of my decision-making process? Will people have peer pressure if their friends are buying houses? Are they searching locations where their friends and relatives live?

These will be examples of questions to be designed if a survey is drafted.

Thinking with mental models

The term "mental models" means that people tend to make decisions under the acceptance of the "agreed" beliefs among people from the same community. "Culture", defined by the anthropologists and sociologists, is "the collection of mental models" [37, 136].

Australian housing market is an interesting research field to find out the influence of cultures. People from different ethnic groups contain their own culture in what it means of "home", or "house". This can influence their "tenure choice", the choice between renting and owning a property [147], when to buy houses, family structure and house preferences, etc. There may be decision making differentiate between people living in Australia for generations and new migrants. Even the housing locations can be influenced by ethnic backgrounds.

Online social media, such Facebook and Twitter may provide tremendous information about what people are thinking when they buy or sell houses, instead of the traditional survey approach.

5.3.3 The gap between behavioral theories and macro-level forecasting models, and big data

Historically, there has been a tension between traditional macroeconomic theories of forecasting and those theories arising from behavioural economics. Assumptions in traditional macroeconomics such as the EMH, introduced in Section 3.1, and rational economic actors ("homo economicus"), do not resolve well with the models used in behavioural economic analysis. In the 20th century, there is a gap between these two views of economics without being able to go much further.

Since the 2008 Global Financial and Economic Crisis, a few economists started to work on the improvement of business cycle model by introducing more realistic factors. For example, Kiyotaki [81] has brought financial credit constraints in the interaction of firms and households. He argued that the financial credit constraints can create the economic fluctuation in reality. But these are the first attempts in the theoretical basis, there is still a gap to transfer these new models in the empirical testing models. In addition, these theoretical modellings are in the beginning stage of inputting more realistic factors, with very limited attempts and development. It is far from complete, so these theories have very limited impact on the real forecasting modelling.

However, as we shall describe in the next section, the advent of social data from Internet sources such as Google Trends and Twitter, provides for the first time an empirical basis upon which we may compare predictions arising from these two different views of economics, in particular with respect to the housing market. In section 4, we provide the theoretical framework and methodology for using big data to analyse the housing market. Such type of research is still at its early stages globally, with only a small number of studies being done . There is plenty of room for original research in this area, especially for the housing market in Australia.

5.4 Methodology and Theoretical Framework

5.4.1 Current literature of big data methods for housing market forecasting

A few decades ago, both Leontief and Simon pointed out that a better business cycle forecasting cannot be reached by applying more and more sophisticated econo-

metric models, the solution should come from a higher level of new information input, which means a painful empirical micro-level data collection is the way to lead to a higher solution [90, 128]. This emphasizes the necessity to study individual decision making, which contributes to the business cycle.

Leontief predicted that the future economic research will be based on surveys in larger scales and in multiple dimensions. Case et al. [28] have done lots of pioneering work in this area. They have kept sending thousands of surveys to homebuyers yearly for three decades and try to understand their economic reasoning for home buying behaviours. As we are coming to the big data era, some researchers have realized that the online big data is a very powerful new survey tool. Leontief's prediction is gradually realized.

Some pioneer studies used google trend index as an indicator to do the economic indicators' forecasting. In [153], the authors studied the real estate cycle in the US. The work in [29] did a similar research in the forecasting of other economic indicators, such as automobile sales, unemployment claims in the US. But the relationship between the behavioural economic theories and the big data methods in forecasting is not discussed in both papers. This discussion is essential for the theoretical framework, because this is to explain why big data method should be effective in the forecasting modelling. This is another gap in the theoretical framework area when analysing the current literature.

Other researchers worked on the forecasting ability of online news, Twitter and Facebook text streams analysis. Soo [130] studied 34 cities news about housing market sentiment and found out that this information has a predictive power about the future housing prices. Sun et al. [133] combined the information from online news articles and search engine to predict the real estate cycle in China.

There hasn't been any literature on using Twitter or Facebook text streams

of public sentiment to predict the housing market cycle. This is another gap. One of the relevant inspirational research is done by [127]. Their research demonstrated the forecasting ability of public sentiment from Twitter text streams on the H1N1 disease levels. Social network analysis methods in computer science provides advanced theoretical framework, algorithms that can be adapted in the housing market studies using Twitter or Facebook text streams [113, 150].

In the following sections, methodologies of Google Trend Index, online news and text streams from Facebook and Twitter will be discussed.

5.4.2 Forecasting model using Google Trend Index

Background of Google Trend Index

In 2015, Google company made real time Google Trends data available publicly. There are trillions of search clicks every year through Google search engine. This data provides a unique opportunity to find out people's searching interest globally or granulated down to the city level in a specific period.

Google Trends index is a normalized data to look at the search interest about a topic relatively comparing to the total clicks in the same place and time. After selecting the time period and location, the highest search interest is counted as 100. Other search interests in the same period will adjust to compare with the maximum figure and generate a ratio.

Selection of housing market performance indicators

The purpose of this model is to test the interpretation power of Google trend index in forecasting the future housing market performance.

There are a few indicators to evaluate the housing market performance, such as sales quantity, auction clearance rate, housing price index (HPI). Sales Quantity

and HPI were used by [153] for housing market indicators. Auction Clearance Rate (ACR) is an original indicator proposed in this paper.

ACR is included as an indicator because the clearance rate indicates the percentage of sellers are satisfied with the auction price. If a house auction is passed in, it means the auction price hasn't reached the vendor's reserve price.

Street auction is quite unique and popular in Australia, which is rare to see in other countries. Large number of houses are sold in the auction, especially the high value properties in Australia. Both in UK and US, the auction of houses is not the main selling method. And the auction is not held in front of the selling houses, but gathered in an auction room, or in front of some court areas.

The assumption is that people in Australia have more chance to observe or experience a house auction than people in most other countries. Based on the endowment effect, we assume people are more attached to the property they are willing to bid if they are standing in front of the property for auction. The feeling of losing an auction is more real and stronger comparing to quietly sitting in an auction room. Therefore, the clearance rate provides a unique opportunity to understand the behaviour of Australian housing market.

Google Trend Index forecasting model design

The model design follows the model structure described in [153]. A simple seasonal autoregressive time series model is chosen to test the interpretation power of search indices for the housing market performance. The major difference from the model in [153] is the Auction Clearance Rate (ACR) indicator introduced above.

The basic auto-regression home sales quantity model is shown in Equation 5.1. Then Google trend data –Search Frequency –is introduced into the equation to find out if Google Trend Index improves the prediction of the current Home Sales figure. Specifically, the current and one lag behind search frequencies are added in

the regression equation to form Equation 5.2. All the performance indicators data are collected in a seasonal basis.

$$\begin{aligned} HomeSales_{it} = \alpha &+ \beta_1 HomeSales_{i,t-1} + \beta_2 HPI_{i,t-1} \\ &+ \beta_3 ACR_{i,t-1} + \beta_4 Population_{it} \\ &+ \sum S_i + \sum R_j + \sum T_t + \varepsilon_{it} \end{aligned} \tag{5.1}$$

$$\begin{aligned} HomeSales_{it} = \alpha &+ \beta_1 HomeSales_{i,t-1} + \beta_2 HPI_{i,t-1} \\ &+ \beta_3 ACR_{i,t-1} + \beta_4 SearchFreq_{it} \\ &+ \beta_5 SearchFreq_{i,t-1} + \beta_6 Population_{it} \\ &+ \sum S_i + \sum R_j + \sum T_t + \varepsilon_{it} \end{aligned} \tag{5.2}$$

The logic to choose a simple linear regression structure

First, the main focus of this paper is to test the interpretation power of Google Trend Index in Australia. Therefore, we use a simple linear regression to distinguish this variable.

Second, based on the studies in [25] and [153], past housing market indicators have the power to predict the future. Therefore, a simple auto regression is appropriate.

Third, [153] have demonstrated the interpretation power of search indices using US data. Their results were even better than the predictions from the National Association of Realtors.

Fourth, this paper also aims to test if Auction Clearance Rate is a good indicator for housing market performance.

Fifth, Wu and Brynjolfsson [153] have also found that simple linear regression had even better results than more sophisticated nonlinear models. This finding is coherent with the theoretical assumptions made in the previous section, that a

higher level of information input by adding the search information is the solution of current forecasting problems, not how sophisticated the econometric model is.

Future prediction model

The next step is to test the forecasting ability of the search data for the future home sales quantity. We follow the model in [153], the current, one-period and two-period lags search frequencies in the model. We could get the current search frequency for the future prediction. Two-period lags cover nine months of searching period prior house purchasing. We assume that this is a practical maximum searching period. The difference from the model in [153] is that we introduce the auction clearance rate into the model (Equation 5.3).

$$\begin{aligned} HomeSales_{it} = \alpha &+ \beta_1 HomeSales_{i,t-1} + \beta_2 HPI_{i,t-1} \\ &+ \beta_3 ACR_{i,t-1} + \beta_4 SearchFreq_{it} \\ &+ \beta_5 SearchFreq_{i,t-1} \\ &+ \beta_6 SearchFreq_{i,t-2} + \beta_7 Population_{it} \\ &+ \sum S_i + \sum R_j + \sum T_t + \varepsilon_{it} \end{aligned} \tag{5.3}$$

Same modelling process with other performance indicators

Following the same structure, we can predict the current and future HPI. Then, we can also predict the current and future Auction Clearance Rate (ACR). Examples are shown in Equations 5.4 and 5.5.

$$\begin{aligned} HPI_{it} = \alpha &+ \beta_1 HPI_{i,t-1} + \beta_2 Homesales_{i,t-1} \\ &+ \beta_3 ACR_{i,t-1} + \beta_4 SearchFreq_{it} \\ &+ \beta_5 SearchFreq_{i,t-1} + \beta_6 Population_{it} \\ &+ \sum S_i + \sum R_j + \sum T_t + \varepsilon_{it} \end{aligned} \tag{5.4}$$

$$
\begin{aligned}
ACR_{it} = \alpha &+ \beta_1 ACR_{i,t-1} + \beta_2 Homesales_{i,t-1} \\
&+ \beta_3 HPI_{i,t-1} + \beta_4 SearchFreq_{it} \\
&+ \beta_5 SearchFreq_{i,t-1} + \beta_6 Population_{it} \\
&+ \sum S_i + \sum R_j + \sum T_t + \varepsilon_{it}
\end{aligned}
\tag{5.5}
$$

5.4.3 Forecasting model with online news

The key element to predict housing market cycle with the information from online news is to build the relationship through "sentiment".

Barberis et al. [10] studied stock market investors sentiment on the stock performance news, such as earnings announcements. And he generalized the sentiment as "overreaction" and "underreaction" with empirical evidence.

Tetlock [139] also quantified the interactions between Wall Street Journal column and the stock market performance. A pessimistic view can drive down the market prices as an overreaction, which will follow a reversion to fundamentals.

With the development in the stock market sentiment study, the housing market study follows the trend. Soo [130] did a pioneer study of American housing market sentiment study by analysing the newspaper tones as "positive" or "negative".

The textual analysis is used to quantify the tone of financial documents. The standard dictionary-based method is used to count the raw frequency of positive and negative words in a text. Soo [130] prepared a housing dictionary and presented the calculation of the overall tone of housing market news sentiment by:

$$
S = \frac{\#pos - \#neg}{\#totalwords} \tag{5.6}
$$

Kou et al. [84] followed this study and calculated the media sentiment index with two different dictionary methods at a suburb-level in Australia. This approach can be continued by extending the suburb study to a macro-level.

5.4.4 Forecasting with text streams from Facebook and Twitter

Articles directly using Facebook or Twitter text streams to predict housing market cycles are not found in the literature. But there are quite a few research papers using the idea of mood or sentiment analysis to predict stock market behaviour [15, 104, 160]. We could reasonably assume that the nature of the prediction algorithm for housing market will be similar to the stock market.

How to find emerging topics is a challenge in social media analysis. Early detection can improve the understanding of people's behaviour towards market movement. Novel tracking method can be found in [94]. Big data analysis could be another challenge. Some technique and solutions include [92, 159].

We assume that the selection of Twitter feeds need to address the location difference, unlike the stock market prediction, because the housing market cycle shows a strong trend difference world widely. But Shiller [126] has shown the big glamorous cities experienced massive boom within similar timeframe (1999-2014). Australian cities, Sydney and Melbourne are recognised as the glamorous cities and following the global boom trends. Therefore, we may test the Tweets in Australia and world-wide for the location hypothesis.

5.4.5 Data source

Google Trend Index is a free publicly available database. The index figure can be downloaded at state level in Australia with monthly interval. Most other housing market indicator data can be found through AURIN data (The Australian Urban Research Infrastructure Network). Online news and samples of text streams from Twitter and Facebook can be captured by Crawler Programs.

The auto-regression time series forecasting model can be run through R or other statistical software. We will compare the results using different variables and find out the forecasting ability and compare the forecasting results of different big data methods.

Data privacy issue also needs to be considered when using text streams from Twitter or Facebook. Certain ways of privacy protection can be applied to avoid personally identifiable information [151, 152].

5.5 Conclusion

This study uses the latest available technology, data and research method to analyse and forecast the Australian housing market. It will provide new insights in economic cycles and in the relationship between behavioural microeconomics and traditional macroeconomics. These methods will provide a significant improvement in the quality, cost-effectiveness and timeliness of Australian housing market forecasts, and become a valuable tool for investors, bankers and policy makers. Predictions from these methods may prove to be the earliest predictors of economic downturn and upturns. Hope this research may also inspire other research to develop more sophisticated methods of using big data in economics.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

Will artificial intelligence make an impact on economic research, and how? This is the question raised in the first chapter. This thesis endeavours to explore empirically on the problem of understanding housing price, in order to seek some strong evidences to support the vision and arguments of multidisciplinary direction of joining machine learning and econometrics, which was discussed deeply in the first chapter.

As the nature of economic data has enlarged to textual, geographical, imaging, sentimental, and behavioural, by the influence of big data revolution, the first big challenge is how to find unique perspective for a valuable economic questions to be enquired with the technologies of artificial intelligence. The second big challenge is how to organise, store, extract, and analyse the high-dimensional data sets, and how to integrate with the traditional numerical data to work for the economic analysis. Housing market research is a good area to test these ideas, due to its nature of being empirical – which means real economic problems can be experimented upon by leveraging the recent available high-dimensional data sets – as well as its inherent and unique difficulty of having heterogeneous features for pricing and evaluation. In addition, one unique challenge is to coordinate traditional and non-traditional data sets by their spatial associations. Due to the strong data processing power of the newly developed technology and emerging big data, we would assume economic research questions that could not be investigated in the past – due to either lack of data, or lack of computing and algorithm power – now have a good opportunity to be explored.

The key challenge is still within the economic domain, to find the economic enquiry that can generate impact on the future. Some of the themes follow the direction of exploring the newly available non-traditional economic data, such as textual, geographical, imaging, and behavioural data, to bring new information for housing market problems. Other themes are embedded in the exploration of the economic and social network in the context of geographic location, where the focus is to enquire on the social network structure and patterns and their impact on resource allocation, including decision-making for choosing the physical locations of homes, businesses, industrial clusters, and research precincts.

The theme of this thesis is to apply data science technology to solve housing market research challenges. Within this thesis, there are a few variations explored. These variations are focused on applying new types of data, specifically newspaper textual data, point level spatial data, behavioural google search data, behavioural ranking data for schools and restaurants, and also applying machine learning algorithms, such as gradient boosting algorithms, K-nearest neighbour algorithms, natural language processing sentiment analysis, into the understanding of housing market problems. By exploring new types of data and algorithms used for housing market research questions, the social and economic network is also investigated. Specifically, the interest is on identifying economic clusters regionally and locally and its interaction with the decisions of dwelling locations, business locations and housing price.

Specifically, in Chapter 3, the novelty of this study is focused on establishing point level housing appraisal model, by utilising location-based social network (LBSN) point level spatial economic data. In detail, 13 categories of open street map point level data are used to calculate the convenience level of living within 1km walking distance of a house. Behavioural ranking data is scripted for schools and

restaurants to understand the quality index of the service in the neighbourhood. By leveraging these non-traditional spatial data sets, the aim is to study how factors beyond neighbourhood impact housing values. Specifically, regional economic clusters are established as the significant source of impact beyond neighbourhood. We presented our housing price appraisal model that combined housing attributes, neighbourhood characteristics, and demographic factors. Our model using the XGBoost algorithm has reached 0.88 in R^2, showing the significant impact of regional economic clusters. This study shows the potential of utilising spatial data with high granularity to build precise appraisal models. In addition, the conclusion of this study provides useful recommendations for decision making in urban planning and policy studies.

In Chapter 4 and Chapter 5, the novelty of these studies are to utilise data science methods to collect micro-level economic behavioural data sets and integrate these new data sets into the new economic prediction modellings. In summary, the micro-level human behavioural data sets are used for the macro level of housing price trend prediction. The gap between the traditional macro housing market modelling and new development of irrationality of micro-economic theories is filled, by collecting and analysing economic behavioural data, such as real estate opinions in local newspaper articles, people's searching behaviour captured by Google Trend Index.

In Chapter 4, micro-level behavioural data is collected and analysed for understanding macro-level housing market behaviour. Sentiment analysis is adopted to examine local newspaper articles discussing real estate at a suburb level in inner-west Sydney, Australia. The media sentiment index is calculated by using two different methods, and compare them with each other and the housing price index. The use of media sentiment index can serve as a finer-grained guiding tool to facili-

tate decision making for home buyers, investors, researchers and policy makers. In Chapter 5, the discussion is around the gap that new developments of behavioural economic theory indicate that the information from micro-level's decision making can bring new solution to the age-old problem of economic forecasting. It provides the theoretical link between irrationality and big data methods. Specifically, Google Trend Index is included as a new variable in a time series auto-regression model to forecast housing market cycles.

In summary, to our knowledge, this is a series of pioneering empirical economic studies in Australia to apply data science methods in understanding housing market problems. This thesis uses the latest available data science technology, new types of data, machine learning algorithms, and data processing methods to analyse and forecast Australian housing market. The challenge and innovation is to join data science and econometrics together. It provide new insights in economic cycles and in the relationship between behavioural microeconomics and traditional macroeconomics. It contributes more powerful housing appraisal models. These methods provide a significant improvement in the quality, cost-effectiveness and timeliness of Australian housing market forecasts, and become a valuable tool for investors, bankers and policy makers. Prediction from these methods claims to be early predictors for economic downturn and upturns. Hope this research may inspire other research, to develop more sophisticated methods of using big data in economics.

6.2 Future Work

Economic problems by the nature are complex, as discussed in the introduction section. Therefore, increasing the dimension of investigation is one of the solutions to discover new variables, unknown relationships and to improve predictions. This

goal can be achieved by continuously leveraging new available data, enlarging the data sets and exploring machine learning and deep learning algorithms, combined with insights and knowledge from real estate economics, other related social science fields, and urban planning.

Follow this logic, in order to improve the housing price prediction models and discover latent variables, other types of data can be explored, more relevant algorithms can be applied. For example, satellite data and image processing algorithms can be used for understanding features associated with housing price and the location based social network (LBSN). This can generate cross comparison of different cities from different countries. Current research focuses on one city or multiple cities from one country due to geographic proximity. This may help to explain the boom and bust of housing prices of big cities across the world. Another example is to explore the location based social network by collecting data sets from different angles and continue to understand the clustering effect to achieve to goals of affordable living and building better community.

The other approach for understanding LBSN and its connection with housing price is to explore graph theory in the building of LBSN. The understanding of network pattern and behaviours can be reached into next levels. This may integrate the geographic graph of routes and traffic with knowledge graph of economic and social activity clusters as different layers on top of the geographic graph.

One immediate following work can focus on the further development of the feature: *Regional Economic Clusters*. For example, systematically discuss the connection with the theoretical development of economic clusters in the economic domain, and further expand our model by adding features of office work space, local business quantities and medical clusters.

One extension of studying the economic clusters is to investigate the agglom-

eration effects at a micro-level. Most of agglomeration study focuses on the big node at country or global level, such as study of Tokyo, New York, or Silicon valley. With our granulated data sets and multiple layers of data from different fields, such as economy, housing, demographic characteristics, traffic, environmental factors, the study of agglomeration can be extended to the suburb level. This may bring insights about clusters emerging, and the interaction between business-to-business and business-to-living at a micro-level, plus policy impact for future urban planning and community building.

In order to improve the trend prediction, following the sentiment analysis method used in [84], more work can be done by collecting textual data across different cities to generate a comprehensive prediction analysis, or combining other type of data, such as Google Search Index.

Other challenges include bringing dimension of time for housing price appraisal. With rich housing price data, we could explore features that have strong influence of long term capital gain by applying deep learning methods, such as long short-term memory (LSTM), and dynamic network analysis.

Other directions can involve collaborations with other disciplines, such as a joint study of the relationship between housing, health and well-being. Smart cities research is also a good direction; for example, research on how innovation in smart infrastructure could create economic growth.

BIBLIOGRAPHY

[1] George A Akerlof and Robert J Shiller. *Animal spirits: How human psychology drives the economy, and why it matters for global capitalism*. Princeton University Press, 2010.

[2] Ashik Alvi, Siuly Siuly, and Hua Wang. *Developing a Deep Learning Based Approach for Anomalies Detection from EEG Data*, pages 591–602. International Conference on Web Information Systems Engineering, 01 2021.

[3] Ashik Alvi, Siuly Siuly, Hua Wang, Kate Wang, and Frank Whittaker. A deep learning based framework for diagnosis of mild cognitive impairment. *Knowledge-Based Systems*, 248:108815, 07 2022.

[4] Concha Artola and Enrique Martínez-Galán. Tracking the future on the web: construction of leading indicators using internet searches. *Banco de Espana Occasional Paper No. 1203*, 2012.

[5] Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 492–499, Washington, DC, USA, 2010. IEEE Computer Society.

[6] Susan Athey. Machine learning and causal inference for policy evaluation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 5–6, New York, NY, USA, 2015. ACM.

[7] Susan Athey et al. The impact of machine learning on economics. *The economics of artificial intelligence: An agenda*, pages 507–547, 2018.

[8] Malcolm Baker and Jeffrey Wurgler. Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61(4):1645–1680, 2006.

[9] World Bank. *World development report 2015: Mind, society, and behavior*. The World Bank, 2014.

[10] Nicholas Barberis, Andrei Shleifer, and Robert Vishny. A model of investor sentiment. *Journal of financial economics*, 49(3):307–343, 1998.

[11] Mirosław Bełej. Does google trends show the strength of social interest as a predictor of housing price dynamics? *Sustainability*, 14(9):5601, 2022.

[12] Archith J Bency, Swati Rallapalli, Raghu K Ganti, Mudhakar Srivatsa, and BS Manjunath. Beyond spatial auto-regressive models: Predicting housing prices with satellite imagery. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 320–329. IEEE, 2017.

[13] Yoshua Bengio, Yann Lecun, and Geoffrey Hinton. Deep learning for ai. *Communications of the ACM*, 64(7):58–65, 2021.

[14] European Systemic Risk Board. Vulnerabilities in the eu residential real estate sector, 2016.

[15] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1 – 8, 2011.

[16] Federico Botta and Mario Gutiérrez-Roig. Modelling urban vibrancy with mobile phone and openstreetmap data. *Plos one*, 16(6):e0252015, 2021.

[17] Jean-Charles Bricongne, Baptiste Meunier, and Pouget Sylvain. Web scraping housing prices in real-time: the covid-19 crisis in the uk. 2021.

[18] Paolo Buonanno, Daniel Montolio, and Josep Maria Raya-Vílchez. Housing prices and crime perception. *Empirical Economics*, 45(1):305–321, 2013.

[19] Ernest W Burgess. The growth of the city: an introduction to a research project. In *Urban ecology*, pages 71–78. Springer, 2008.

[20] Anastasia Buyalskaya, Marcos Gallo, and Colin F Camerer. The golden age of social science. *Proceedings of the National Academy of Sciences*, 118(5), 2021.

[21] Ayse Can. The measurement of neighborhood dynamics in urban house prices. *Economic geography*, 66(3):254–272, 1990.

[22] Ayse Can. Specification and estimation of hedonic housing price models. *Regional science and urban economics*, 22(3):453–474, 1992.

[23] Yan Carrière-Swallow and Felipe Labbé. Nowcasting with google trends in an emerging market. *Journal of Forecasting*, 32(4):289–298, 2013.

[24] Bradford Case, Henry O. Pollakowski, and Susan M. Wachter. On choosing among house price index methodologies. *Real Estate Economics*, 19(3):286–307, 1991.

[25] Karl E Case and Robert J Shiller. The efficiency of the market for single-family homes, 1988.

[26] Karl E Case and Robert J Shiller. Forecasting prices and excess returns in the housing market. *Real Estate Economics*, 18(3):253–273, 1990.

[27] Karl E Case and Robert J Shiller. Is there a bubble in the housing market? *Brookings papers on economic activity*, 2003(2):299–362, 2003.

[28] Karl E Case, Robert J Shiller, and Anne Thompson. What have they been thinking? home buyer behavior in hot and cold markets. Technical report, National Bureau of Economic Research, 2012.

[29] Hyunyoung Choi and Hal Varian. Predicting the present with google trends. *Economic record*, 88:2–9, 2012.

[30] Joe Cortright. *Walking the walk: How walkability raises home values in US cities*. CEOs for Cities, 2009.

[31] Kent Daniel, David Hirshleifer, and Avanidhar Subrahmanyam. Investor psychology and security market under-and overreactions. *the Journal of Finance*, 53(6):1839–1885, 1998.

[32] Werner F. M. De Bondt and Richard H. Thaler. Anomalies: A mean-reverting walk down wall street. *Journal of Economic Perspectives*, 3(1):189–202, March 1989.

[33] Marco De Nadai and Bruno Lepri. The economic value of neighborhoods: Predicting real estate prices from the urban environment. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 323–330. IEEE, 2018.

[34] Marco De Nadai, Jacopo Staiano, Roberto Larcher, Nicu Sebe, Daniele Quercia, and Bruno Lepri. The death and life of great italian cities: a mobile phone data perspective. In *Proceedings of the 25th international conference on world wide web*, pages 413–423, 2016.

[35] Marco De Nadai, Radu Laurentiu Vieriu, Gloria Zen, Stefan Dragicevic, Nikhil Naik, Michele Caraviello, Cesar Augusto Hidalgo, Nicu Sebe, and Bruno Lepri. Are safer looking neighborhoods more lively? a multimodal investigation into urban life. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1127–1135, 2016.

[36] Guy Debelle. Housing and the economy. Technical report, Tech. rept. Reserve Bank of Australia, 2019.

[37] Paul DiMaggio. Culture and cognition. *Annual review of sociology*, 23(1):263–287, 1997.

[38] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Knowledge-driven event embedding for stock prediction. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, pages 2133–2142, 2016.

[39] Dave Donaldson and Adam Storeygard. The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, 30(4):171–98, November 2016.

[40] Robin A Dubin. Predicting house prices using multiple listings data. *The Journal of Real Estate Finance and Economics*, 17(1):35–59, 1998.

[41] Liran Einav, Dan Knoepfle, Jonathan Levin, and Neel Sundaresan. Sales taxes and internet commerce. *American Economic Review*, 104(1):1–26, January 2014.

[42] Liran Einav and Jonathan Levin. The data revolution and economic analysis. *Innovation Policy and the Economy*, 14(1):1–24, 2014.

[43] Liran Einav and Jonathan Levin. Economics in the age of big data. *Science*, 346(6210):1243089, 2014.

[44] Joseph E. Engelberg and Christopher A. Parsons. The causal impact of media in financial markets. *The Journal of Finance*, 66(1):67–97, 2008.

[45] Mariia Ermilova and Sergey Laptev. Information technology as a tool to improve the efficiency of the housing market. In *E3S Web of Conferences*, volume 234, page 00047. EDP Sciences, 2021.

[46] Cem Ertur, Wilfried Koch, et al. Convergence, human capital and international spillovers. *Laboratoire d'Economie et de Gestion Working Paper*, 2006.

[47] Jonathan St BT Evans and Keith E Stanovich. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3):223–241, 2013.

[48] Eugene F Fama. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25(2):383–417, 1970.

[49] Fuli Feng, Xiangnan He, Xiang Wang, Cheng Luo, Yiqun Liu, and Tat-Seng Chua. Temporal relational ranking for stock prediction. *ACM Transactions on Information Systems (TOIS)*, 37(2):1–30, 2019.

[50] IHME COVID-19 forecasting team. Modeling covid-19 scenarios for the united states. *Nature medicine*, 2020.

[51] Yanjie Fu, Yong Ge, Yu Zheng, Zijun Yao, Yanchi Liu, Hui Xiong, and Jing Yuan. Sparse real estate ranking with online user reviews and offline moving behaviors. In *2014 IEEE International Conference on Data Mining*, pages 120–129. IEEE, 2014.

[52] Yanjie Fu, Hui Xiong, Yong Ge, Zijun Yao, Yu Zheng, and Zhi-Hua Zhou. Exploiting geographic dependencies for real estate appraisal: a mutual perspective of ranking and clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1047–1056, 2014.

[53] Yong-Feng Ge, Jinli Cao, Hua Wang, Zhenxiang Chen, and Yanchun Zhang. Set-based adaptive distributed differential evolution for anonymity-driven database fragmentation. *Data Science and Engineering*, 6, 12 2021.

[54] Yong-Feng Ge, Maria Orlowska, Jinli Cao, Hua Wang, and Yanchun Zhang. Mdde: multitasking distributed differential evolution for privacy-preserving database fragmentation. *The VLDB Journal*, pages 1–19, 01 2022.

[55] Yong-Feng Ge, Wei-Jie Yu, Jinli Cao, Hua Wang, Zhi-Hui Zhan, Yanchun Zhang, and Jun Zhang. Distributed memetic algorithm for outsourced database fragmentation. *IEEE Transactions on Cybernetics*, PP:1–14, 11 2020.

[56] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, 114(50):13108–13113, 2017.

[57] Dirk Geeraerts. *Theories of lexical semantics*. Oxford University Press, 2010.

[58] Eric Gilbert and Karrie Karahalios. Widespread worry and the stock market. In *ICWSM 2010 - Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 58–65, 2010.

[59] Edward L. Glaeser, Scott Duke Kominers, Michael Luca, and Nikhil Naik. Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry*, 56(1):114–137, 2018.

[60] Sharad Goel, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, and Duncan J. Watts. Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences*, 107(41):17486–17490, 2010.

[61] Daniel Gruhl, R. Guha, Ravi Kumar, Jasmine Novak, and Andrew Tomkins. The predictive power of online chatter. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 78–87, New York, NY, USA, 2005. ACM.

[62] Christopher Hannum, Kerem Yavuz Arslanli, and Ali Furkan Kalay. Spatial analysis of twitter sentiment and district-level housing prices. *Journal of European Real Estate Research*, 2019.

[63] Shaobo He, Yuexi Peng, and Kehui Sun. Seir modeling of the covid-19 and its dynamics. *Nonlinear dynamics*, 101(3):1667–1680, 2020.

[64] J. Vernon Henderson, Adam Storeygard, and David N. Weil. Measuring economic growth from outer space. *American Economic Review*, 102(2):994–1028, April 2012.

[65] Desislava Hristova, Luca M Aiello, and Daniele Quercia. The new urban success: How culture pays. *Frontiers in Physics*, 6:27, 2018.

[66] Hong Hu, Jiuyong Li, Hua Wang, and G.E. Daggard. Combined gene selection methods for microarray data analysis. In *International conference on knowledge-based and intelligent information and engineering systems*, volume 4251, pages 976–983, 10 2006.

[67] Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 261–269, 2018.

[68] Hong Huang, Bo Zhao, Hao Zhao, Zhou Zhuang, Zhenxuan Wang, Xiaoming

Yao, Xinggang Wang, Hai Jin, and Xiaoming Fu. A cross-platform consumer behavior analysis of large-scale mobile shopping data. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1785–1794. International World Wide Web Conferences Steering Committee, 2018.

[69] Jiajia Huang, Min Peng, Hua Wang, Jinli Cao, Wang Gao, and Xiuzhen Zhang. A probabilistic method for emerging topic tracking in microblog stream. *World Wide Web*, 20(2):325–350, Mar 2017.

[70] C.J. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM)*, pages 216–225, 2014.

[71] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

[72] Narasimhan Jegadeesh and Sheridan Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1):65–91, 1993.

[73] Haixin Jiang, Rui Zhou, Limeng Zhang, Hua Wang, and Yanchun Zhang. Sentence level topic models for associated topics extraction. *World Wide Web*, 22:2545–2560, 11 2019.

[74] Daniel Kahneman. Maps of bounded rationality: Psychology for behavioral economics. *American economic review*, 93(5):1449–1475, 2003.

[75] Daniel Kahneman, Jack L Knetsch, and Richard H Thaler. Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic perspectives*, 5(1):193–206, 1991.

[76] Daniel Kahneman and Amos Tversky. On the study of statistical intuitions. *Cognition*, 11(2):123–141, 1982.

[77] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

[78] John F Kain and John M Quigley. Measuring the value of housing quality. *Journal of the American statistical association*, 65(330):532–548, 1970.

[79] Deanna M Kennedy, Gustavo José Zambrano, Yiyu Wang, and Osmar Pinto Neto. Modeling the effects of intervention strategies on covid-19 transmission dynamics. *Journal of Clinical Virology*, 128:104440, 2020.

[80] Faten Khalil, Jiuyong Li, and Hua Wang. Integrating markov model with clustering for predicting web page accesses. *AusWeb 2007: 13th Australasian World Wide Web Conference*, 01 2007.

[81] Nobuhiro Kiyotaki. A perspective on modern business cycle theory. *FRB Richmond Economic Quarterly*, 97(3):195–208, 2011.

[82] Zona Kostic and Aleksandar Jevremovic. What image features boost housing market predictions? *IEEE Transactions on Multimedia*, 2020.

[83] S. P. Kothari, Xu Li, and James E. Short. The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *The Accounting Review*, 84(5):1639–1670, 2009.

[84] Jiaying Kou, Xiaoming Fu, Jiahua Du, Hua Wang, and Geordie Z Zhang. Understanding housing market behaviour from a microscopic perspective. In *2018 27th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–9. IEEE, 2018.

[85] Paul Krugman. How did economists get it so wrong? *New York Times*, 2(9):2009, 2009.

[86] Theresa Kuchler, Dominic Russel, and Johannes Stroebel. Jue insight: The geographic spread of covid-19 correlates with the structure of social networks as measured by facebook. *Journal of Urban Economics*, page 103314, 2021.

[87] Stephen Law, Brooks Paige, and Chris Russell. Take a look around: using street view and satellite images to estimate house prices. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5):1–19, 2019.

[88] Edward E Leamer. Housing is the business cycle. Technical report, National Bureau of Economic Research, 2007.

[89] Chaeyoung Lee, Soobin Kwak, and Junseok Kim. Controlling covid-19 outbreaks with financial incentives. *International journal of environmental research and public health*, 18(2):724, 2021.

[90] Wassily Leontief. Theoretical assumptions and nonobserved facts. *American Economic Review*, 61(1):1–7, March 1971.

[91] James P LeSage. An introduction to spatial econometrics. *Revue d'économie industrielle*, 3(123):19–44, 2008.

[92] Hu Li, Ye Wang, Hua Wang, and Bin Zhou. Multi-window based ensemble learning for classification of imbalanced streaming data. *World Wide Web*, 20(6):1507–1525, 2017.

[93] Hu Li, Ye Wang, Hua Wang, and Bin Zhou. Multi-window based ensemble learning for classification of imbalanced streaming data. *World Wide Web*, 20:1–19, 11 2017.

[94] Jian-Yu Li, Ke-Jing Du, Zhi-Hui Zhan, Hua Wang, and Jun Zhang. Distributed differential evolution with adaptive resource allocation. *IEEE transactions on cybernetics*, PP, 03 2022.

[95] A Christopher Limnios and Hao You. Can google trends improve housing market forecasts? *Curiosity: Interdisciplinary Journal of Research and Innovation*, 1(2):21987, 2021.

[96] Xiaobai Liu, Qian Xu, Jingjie Yang, Jacob Thalman, Shuicheng Yan, and Jiebo Luo. Learning multi-instance deep ranking and regression network for visual house appraisal. *IEEE Transactions on Knowledge and Data Engineering*, 30(8):1496–1506, 2018.

[97] George Loewenstein and Drazen Prelec. Anomalies in intertemporal choice: Evidence and an interpretation. *The Quarterly Journal of Economics*, 107(2):573–597, 1992.

[98] Tim Loughran and Bill Mcdonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

[99] Burton G Malkiel and Eugene F Fama. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25(2):383–417, 1970.

[100] Ulrike Malmendier and Young Han Lee. The Bidder's Curse. *American Economic Review*, 101(2):749–787, April 2011.

[101] Nick McLaren and Rachana Shanbhogue. Using internet search data as

economic indicators. *Bank of England Quarterly Bulletin*, 51(2):134–140, 2011.

[102] Daniel P McMillen. Changes in the distribution of house prices over time: Structural characteristics, neighborhood, or coefficients? *Journal of Urban Economics*, 64(3):573–589, 2008.

[103] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[104] Anshul Mittal and Arpit Goel. Stock prediction using twitter sentiment analysis. *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)*, 15, 2012.

[105] Sendhil Mullainathan and Jann Spiess. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, May 2017.

[106] Nikhil Naik, Scott Duke Kominers, Ramesh Raskar, Edward L Glaeser, and César A Hidalgo. Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences*, 114(29):7571–7576, 2017.

[107] United Nations. The world's cities in 2018. *Department of Economic and Social Affairs, Population Division, World Urbanization Prospects*, pages 1–34, 2018.

[108] Gregory B Northcraft and Margaret A Neale. Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational behavior and human decision processes*, 39(1):84–97, 1987.

[109] Liv Osland. An application of spatial econometrics in relation to hedonic house price modeling. *Journal of Real Estate Research*, 32(3):289–320, 2010.

[110] Scott. E. Page. Computational models from a to z. *Complexity*, 5(1):35–41, 1999.

[111] Dinesh Pandey, Hua Wang, Xiaoxia Yin, Kate Wang, Yanchun Zhang, and

Jing Shen. Automatic breast lesion segmentation in phase preserved dce-mris. *Health Information Science and Systems*, 10, 05 2022.

[112] Carita Paradis. Lexical semantics. In *The encyclopedia of applied linguistics*. Wiley-Blackwell, 2012.

[113] Min Peng, Jiahui Zhu, Hua Wang, Xuhui Li, Yanchun Zhang, Xiuzhen Zhang, and Gang Tian. Mining event-oriented topics in microblog stream with unsupervised multi-view hierarchical embedding. *ACM Trans. Knowl. Discov. Data*, 12(3):38:1–38:26, April 2018.

[114] Henry O Pollakowski and Traci S Ray. Housing price diffusion patterns at different aggregation levels: an examination of housing market efficiency. *Journal of Housing Research*, pages 107–124, 1997.

[115] Michael E Porter and Michael P Porter. Location, clusters, and the" new" microeconomics of competition. *Business Economics*, pages 7–13, 1998.

[116] Omid Poursaeed, Tomáš Matera, and Serge Belongie. Vision-based real estate price estimation. *Machine Vision and Applications*, 29(4):667–676, 2018.

[117] David Powers. Evaluation: From precision, recall and f-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

[118] You Ren. *Bayesian Modeling of a High Resolution Housing Price Index*. PhD thesis, University of Washington, 2015.

[119] Rubina Sarki, Khandakar Ahmed, Hua Wang, and Yanchun Zhang. Automated detection of mild and multi-class diabetic eye diseases using deep learning. *Health Information Science and Systems*, 8, 10 2020.

[120] Rubina Sarki, Khandakar Ahmed, Hua Wang, Yanchun Zhang, and Kate Wang. Convolutional neural network for multi-class classification of diabetic eye disease. *ICST Transactions on Scalable Information Systems*, page 172436, 12 2021.

[121] Rubina Sarki, Khandakar Ahmed, Hua Wang, Yanchun Zhang, and Kate Wang. Automated detection of covid-19 through convolutional neural network using chest x-ray images. *PLOS ONE*, 17:e0262052, 01 2022.

[122] Kevin Schawinski, M Dennis Turp, and Ce Zhang. Exploring galaxy evolution with generative models. *Astronomy & Astrophysics*, 616:L16, 2018.

[123] Robert P. Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Trans. Inf. Syst.*, 27(2):12:1–12:19, March 2009.

[124] Wen Shi, Wei-neng Chen, Sam Kwong, Jie Zhang, Hua Wang, Gu Tianlong, Huaqiang Yuan, and Jun Zhang. A coevolutionary estimation of distribution algorithm for group insurance portfolio. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, PP:1–15, 07 2021.

[125] Robert J Shiller. From efficient markets theory to behavioral finance. *Journal of economic perspectives*, 17(1):83–104, 2003.

[126] Robert J Shiller. *Irrational exuberance: Revised and expanded third edition*. Princeton university press, 2015.

[127] Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5):e19467, 2011.

[128] Herbert A Simon. On the behavioral and rational foundations of economic dynamics. *Journal of Economic Behavior & Organization*, 5(1):35–55, 1984.

[129] Ravinder Singh, Yanchun Zhang, Hua Wang, Yuan Miao, and Khandakar Ahmed. Investigation of social behaviour patterns using location-based data – a melbourne case study. *ICST Transactions on Scalable Information Systems*, 8:166767, 10 2020.

[130] Cindy K Soo. Quantifying sentiment with news media across local housing markets. *The Review of Financial Studies*, 31(10):3689–3719, 2018.

[131] Stefan Steiniger, Mohammad Ebrahim Poorazizi, Daniel R Scott, Cristian Fuentes, and Ricardo Crespo. Can we use openstreetmap pois for the evaluation of urban accessibility? In *International Conference on GIScience Short Paper Proceedings*, volume 1, 09 2016.

[132] James H Stock, Mark W Watson, et al. *Introduction to econometrics*, volume 3. Pearson New York, 2012.

[133] Daoyuan Sun, Yudie Du, Wei Xu, Mei Yun Zuo, Ce Zhang, and Junjie Zhou.

Combining online news articles and web search to predict the fluctuation of real estate market in big data context. *Pacific Asia Journal of the Association for Information Systems*, 6(4):2, 2014.

[134] Xiaoxun Sun, Hua Wang, Jiuyong Li, and Jian Pei. Publishing anonymous survey rating data. *Data Min. Knowl. Discov.*, 23:379–406, 11 2011.

[135] Supriya Supriya, Siuly Siuly, Hua Wang, and Yanchun Zhang. Automated epilepsy detection techniques from electroencephalogram signals: a review study. *Health Information Science and Systems*, 8, 10 2020.

[136] Ann Swidler. Culture in action: Symbols and strategies. *American sociological review*, pages 273–286, 1986.

[137] Mark Junjie Tan and ChengHe Guan. Are people happier in locations of high property value? spatial temporal analytics of activity frequency, public sentiment and housing price using twitter data. *Applied Geography*, 132:102474, 2021.

[138] Md. Nurul Ahad Tawhid, Siuly Siuly, Kate Wang, and Hua Wang. *Data Mining Based Artificial Intelligent Technique for Identifying Abnormalities from Brain Signal Data*, pages 198–206. International Conference on Web Information Systems Engineering, 01 2021.

[139] Paul C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007.

[140] Paul C. Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467, 2008.

[141] Richard H Thaler. Anomalies: The winner's curse. *Journal of Economic Perspectives*, 2(1):191–202, 1988.

[142] Marius Thériault, François Des Rosiers, Paul Villeneuve, and Yan Kestens. Modelling interactions of location with specific value of housing attributes. *Property Management*, 2003.

[143] Aviral Kumar Tiwari, Rangan Gupta, and Mark E Wohar. Is the housing market in the united states really weakly-efficient? *Applied Economics Letters*, 27(14):1124–1134, 2020.

[144] Michael Tomasello. *A natural history of human thinking*. Harvard University Press, 2018.

[145] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. Highly accurate protein structure prediction for the human proteome. *Nature*, pages 1–9, 2021.

[146] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

[147] Maarten Van Ham and David Manley. The effect of neighbourhood housing tenure mix on labour market outcomes: a longitudinal investigation of neighbourhood effects. *Journal of Economic Geography*, 10(2):257–282, 2009.

[148] Hal R. Varian. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28, May 2014.

[149] Pasupathy Vimalachandran, Hong Liu, Yongzheng Lin, Ke Ji, Hua Wang, and Yanchun Zhang. Improving accessibility of the australian my health records while preserving privacy and security of the system. *Health Information Science and Systems*, 8, 10 2020.

[150] Guangyuan Wang, Hua Wang, Xiaohui Tao, and Ji Zhang. A self-stabilizing algorithm for finding a minimal positive influence dominating set in social networks. In *Proceedings of the 24th Australasian Database Conference (ADC 2013)*, pages 93–100. Australian Computer Society Inc., 2013.

[151] Hua Wang, Xiaohong Jiang, and Georgios Kambourakis. Special issue on security, privacy and trust in network-based big data. *Information Sciences—Informatics and Computer Science, Intelligent Systems, Applications: An International Journal*, 318(C):48–50, 2015.

[152] Hua Wang and Lili Sun. Trust-involved access control in collaborative open social networks. In *2010 Fourth International Conference on Network and System Security*, pages 239–246. IEEE, 2010.

[153] Lynn Wu and Erik Brynjolfsson. The future of prediction: How google searches foreshadow housing prices and sales. In *Economic Analysis of the Digital Economy*, pages 89–118. University of Chicago Press, April 2015.

[154] Yumo Xu and Shay B Cohen. Stock movement prediction from tweets and

historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979, 2018.

[155] Jiao Yin, MingJian Tang, Jinli Cao, Hua Wang, Mingshan You, and Yongzheng Lin. Adaptive online learning for vulnerability exploitation time prediction. In *Web Information Systems Engineering – WISE 2020*, pages 252–266, Cham, 2020. Springer International Publishing.

[156] Jiao Yin, MingJian Tang, Jinli Cao, Hua Wang, Mingshan You, and Yongzheng Lin. Vulnerability exploitation time prediction: An integrated framework for dynamic imbalanced learning. *World Wide Web*, 25(1):401–423, 2022.

[157] Mingshan You, Jiao Yin, Hua Wang, Jinli Cao, and Yuan Miao. *A Minority Class Boosted Framework for Adaptive Access Control Decision-Making*, pages 143–157. International Conference on Web Information Systems Engineering, 01 2021.

[158] Quanzeng You, Ran Pang, Liangliang Cao, and Jiebo Luo. Image-based appraisal of real estate properties. *IEEE transactions on multimedia*, 19(12):2751–2759, 2017.

[159] Ji Zhang, Xiaohui Tao, and Hua Wang. Outlier detection from large distributed databases. *World Wide Web*, 17(4):539–568, 2014.

[160] Xue Zhang, Hauke Fuehres, and Peter A Gloor. Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Procedia - Social and Behavioral Sciences*, 26:55 – 62, 2011. The 2nd Collaborative Innovation Networks Conference - COINs2010.

[161] Hao Zhao, Qingyuan Gong, Yang Chen, Jingrong Chen, Yong Li, and Xiaoming Fu. This place is swarming: Using a mobile social app to study human traffic in cities. In *5th IEEE International Workshop on Crowd Assisted Sensing, Pervasive Systems and Communications (CASPer'18), in conjunction with IEEE PerCom*, 2018.