2016

Predicting Intake of Applications for First Registration in the Property Registration Authority

Orlaith Mernagh
Technological University Dublin, orlaith8@hotmail.com

Recommended Citation

Mernagh, O. (2016). Predicting Intake of Applications for First Registration in the Property Registration Authority [Technological University Dublin]. DOI: 10.21427/MGZD-5667

Predicting Intake of Applications for First Registration in the Property Registration Authority

Orlaith Mernagh

C99391341

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computing (Data Analytics)

2016

DECLARATION

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the test of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research..

Signature of candidate:………………………………………….

Date: 15 September 2016

ABSTRACT

The motivation for this dissertation is rooted in a real business need. The Property Registration Authority is the state organisation tasked with maintaining a register of land ownership on the island of Ireland. The PRA currently faces a series of challenges; a high level of staff retiring and the inherent loss of knowledge associated with this trend, a lack of recruitment in recent years and a large increase in lodgement of applications for first registration as a result of legislation. The organisation therefore requires a reliable system for predicting future intake. Prior to this project, there has also been a lack of understanding of the factors that influence intake, and that go much of the way to explaining the peaks and troughs in intake levels that have been seen over recent years. Therefore, this dissertation seeks to identify the factors that influence intake of applications for first registration, and to ascertain if these features may be used to build models to predict future intake.

To answer these questions, an exercise in data analytics has been designed and implemented, following the industry standard CRISP-DM methodology. As part of this process, a review of contemporary literature has been carried out, on the subjects of the Irish property market, the factors that influence the level of demand for registration, and modelling approaches applied to variable selection and predictive modelling. Using the insights gleaned from this research, a varied dataset has been sourced, assembled, explored and prepared which includes property registration data, house sale data and economic indicator data. The final dataset has been used to build a series of predictive models, and after evaluation the results show the Random Forest model to be the most effective. A further finding is that the combined outcome of all of the models indicates that the number of houses sold is the single most important factor in predicating volume of applications for registration.

The series of experiments conducted and the body of research analysed have presented several valuable insights into the housing market and the factors that influence it, the modelling techniques that can be applied to intake prediction and the key prediction factors that influence intake. The overall conclusion of the study is that the null hypothesis has been rejected and that intake of ‘first registration’ applications in the Property Registration Authority can be predicted through analysis of historical intake data and external factors. However, it is acknowledged that further work will be required to develop a data gathering and analysis process that can be operationalised in the PRA context.

Keywords: Housing Market, Intake, Ireland, Predictive Analytics, Property, Registration

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my project supervisor, Damian Gordon, for his exceptional guidance, patience, advice and motivation.

Thanks also to my colleagues in the Property Registration Authority for their support and insight, and to the staff of the Central Statistics Office, Property Services Regulatory Authority and Central Bank for their invaluable advice and information.

Finally, special thanks to my family, for their perpetual support in a myriad of ways.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

GLOSSARY OF TERMS

ABT	Analytical Base Table
AIC	Akaike Information Criterion
ANN	Artificial Neural Network
CDS	Clinical Decision Support
CPI	Consumer Price Index
cp	Cost Complexity Factor
CSO	Central Statistics Office
DFM	Dynamic Factor Model
ECF Unit	Examiners Cross Functional Unit
ESRI	The Economic and Social Research Institute
GDP	Gross Domestic Product
LBVAR model	Large-scale Bayesian Vector Autoregressive model
%MAE	Percentage Mean Absolute Error
MSB	Multi-Storey Building
OECD	The Organisation for Economic Co-Operation and Development
PRA	Property Registration Authority
PSRA	Property Services Regulatory Authority
RMSE	Root Mean Square Error
%RMSE	Percentage Root Mean Squared Error
REAP	Risk Analysis Regression Programme (Revenue)
ROD	Registry of Deeds
SQL	Structured Query Language
SVD model	Single Value Decomposition model
SVR	Standard Variable Rate
UCCh framework	User Cost of Capital framework

1. INTRODUCTION

The aim of this research is to analyse the factors that influence intake of applications to the Property Registration Authority, the organisation which is tasked with maintaining a register of land ownership on the island of Ireland. This study will include profiling and exploration of the system of registration, a review of contemporary literature on the subject of the property markets both in Ireland and internationally, and an examination of modelling techniques that can potentially be employed to predict the volume of incoming applications for registration. This chapter will introduce the research questions to be explored, set out the structure of the document, and discuss the methodology to be employed.

1.1. Background & Context

At present, there are two systems of registration in Ireland - the Land Registry and the Registry of Deeds - under the common umbrella of the Property Registration Authority, or 'PRA'.

The Land Registry operates a registration of title system and maintains a 'Land Register' recording the ownership of land and other rights through a combination of a textual document (the 'folio') and a map recording location. In comparison, the Registry of Deeds operates a registration of deeds system, maintaining a combination of manual, micro-film and electronic documents that record the existence of deeds affecting property. These two systems currently operate concurrently in Ireland. However, since 2011 compulsory first registration has been extended to all counties, and thus it is envisaged that eventually the vast majority of property will be registered in the Land Registry. The structure and functions of the two organisations will be discussed in further detail in Chapter 3.

The principal objective of the PRA in recent years has been the completion of the Irish Land Register. At present, approximately 93% of all land is registered (The Property Registration Authority, 2015). Thus, the strategic emphasis within the organisation is on targeting the remaining 7% of unregistered titles. Several tactics have been pursued to date, but the most important has been the introduction of nationwide compulsory first registration for sold lands, under Statutory Instrument 516 of 2010. This legislation requires that any unregistered property purchased in the state after 1 June 2011 is subject to compulsory first registration[1].

[1] Registration of Title Act 1964 (Compulsory Registration of Ownership) (Cork and Dublin) Order 2010 (S.I. No. 516/2010), Pub. L. No. 516 (2010). Ireland. Retrieved from http://www.irishstatutebook.ie/eli/2010/si/516/made/en/print

As a result of this legislation, since 2011 the PRA has experienced greatly increased volumes of applications for first registration; up 450% from 4,527 applications in 2010 to 20,559 in 2015. The Senior Management Team requires the facility to anticipate the number of applications that will be lodged in the near future. This would enable the team to redirect staff as required, manage resources efficiently and reduce the likelihood of backlogs building up.

Concurrent to this large increase in applications for registration, the housing market in Ireland has undergone a period of huge change over the past twenty years. In this time, the country has gone through a series of economic peaks and troughs, and the property market has echoed these patterns. The two markets appear to be intertwined; when the economy is booming and unemployment rates are low, financial institutions lend freely, an increased number of houses are bought and sold, and sale prices steadily increase. Conversely, during times of economic recession the property market tends to collapse as demand and prices fall and the availability of mortgages is greatly reduced. Thus it will be seen that the economy and housing market are strongly correlated; this project will seek to explore the extent to which the housing market also affects intake of first registration applications.

The PRA faces a series of significant challenges in the current economic climate. At a time of greatly increased intake, the organisation has yet to fully recover from the impact of the recent recession. Staffing levels are at a critical level as a result of natural loss due to retirements and the moratorium on recruitment that has only recently been lifted. In addition, the demographic profile of the PRA workforce means that many essential staff are fast approaching retirement age. The PRA must continue to evolve and adapt in order to meet these new challenges, in order to cope with increased intake while maintaining standards of accuracy and protecting the integrity of the Register.

1.2. Research Problem

The aim of this research document is to establish if it is feasible to develop a model to predict the quantity of applications for first registration that will be lodged in the near future. Currently, there is an extremely basic method in operation, basing the targets for the coming month on the number of applications lodged the previous month. This system is not fit for purpose since it does not provide reliable figures that can be used for planning. Therefore the key research question is:

> *"What are the factors that influence intake of applications for first registration to the Property Registration Authority, and can these features be used to build models to predict future intake?"*

1.3. Research Hypothesis and Objectives

H_1: Intake of 'first registration' applications in the Property Registration Authority can be predicted to a set degree of accuracy through analysis of historical intake data and external factors.

H_0: Intake of 'first registration' applications in the Property Registration Authority cannot be predicted to a set degree of accuracy through analysis of historical intake data and external factors.

In order to test this hypothesis, quantitative research will be used to conduct a systematic empirical investigation. The research objective is to use machine learning techniques to determine if there is a correlation between various factors as suggested by previous literature and research and the number of first registration applications lodged. If so, can these factors be harnessed in the form of a model in order to accurately predict future intake?

The experiment will require the sourcing and preparing of a dataset, building a series of models and comparing the outputs to a test set. If any of the models achieve a minimum threshold of success in predicting intake, then the hypothesis can be deemed to be proved.

1.4. Scope and Limitations

In the course of the literature review carried out, it has not been possible to find any papers directly dealing with prediction of intake of applications on a monthly basis; therefore the gap identified in this instance is rooted in a real business need. This presents an opportunity to exploit the wealth of historical data available in the PRA to both reveal trends and to build prediction models that can be used for future planning.

Potential challenges can be foreseen at all stages of the proposed experiment, from gaining a full understanding of what the Senior Management Team hope to achieve, to integrating multiple datasets from various sources and processing them into a single usable dataset. The choice of algorithms to apply will also be key to the project, as will accurate identification of the features that have the biggest impact on application intake.

1.5. Research Methodology

This section outlines the software to be used in the course of the experiment, describes the dataset in broad terms and discusses the research methodology to be used, which is CRISP-DM.

1.5.1. Software

The software used will include a mixture of proprietary and open source software:

- Microsoft Excel for accessing the raw data and collating sources into a single dataset. Excel is often overlooked in favour of more specific tools in Data Analysis, but it remains a familiar and powerful programme capable of storing and processing up to a million rows. In this project, the filtering, searching and sorting tools were used most often.
- Tableau for creating exploratory and explanatory visualisations. This proprietary software enables quick and effective graphical analysis, helping the analyst to gain a better understanding of the data and to express this in a visual manner.
- R-Studio for building models and evaluating the performance of the final models against the test set. R Studio is a user interface for the open source programming language and environment R, and consists of a console and tools for plotting and managing the R workspace. R itself is a language specifically designed for statistical and data analysis. It contains all the tools needed to build and analyse models.

1.5.2. Dataset

The gathering and processing of data will be discussed further in Chapter 4. In broad terms, the sources identified for accessing raw data are:

- transactional data extracted directly from the PRA database
- open data from the Central Statistic Office's 'StatBank' repository – specifically Consumer Price Index data and unemployment data
- property sale data downloaded from the Property Services Regulatory Authority website

1.5.3. CRISP-DM

The experiment will follow the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology, in order to give the experiment a logical structure that aligns with industry standards (Shearer, 2000). The diagram and list on the following page describes the planned structure of the experiment, and the corresponding chapter structure:

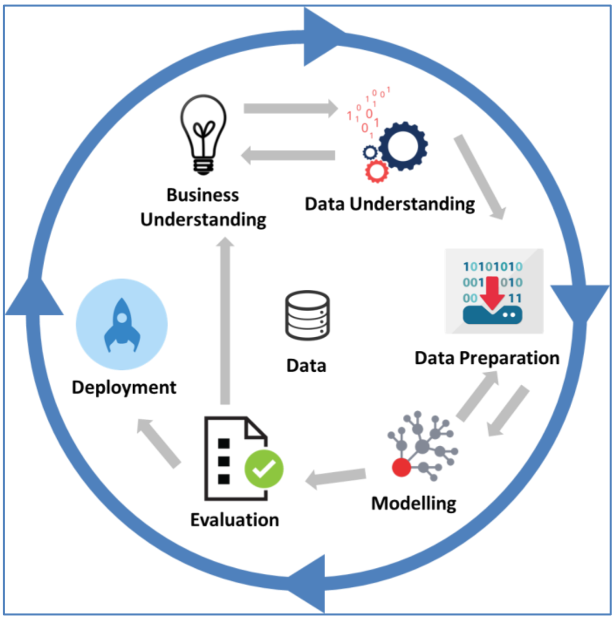


Figure 1.1 - Diagram describing structure of experiment

Business Understanding (Chapters 2 & 3):

- Chapter 2 will comprise a review of published contemporary literature in this field. The aim is to identify gaps and limitations and discuss how these findings are relevant to this dissertation.
- Chapter 3 will provide an insight into the work and structure of the PRA, the organisation's future plans, and discuss how this research project has the potential to assist the PRA in achieving their aims.

Data Understanding (Chapters 4 & 5):

- The exploration of each feature in the dataset and creation of exploratory visualisations and statistical reports to gain insight into the data.

Data Preparation (Chapter 4):

- Integrating the various datasets into one unified dataset and dealing with missing values, outliers and any other inconsistencies.

Modelling (Chapter 6):

- The creation of a series of models based on a variety of algorithms, being guided in choice of model by the literature review in Chapter 2.

Evaluation (Chapter 7):

- Evaluation of the models built, using various methods to examine the strengths and weaknesses of each and to identify the optimal model.
- This chapter will also evaluate the experiment and assess the extent to which the aims have been achieved.

Deployment (Chapter 8):

- Chapter 8 will report on the outcome of the overall process, determine if the hypothesis can be deemed to be proven or not, and suggest potential avenues for further research.

2. PREDICTION & THE PROPERTY MARKET

2.1. *Introduction*

This chapter records the learning from a wide exploration of previously conducted research, which will inform the approach taken in the present study to modelling and evaluating the dataset. This exploration of the literature relevant to the prediction of quantity of property sales focuses on two main areas of commentary: housing markets and the factors that influence them; and various approaches taken to modelling datasets in housing market or other similar trend research.

2.2. *Prediction and the Property Market*

There are many reasons why analysts are asked to make predictions with relation to property markets, be it predicting future house prices or potential quantity of sales. Home owners may wish to gauge the optimum time to sell and maximise their investment, and potential buyers want to avoid buying at the height of a boom and then falling foul of negative equity. Financial institutions require an accurate view of the property market in order to manage risk and control lending (Bin, 2004), and economists monitor the housing market as an indicator of the health of the wider economy.

There is, therefore, a wide range of literature on the subject. Most of these conference papers and journal articles focus on the markets in other countries, but many of the features and trends identified are applicable in the Irish and European markets also.

2.3. *The Irish Housing Market*

As is the case for many small nations, the Irish economy and housing market are intertwined. When the economy is booming - as during the Celtic Tiger years from the mid-1990s to the mid-2000s - unemployment rates tend to be low and financial institutions lend freely. During such times new houses are built, an increased number of houses are bought and sold, and sale prices steadily increase. In turn, the construction sector provides employment and stamp duty returns fuel the economy. Conversely, during times of economic recession the property market collapses as demand and prices fall and the availability of mortgages and loans dries up. This trend was clearly seen during the recent economic crash, as residential house prices in Ireland tripled between 1997 and 2007 during the boom, and then fell by 50% nationally in the following five years (Goodbody Stockbrokers, 2013).

2.3.1. Cyclical Nature of the Market

Property prices and sales are now increasing once again, as the economy picks up and lending increases. This cyclical nature is one of the key trends that characterises the Irish property market. Lyons (2015) analysed property listing and sale prices for Dublin throughout the 20th century, and determined that house prices increased consistently from the 1940s on, with the exception of a downturn in the 1950s. However, these years were all times of high inflation; when adjusted to allow for inflation, peaks in house prices were identified in the 1920s, 1940s, 1960s-70s and 1990-2007 (during the "Celtic Tiger" years).

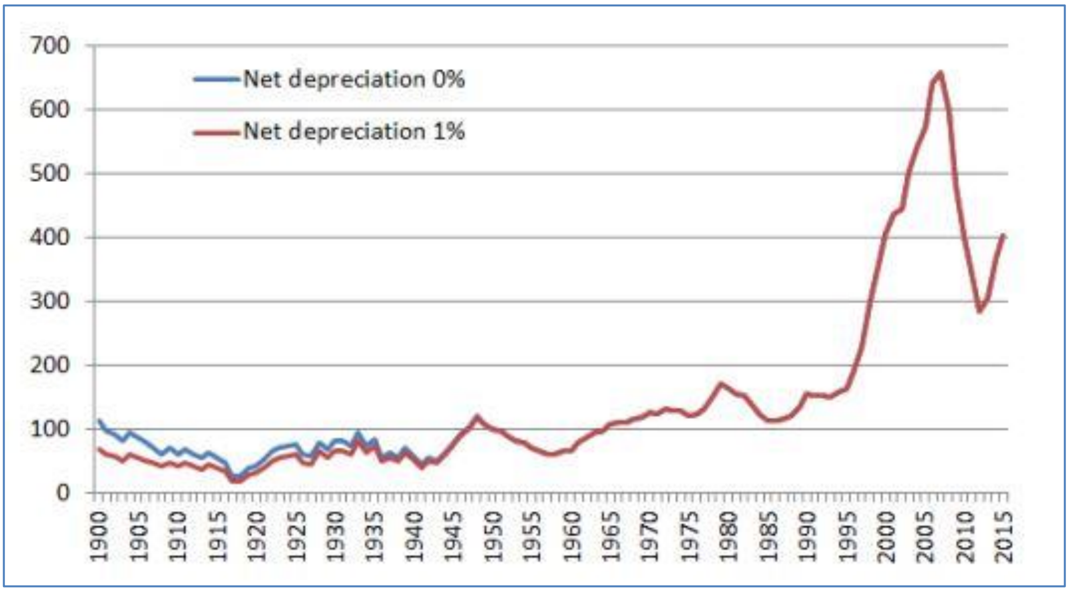


Figure 2.1 - Dublin House Price Index, 1900-2015 (Lyons, 2015)

These findings correlate with those discovered by Deeter, Quinn, & Duffy (2016) in their comprehensive examination of Dublin property prices between 1708 and 1949. By analysing 20,252 memorials of property transactions recorded in the Registry of Deeds, an annual property price index was derived. This index reveals a series of peaks and troughs in the Dublin market, with the trend being generally upwards until 1810, followed by a 50 year decline and then a gradual recovery until a large leap in the early 20th century.

Similarly, a recent report by Goodbody Stockbrokers (2013) highlights the "*inherent cyclical nature of property*" (p.2), both in the Irish context and further afield. While the time frame for a cycle can vary greatly, with the recent Celtic Tiger constituting quite a long growth phase, the stages of a cycle tend to follow a set pattern.

The catalyst for the cycle is economic growth and low interest rates, which lead to increased demand for property. As prices rise the construction industry responds, and building increases. Once the property market becomes more profitable, speculators invest heavily in the industry, with the expectation of large profits. Eventually incentives are reduced in an effort to stabilise the economy, and as investing in property starts to become less lucrative the market slows down. The economy begins to suffer, housing demand reduces and prices start to fall. Finally, a new cycle begins once prices have dropped to such a level as to be an incentive to commencing investment once more (Goodbody Stockbrokers, 2013).

Thus, these peaks and troughs in the property market appear to be strongly influenced by government economic policy and initiatives. For example, Ireland in the 1980s had a massive national debt, which the governments of the time attempted to manage through increased taxation (Whelan, 2013). It is estimated that 90% of all revenue from income tax at this time was being used just to service this debt (Dorgan, 2006). As a result, investment in industry was low and unemployment was extremely high -15% in 1985- as were levels of emigration (Dorgan, 2006). Without jobs and disposable income there was little public demand for housing, and therefore house prices were stagnant and rates of residential construction were low.

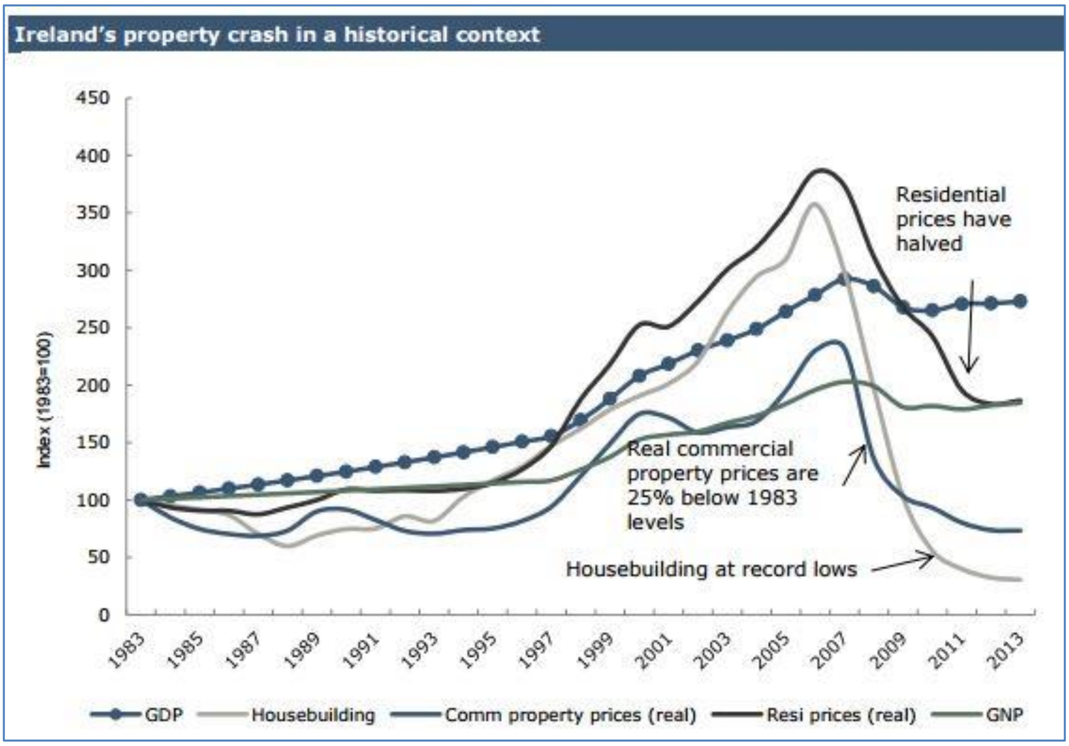


Figure 2.2 - Property Market Indices 1983-2013 (Goodbody Stockbrokers, 2013)

In response, a series of initiatives were introduced that kick-started the boom years of the Celtic Tiger. First, spending on services was reduced and a number of government agencies disbanded. Several industries were targeted for investment, including the financial services, telecommunications and electronics sectors (Dorgan, 2006).

The nation also benefitted at this time from a unique demographic composition, with a high percentage of the population of working age when compared to the UK or the US. Added to this, in the 1990s a large proportion of Irish women entered the workforce for the first time, in a rejection of traditional domestic roles, and rates of emigration fell as more jobs became available at home (Whelan, 2013). All of these factors combined to provide the new economy with a large, young, skilled workforce.

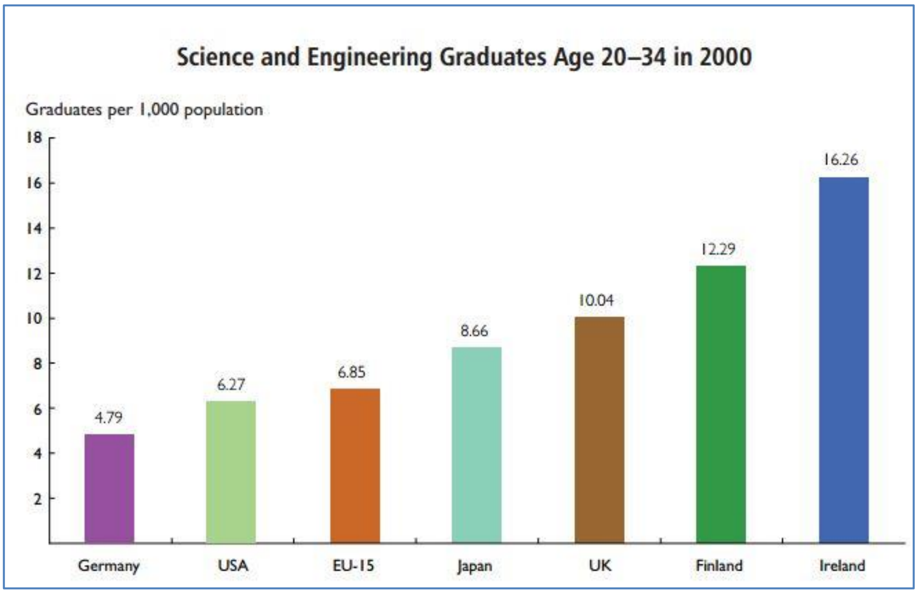


Figure 2.3 - Science & Engineering graduates per 1,000 population (Dorgan, 2006)

As the economy grew and employment increased, the government were able to lower tax rates and increase spending (Whelan, 2013). At the same time, banks and lending institutions reduced interest rates and stimulated the housing market, as the workforce, with greater disposable income than ever before, invested heavily in residential property. The construction industry boomed in response with more and more houses built to satisfy demand, and sale prices soared.

However, while clever policy making may have initiated the Celtic Tiger years, a lack of foresight and timely intervention from the political arena also contributed to its demise, and to the scale of its demise. Several factors combined to cause the catastrophic collapse from 2007 onwards. The natural aging of the population meant that the workforce could no longer be relied on to maintain the economic advantage alone (Whelan, 2013). Added to this, house prices had eventually risen to an unsustainable level, and eventually in 2008 the housing market slowed to a halt. Unfortunately the Irish economy had become dependent on the income from the construction sector, and when it collapsed it had a devastating effect on the nation’s finances. As unemployment soared as a result of the collapse, the government had to deal with increased expenditure as well as reduced income from taxation and stamp duty. The last nails in the coffin were the banking crisis[2] and the advent of an international recession. By 2009, after a decade of budget surpluses, fiscal deficits had risen to 20% of GDP (Whelan, 2013).

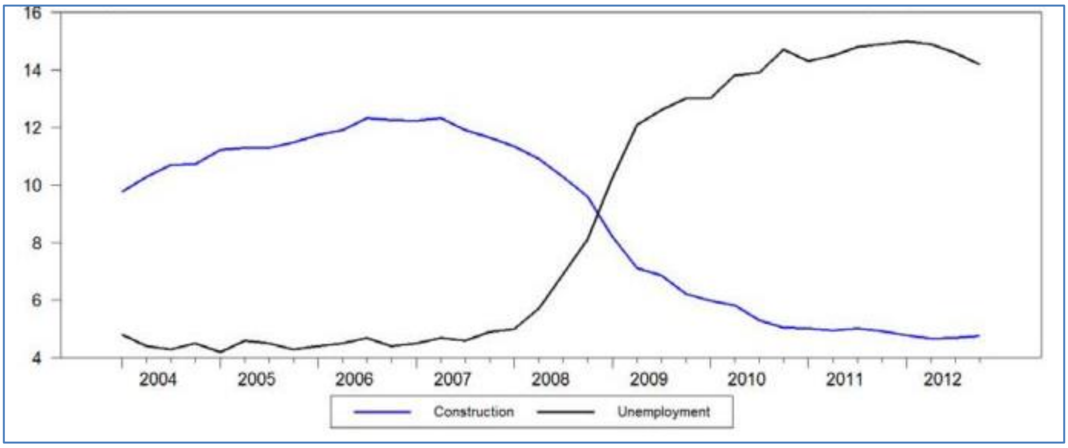


Figure 2.4 - Fraction of Labour Force in Construction or Unemployment (Whelan, 2013)

The government of the time were warned frequently in the years before the crash about the risk of house prices overtaking rates of economic growth. As early as 1999, an Organisation for Economic Co-operation and Development (OECD) report criticised the government for promoting house purchases and suggested the implementation of property taxes or other deterrents to slow down the property market (Casey, 2014). However, it is difficult for a ruling government to make potentially unpopular decisions for the benefit of economic stability but at the risk of political ruin.

Both the economy and the property market have begun to make a recovery, although the national house price index in January 2016 was still 33.8% lower than the peak in 2007[3]. However, a more gradual recovery may result in a more sustainable and diverse economic climate in the future, rather than one overly reliant on a single market.

Understanding these cycles in the property market is the key to limiting the impact of such fluctuations and to informing future government policy in order to avoid repetition of the

[2] The Irish banking crisis began in 2008, when many of the nation's banks and other financial organisations found themselves at risk of insolvency, due to the collapse in the Irish property market. In response, the government of the time provided a state guarantee for existing and future liabilities, and a €63 billion bailout. Although the banking crisis cannot be entirely blamed for the necessity to ask the International Monetary Fund (IMF) for assistance in 2010, it was certainly a contributing factor (Whelan, 2013).

[3] Central Statistics Office. (2016, Feb 29). *Residential Property Price Index*. Retrieved from Central Statistics Office: http://www.cso.ie/en/releasesandpublications/er/rppi/residentialpropertypriceindexjanuary2016/

mistakes made during the Celtic Tiger years. It has particular relevance to the PRA, since the state of the property market has the capacity to directly influence application intake.

2.3.2. Urban/Rural Divide

The Irish property market is characterised by a marked regionality. Whereas once Ireland was a predominately rural nation, since the mid-1950s the gradual but consistent trend towards urbanisation has increased. In 1961 less than 50% of Irish people lived in cities compared to over 62% in 2011 (Central Statistics Office, 2012). Some estimations suggest that greater than 50% of the population will have migrated to the Greater Dublin Area by 2070 (Sirr, 2014). This general movement towards towns and cities, which is echoed in most developed countries worldwide, is to access employment, educational opportunities and services.

During the Celtic Tiger era, house prices rose faster and higher in cities like Dublin, Galway and Cork, and also in the commuter belt of the greater Dublin Region, than in rural counties. During the subsequent recession, the largest collapse in residential prices was also in Dublin, with apartments falling by 52% over four years from 2007-2011. The rest of Ireland suffered a slightly less extreme decline, at an average of 35% down from 2007 prices[4].

In terms of economic recovery, there is also a regional divide. The construction industry was distributed widely around the island of Ireland before the crash, and this sector has yet to fully recover any significant momentum. In 2006, it was estimated that 15% of the workforce outside the Greater Dublin area were employed in construction, compared to 10% inside the region. The economic recovery is coming from other areas of industry, and economists have identified that urban areas are more likely to attract foreign investment and multi-national companies (Goodbody Stockbrokers, 2013).

Likewise, the recent increase in house prices has been more significant in urban centres, where demand is higher. The Goodbody Stockbrokers (2013) report showed that while on average house prices in Ireland rose in the second quarter of 2013 for the first time since 2007, this was disproportionately skewed by the Dublin market, where prices rose by 11% over a year. Outside of the Dublin area, in 2013 house prices were still in decline.

[4] Central Statistics Office. (2011, May 13). *Residential Property Price Index March 2011.* Retrieved from Central Statistics Office: http://www.cso.ie/en/newsandevents/pressreleases/2011pressreleases/residentialpropertypriceindexmarch2011/

2.3.3. Tenure

Ireland has higher levels of property ownership than in many other European countries. In 2011, 71% of properties were owner occupied, compared to the EU average of 67%. Other European nations such as Germany have traditionally had a much greater history of long term renting, and in 2011 only 53.4% of properties were owner occupied (National Economic Social Council of Ireland, 2014).

This desire to own property stems at least partially from our history. In the early 20th century, approximately 90% of people in Ireland rented their homes. Much of the housing stock was of poor quality and many people lived in overcrowded, unsanitary conditions. A combination of factors contributed to the change to property ownership over a relatively short period of time, including the clearing of tenements, building of council housing and incentivised sale and taxation schemes (Sirr, 2014). There was also an element of nationalist pride after so many centuries of rule by foreign landlords. As Sirr (2014) states:

> *"For the past century and a half, perhaps longer, the history of Ireland has been about a struggle to own property, to escape from and then avoid the need to rent."*

This traditional desire for independence and security came to a head during the Celtic Tiger years. As the nation rushed to "*get a foot on the property ladder*"[5] and construction increased, the total housing stock rose from 1.4 million homes in 2000 to 1.9 million in 2008. In 2006 alone, 93,000 new houses were completed (Whelan, 2013).

However, this upwards trend was curbed by the recession, as house sales collapsed. From 2011 to 2016, the number of households in Ireland rose by nearly 13%. However, when this increase is analysed by tenure, nearly 11% were rental properties, compared to 2% of privately owned households (Duffy & Fitzgerald, 2012). This trend towards renting appears to be due to a combination of factors; the difficulty of obtaining a mortgage in response to higher personal debt or lack of secure employment, and a lack of trust in the reliability of the property market leading to a reluctance to invest (Duffy & Fitzgerald, 2012).

Also, as before, the trend is more pronounced in urban than rural areas. Duffy & Fitzgerald (2012) point out that between the 2006 and 2011 censuses, the rental rate in Dublin City rose from 19.2% to 37.5%, compared to an 8.6% increase nationwide. Whether this is a long-term

[5] During the Celtic Tiger years, this phrase was often repeated by politicians and media alike. Young people were consistently advised to "get their foot on the ladder" as quickly as possible as house prices soared. Parents were also encouraged to assist their children in buying their first home; for example in this article from the Irish Times in 2002 - http://www.irishtimes.com/business/it-s-a-family-affair-for-first-time-buyers-1.1090620

trend or a temporary reaction to the recent property crash remains to be seen, particularly as the cost of renting has increased greatly in recent years.

2.3.4. Availability of Data

Until recently, there has been a dearth of official records and analysis pertaining directly to the property market in Ireland. The Property Price Register operated by the Property Services Regulatory Authority (PSRA), which records the sale price of each residential property sold in the state, was only established in 2010. Since 2005, house price indices by quarter have been available from the CSO, and a similar index was issued by the Economic & Social Research Institute (ESRI) for the years 1996-2005 (Lyons, 2015). However, both of these indices draw their data from mortgage transaction returns issued by financial institutions such as Permanent TSB, and therefore give a broad indication of property sales rather than an accurate dataset suitable for analysis.

Prior to 1996, available data is even less reliable. Lyons (2015) attempts to construct a housing price index for Dublin for the twentieth century using other sources, which includes trawling through Irish Times newspaper property listings to record the list prices of residential properties, conducting searches in the Registry of Deeds and accessing mortgage transaction data released by the Department of the Environment.

While such alternative sources can provide useful insight into historical markets, the data is incomplete and cannot be relied on as a legitimate source on which to base predictions for future market activity. It is therefore imperative that accurate, complete property data continues to be recorded by the PSRA for use by economists and analysts in the future.

Synthesising the conclusions drawn by these commentators, the key message emerging from the literature is that the Irish housing market can be characterised as cyclical and heavily influenced by location and economic fluctuation. The following section explores how these key influencing factors operate in housing markets both in Ireland and the rest of the world.

2.4. Factors that Influence Property Markets

Certain key factors can be identified from the literature gathered that have the capacity to influence property markets. The factors fall into three main categories – Economic Factors, Location and Seasonality.

2.4.1. Economic Factors

The literature explored shows that various economic factors have been analysed and tested as potential drivers of the property market. These include GDP, rates of inflation, interest rates, population demographics and supply and demand. A selection of the results are discussed herein.

GDP, or Gross Domestic Product, is used by economists as a measure of the health of an economy. GDP represents the aggregate value of all of the goods and services manufactured during a particular date range, and is usually expressed in terms of percentage change from one time period to another[6].

Valadez (2011) explores the relationship between property prices and GDP over the course of a period of financial instability in the USA. The dataset covers the years of the banking crisis in the USA from 2005-2009, and consists of quarterly house price indices and GDP returns. This data was plotted using linear regression and tested for significance at a level of 0.2.

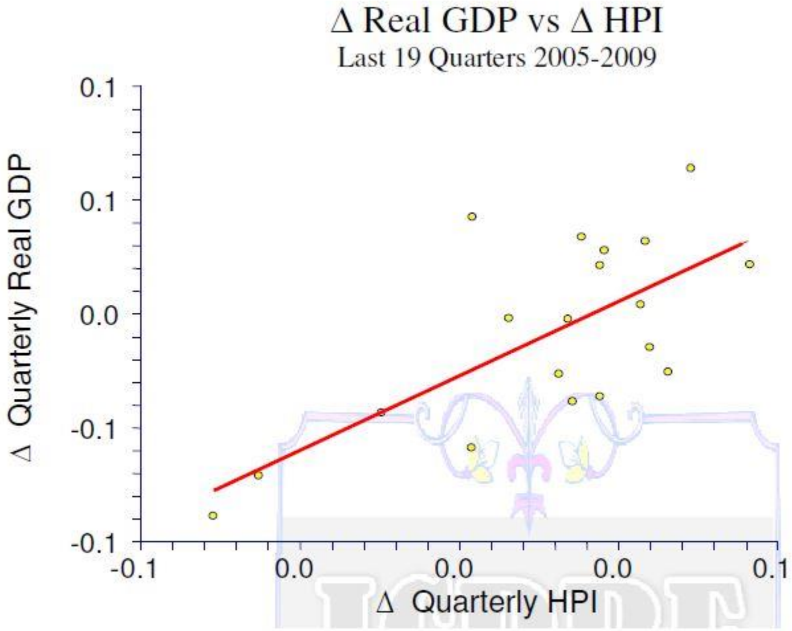


Figure 2.5 - Correlation between House Prices & GDP (Valadez, 2011)

Valadez finds that there is indeed a relationship between GDP and the house price index, and that it can be quantified:

> *"There exists a relationship between the two variables such that a quarterly change in the housing price index may yield a quarterly change in Real GDP".*

[6] Callen, Tim. (2012, Mar 28). *Gross Domestic Product: An Economy's All*. Retrieved from International Monetary Fund: http://www.imf.org/external/pubs/ft/fandd/basics/gdp.htm

However, this correlation merely highlights a relationship, and does not necessarily indicate causality. It may be that rising house prices stimulate the economy and lead to higher GDP, or that a more buoyant economic scene encourages investment in property and increases demand, thus pushing prices higher. Alternatively, there may be other external driving forces that influence both factors.

Demary (2010) performs a more thorough exploration of the relationship between house prices and economic indicators. The dataset assembled from the OECD's databank resource includes house price indices, GDP and interest rates for a range of countries during the period 1970-2005. Demary identifies three economic trends and analyses the dataset to evaluate if the occurrence of these events correlates with changes in house prices. The trends are: increases in interest rates ('interest rate shock'), increases in demand for labour due to increased demand for products and services ('output shock') and increases in rates of inflation ('inflationary shock').

After performing regression analysis, the results are more or less as one would expect. Output shocks are found to have a positive impact on house prices, while interest rate shocks and inflationary shocks have a negative impact. However, this study does reinforce the existence of a relationship between the various factors, and also indicates a converse relationship. As well as being affected by the economy, Demary's results show that increases in demand for housing influences the economy by increasing interest rates and inflation.

Engsted & Pedersen (2014) assess the contribution of interest rates in OECD countries during the late 1990s and early 2000s towards stimulating the housing boom and potentially causing the subsequent recession. The paper finds that relaxed credit requirements probably have a greater impact than interest rates alone, although they may indeed have been a stimulus to the improving property market of the early years of the boom. However, interest rates did not increase significantly during the global recession of the late 2000s:

> *"The decreasing house prices after 2006 in many countries are more naturally explained by higher risk-aversion and tightening of credit constraints following the general economic downturn, especially the global recession beginning in 2008."*
> (Engsted & Pedersen, 2014)

Outside of the OECD, China has experienced massive economic growth since the Reform Period began in the late 1970s, and the housing market has also undergone a period of concurrent change[7]. Because of this, several interesting studies have focused on the property market in China in recent years.

Demographics will always be relevant in any analysis of economic activity in China, since the one-child policy introduced in 1979 has artificially skewed the natural distribution of age profiles across the region ever since. Li (2014) seeks to evaluate the impact of an aging demographic on the housing market and consequentially on the economy. The research contrasts activity in two areas; China, which has seen a fall in population growth due to the one-child policy, and the USA where the 'baby boom' generation are now heading towards retirement. The models created show that the profile of the population is statistically significant: "*a 1% increase with the people between the ages of "15-64" causes a 2.707% increase in the housing prices*" (Li, 2014). Furthermore, the results show that increases in income level or GDP result in higher house prices, while higher unemployment and increased housing stock contribute to prices falling.

Demographic profiling also plays a part in a study conducted on United Kingdom survey data to determine if there is a causal relationship between 'consumption' – the rate of growth in consumer spending – and house prices (Attanasio *et al.*, 2009).

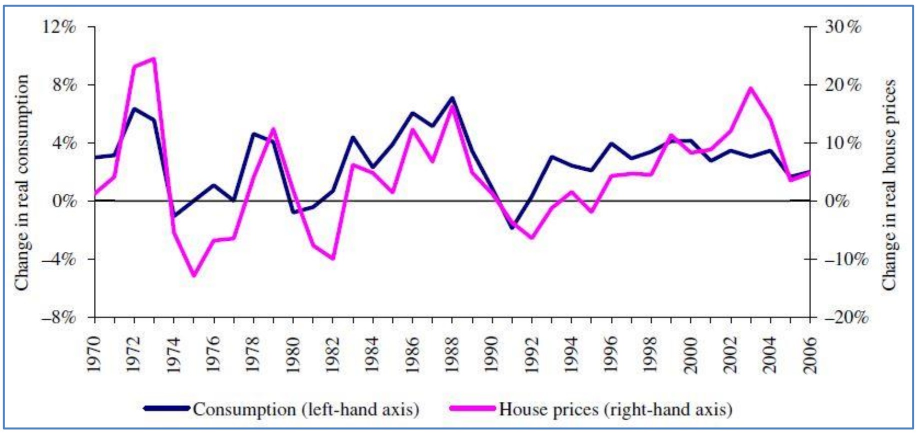


Figure 2.6 - Annual UK real house price and consumption growth (Attanasio *et al.*, 2009)

This paper proposes three variant explanations for the correlation between the two trends:

1. That an increase in house prices leads to the householders having more spare funds, and thus allows them to spend more;

[7] The 'Chinese Economic Reform Period' refers to a long-term program of economic reform instigated by the Communist Party of China. Under this program, agriculture was decollectivised, previously closed markets were opened to foreign investment and the private sector grew significantly, with increases in GDP averaging 10% per year. The World Bank. (2016, Apr 6). China: Overview.

2. An increase in house prices improves the owners' credit score and thus allows them increased access to borrowing from lending institutions and enables them to spend;
3. That both house prices and consumption are influenced by the same external factors, and thus trends follow the same patterns.

Using life cycle prototypes to model the data, the team find that house prices and consumption are most closely linked for younger householders, with the relationship between the two features weakening as the age increases.

Returning to China, Hin Li & Lin Ge (2008) examine the impact of inflation on the Shanghai housing market between 1997 and 2005. Despite its geographic distance from Ireland, several comparisons can be drawn between the property cycles experienced in both locations during this time frame. As in Ireland, Shanghai experienced a property boom which was largely stimulated by policies and incentives implemented by the government. The central bank operated a relaxed approach to credit and loans, and interest rates were low. In response, annual new house completions soared from 5,291 in 1995 to 30,762 in 2004 and rates of home ownership rose concurrently.

After performing regression analysis, the paper finds that while rates of inflation can impact on housing prices, an imbalance between supply and demand actually has a greater bearing on residential property. If incentives such as low interest rates are offered to potential buyers, then an insufficient supply of suitable properties will inevitably lead to higher prices. This trend can also be seen in Hong Kong, where a severe shortage of housing has led to several dramatic price increases between 1980 and 2000 (Ge & Lam, 2001).

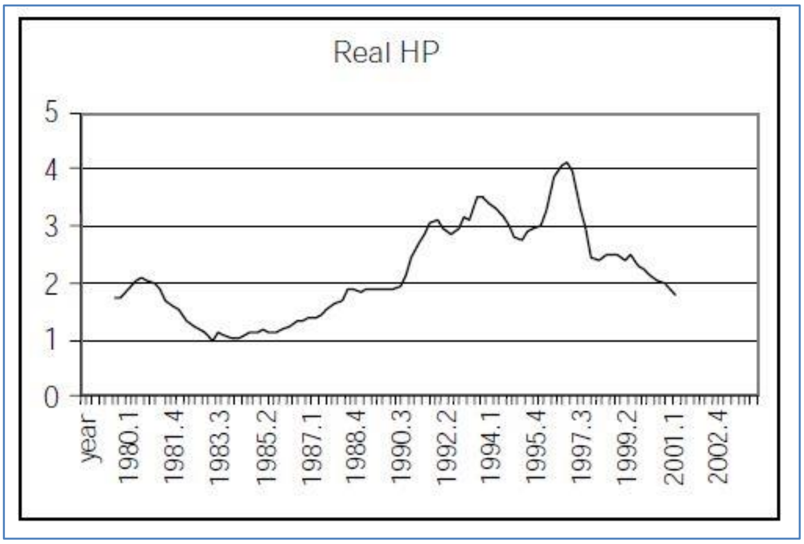


Figure 2.7 - Residential Housing Prices Index in Hong Kong (Ge & Lam, 2001)

These peaks and troughs are due to the volatility of the Asian financial markets. Just as Ireland placed too much emphasis on the construction sector during the Celtic Tiger years, Hong Kong relies heavily on the financial sector as its principal source of employment and income. Any financial crisis - such as that which occurred in 1997 - has an instant effect on property sales and prices. The regression analysis performed shows that economic fluctuation and household income have the greatest impact on house prices, followed by changes in the population in the 20-59 years demographic. However, the author points out the limitations of the linear regression approach in that it can be adversely influenced by multi-collinearity. The use of artificial neural network models is suggested as an alternative approach.

The impact of government incentives is further explored by Snyder & Ekmekjian (2013), who evaluate the impact of a tax credit introduced in the USA in 2008 to stimulate the stagnant housing market. The findings show that while the tax credit cannot be proven to have had a major impact on house prices or sales, it did assist house buyers in terms of supplementing their overall income and thus may have contributed indirectly to steadying the property market.

All of these studies thus far have analysed and mined historical data to extract insight and draw conclusions. Taking this a step further, Li & Leatham (2013) use Dynamic Factor Models (DFMs) and Large-scale Bayesian Vector Autoregressive (LBVAR) models to attempt to forecast the future growth rate of the housing market. The housing market is "*illiquid and heterogeneous in both physical and geographical perspectives*" (Li & Leatham, 2013), which makes it a challenging task.

Using a dataset of house prices indices and economic indicators from 1981-2010, a series of models are built and the results compared to a test set. The findings show that the DFMs consistently outperform the LBVAR models, but that as with all variants of linear regression, the models are only applicable in cases where datasets follow a linear structure.

Dynamic Factor Models

- Dynamic Factor Models (DFMs), also sometimes called 'Diffusion Indexes', are used in macroeconomic analysis to model time-series data and identify turning points and trends in cycles (Y. Li & Leatham, 2013).
- Analysts in this sector often encounter datasets with many variables but limited time series observation instances. For instance, a dataset may cover 20 years, but only contain data on a quarterly basis, thus resulting in only 80 individual observations. Analysis of such datasets using time series models can be difficult, since most algorithms – such as LBVAR models as discussed below – can only process up to ten variables (Stock & Watson, 2002).
- DFMs offer a solution to this problem, since they have the ability to "*model simultaneously and consistently data sets in which the number of series exceeds the*

number of time series observations" (Stock & Watson, 2010). They do this by assessing the entire dataset and identifying the small number of 'dynamic' factors that cause movement and change in economic cycles, and also any "*mean-zero idiosyncratic disturbances*" such as outliers or measurement errors that can cause skew in the analysis (Stock & Watson, 2010).

- Thus, the principal benefit of a DFM is that it can potentially explain the variance in a series of data using only a small selection of factors.

Large-scale Bayesian Vector Autoregressive (LBVAR) Model

- Like DFMs, Vector Autoregression (VAR) models are used in macroeconomics to forecast trends and analyse the structure of complex datasets. As Hyndman & Athanasopoulos (2013) discuss, a VAR model typically consists of a series of equations, one per variable. Each equation assesses the extent of the influence of the other coefficients on this variable and outputs a forecast. Therefore, a series of forecasts is generated in a recursive fashion.
- The benefit of VAR models are that they allow for the input of multiple parameters and therefore have the potential to identify complex relationships between elements in a dataset. However, this high level of generality can also be a disadvantage, as it can be expensive and difficult computationally, and so VAR models are generally restricted to a small number of variables (Bańbura, Giannone, & Reichlin, 2008).
- LBVAR models therefore represent an enhancement of this basic VAR model. By applying principles of Bayesian probability to the VAR framework, it can then be used for multiple time series variables (Bańbura et al., 2008).
- The enlargement of these principles and techniques is outside the capacity of this study, but does represent an interesting avenue for further research in the future.

2.4.2. Location

There is no question that location is an influencing factor on housing markets across the world. House prices, quantity of sales and square footage vary according to address. This is not a new trend; in 1975 Straszheim asserted that "*variation in housing characteristics and prices by location is a fundamental characteristic of the urban housing market*" (as cited in Goodman & Thibodeau, 2003). This variance can be due to proximity to economic centres or essential services, or other desirable factors.

In terms of prediction models, the presence of a particular valued amenity such as a well-regarded school can introduce skew into house price datasets within a relatively small geographic area. Clapp, Kim, & Gelfand (2002) explore the influence of location on house prices, and the inherent difficulty in predicting future prices based on such “unmeasured spatial variables”. Many conventional hedonic pricing models find this variance difficult to handle effectively.

A *hedonic pricing model* is an economic modelling technique used widely in the property marked to estimate pricing or demand. It operates on the assumption that the various attributes of a house can be classified as positive and negative, and that by quantifying the extent to which these features impact on price, the pricing of other properties can be inferred (Cebula, 2009). Hedonic pricing models are discussed further in the Section 3.5 Modelling Approaches.

As a solution to the limitations of conventional modelling techniques, Clapp, Kim, & Gelfand (2002) use semi-parametric modelling with a Bayesian framework to create more accurate outputs. The factors that cause skew in a series of small towns within the same region of Massachusetts are identified. The age of the property is found to have a significant impact, with older properties commanding higher prices – interestingly this is found to be a non-linear relationship, and therefore might not be identified using standard linear regression techniques. Distance from the popular historical centre of the town of Lincoln has a positive impact on house prices, while proximity to a local dump or an airbase has a tendency to depress prices. When these factors can be identified and included in pricing models, a more accurate picture of the local property marked can be deduced.

A Note on Semi-Parametric Modelling with a Bayesian framework

Semi-Parametric Modelling

- In general, statistical methods for modelling data can be classified into two major categories; ‘parametric’ and ‘nonparametric’. The parametric method uses algorithms that assume that data follows a distribution based on a fixed set of parameters. The nonparametric method does not assume that the data follows a set statistical model and instead uses the input data to identify patterns (Han, Kamber, & Pei, 2012, p. 553).
- A ‘semi-parametric’ model falls between these two extremes, and may take characteristics of either. For example, a standard regression model might take the form:

$$Y=\beta_{T}X+g(Z)+\epsilon$$

In this instance, βTX is the standard parametric regression model, but with the addition of $g(Z)+\epsilon$, which represents unknown parameters that could be inferred from the dataset[8].

Bayesian framework

- A Bayesian framework relies on the principles of Bayes Theorem, which evaluates the probability of an event (P[H/X]), based on information about other events.

$$P(H|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|H)P(H)}{P(\boldsymbol{X})}$$

Figure 2.8 - Bayes Theorem (Han et al., 2012)

- For example, if given the results of a telephone survey asking if voters intended to vote for a new law, and if they support the Republican or Democrat parties, Bayes Theorem could be applied to determine the probability that a selected respondee is a Republican, based on the fact that they are in favour of the law.

Given all of this information, a semi-parametric model with a Bayesian framework may be a model with known modelling functions, but unknown distributions for some of the modelling components, that is used for the prediction of some event based on other events (Chow, Tang, Yuan, Song, & Zhu, 2011).

Kestens, Thériault, & Des Rosiers (2004) also explore the effect that location has on house prices by building a set of regression models, each one based on a different spatial scale. Using data derived from aerial photographs and satellite imagery, the first model considers various factors, including location, census data and accessibility. The second model integrates land-use and vegetation variables, and a third includes car journey time from the property to the nearest city centre.

The results of the research reveal some interesting insights. Vegetation levels can be an important factor – the presence of trees on streets and in gardens indicates higher prices, while conversely woodland is associated with lower sale figures. Properties within a certain distance of water features such as rivers and lakes showed a positive price correlation, while proximity

[8] Darmon, David. (2006, Oct). *Parametric, Semiparametric, and Nonparametric Models*. Retrieved from thirdorderscientist: http://thirdorderscientist.org/homoclinic-orbit/2013/10/6/parametric-semiparametric-and-nonparametric-models

to agricultural or industrial activities had a negative impact. Furthermore, the results showed that the effect of distance between a property and economic centres is not linear. House prices declined as distance from the centre increased, but only until a distance of approximately 17 minutes by car from the centre was reached. After this point, prices began to increase once more, suggesting a ring of less desirable properties in the inner suburbs.

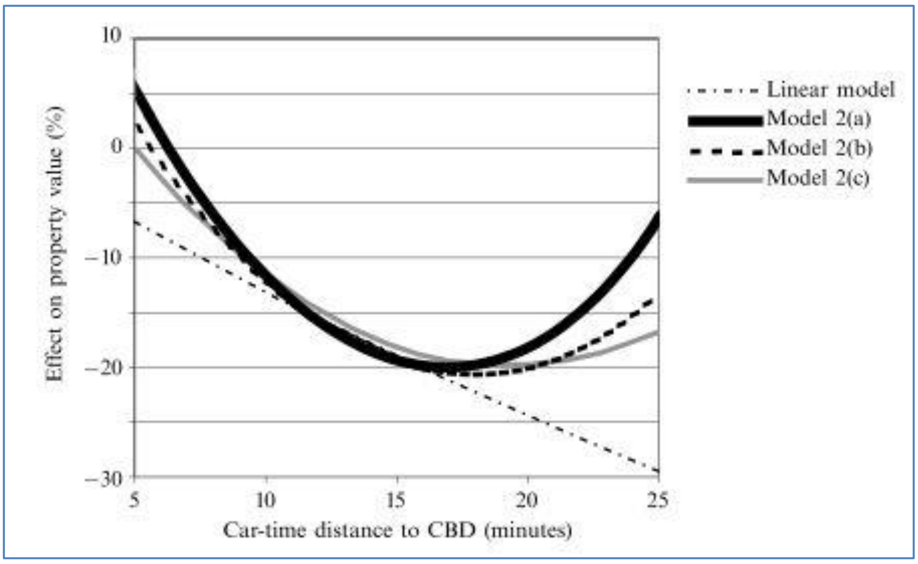


Figure 2.9 – Relationship: car time distance to economic centres & house prices (Kestens, Thériault, & Des Rosiers, 2004)

However, when other factors are also taken into account – such as the density of trees and proximity to other land-uses, as already discussed – the trend flattens out and "*the U-shaped curve becomes a rather linear trend*". This highlights the need to consider multiple factors at once in terms of property transactions, and the dangers of relying too heavily on analysis of individual factors.

While Kestens *et al.* (2004) concentrate on a specific region of Quebec City, Rapach & Strauss (2009) take a wider view by analysing the United States as a whole. This study compares house price forecasts for the years 1995-2006 across the US against the real sale price figures. They discover a wide difference in accuracy between the predictions made for interior and coastal states. Whereas models that integrate economic features were relatively correct for many internal states, the same models did not perform well with coastal locations. This suggests that whereas the housing markets in the central areas of the United States are dependent on the economy, this is less true for the coastal states. Therefore, while modelling house sales in Ireland, it may be interesting to investigate the impact of economic change on coastal counties compared to midland regions.

2.4.3. *Seasonality*

The property market in most if not all nations is seasonal – there are certain times of year more prone to periods of increased market activity, and other times of near stagnation. For example, in the USA and the UK, a 'hot season' of high sales has been identified in the second and third quarters of a year, and a 'cold season' in the first and fourth quarters (Ngai & Tenreyro, 2009). While the exact timing of these peaks and troughs can vary slightly by location, the overall trend remains the same. In the UK, since the early eighties both the number of transactions and the average sale price of properties have increased in the summer months, and dropped back again in the winter.

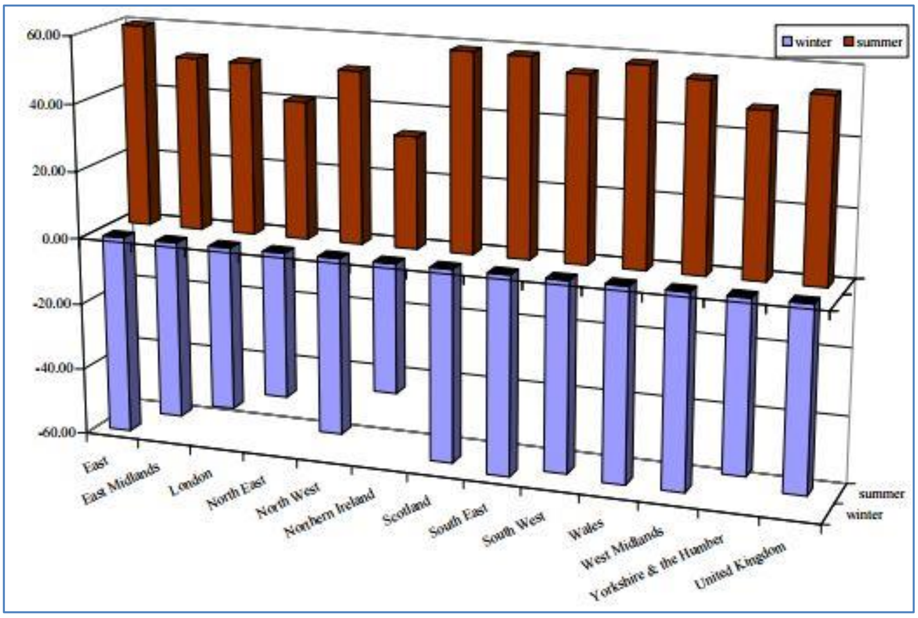


Figure 2.10 - Annualized growth rates in transactions in U.K. & its regions, 1983-2007 (Ngai & Tenreyro, 2009)

Likewise, Cebula (2009) finds some seasonal variation in the housing market in Savannah, Georgia, with house sales closing in the months of May or July tending to achieve higher prices.

However, the exact cause of this seasonal variation is not made clear – is it a factor on its own, or does it operate in conjunction with contributing factors? For instance, perhaps more houses are sold in the summer because lending institutions grant more loans in the early part of the year, with clean loan books at the start of a new fiscal year. Or, perhaps less transactions complete in December because this is traditionally a time of less activity for solicitors and other legal professionals.

There also do not appear to have been any equivalent studies performed specific to the Irish market – this could be an interesting and useful avenue for future research.

2.5. Modelling Approaches

The previous sections have reviewed contemporary literature from the perspective of the property market, its principal features and the influences that affect it. In terms of the modelling techniques used to analyse and process property market datasets, a range of approaches are found to have been employed. The most common is linear regression, with occasional occurrences of decision trees and neural network models. This section will briefly review the algorithms used. It will also suggest some techniques employed in other areas of prediction modelling that could potentially be applied to the property market.

2.5.1. Supervised vs. Unsupervised Learning

Machine learning and data mining are two separate fields within the realm of Data Analysis. However, they often employ the same methods and overlap significantly. They can be roughly distinguished as follows:

- Machine learning focuses on prediction, based on *known* properties learned from the training data - this is 'supervised learning'.
- Data mining focuses on the discovery of (previously) *unknown* properties in the data – this is 'unsupervised learning'.

In the case of supervised learning, some key data in the training set is already labelled, and the model 'learns' from this classification and uses it to label unseen unclassified instances (Han *et al.*, 2012*)*. An application of supervised learning would be in the field of spam detection, where a training set is used to compose a list of 'rules' as to the features/combinations of features in a dataset that indicate spam. These rules are then used to analyse new data to identify probable spam emails.

In contrast, where unsupervised learning is concerned, the model is trained without any prior class labels. The aim of the model is to identify potential groups of classes within the dataset, and thus to reveal previously unknown insights (Han *et al.*, 2012). An application of unsupervised learning is in the field of cluster analysis – for example the use of k-means cluster analysis to build a market segmentation of a customer base and thus target marketing campaigns more efficiently (Foreman, 2014).

In short, supervised learning can be viewed as a classification task and unsupervised learning as a clustering task (Manning & Schütze, 1999).

In the case of this set of experiments, the models will be built using supervised learning techniques, making use of insights from the literature and statistics explored in the previous chapters.

2.5.2. Decision Trees

Decision trees are one of the simplest models to train and to interpret, and yet can be a powerful tool in performing analysis on multiple variables.

A decision tree model consists of a 'root node' or starting point, and a series of 'branches' and 'leaf nodes' leading from it. Each branch of the decision tree represents a possible decision or occurrence. Decision trees can be used to assist decision making, to classify data into useful categories, or as an easily interpretable visual communication tool (Quinlan, 2007).

To create a decision tree, the analyst takes a dataset and recursively divides the training data by measuring the most discriminative dividing criteria. When a feature is a numeric value the measurement used is the variance; when it is a category the measurement will be the 'entropy' (entropy is a measure of the amount of 'variance' or uncertainty associated with a set of probabilities). This process is then repeated until there is no statistically significant gain by further split the tree into more branches. The final leaf nodes hold the prediction for the original query.

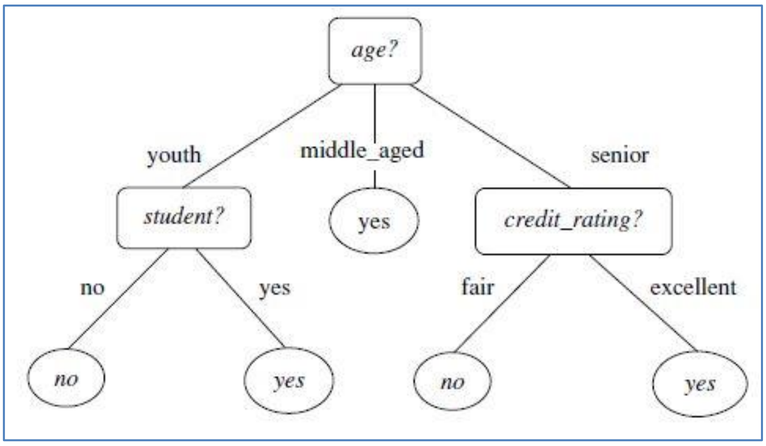


Figure 2.11 - Decision Tree determining customer base likely to purchase a computer (Han et al., 2012)

Decision trees are generally used to identify classes, as in the example above. When the target variable is numeric – for instance when predicting sale price or quantity of sales – then a "regression tree" using a regression algorithm may be employed instead.

The potential benefits of using decision trees are that they are easily read, flexible in terms of the data that can be inputted, and can handle categorical, binary and numeric values. However, they are limited in that they can only consider one attribute at a time, and not a combination of multiple input values. Also, trees cannot be updated if new data is provided; they must be redrawn from scratch each time. Lastly, decision trees can be prone to 'overfitting' the data – i.e. applying too much meaning to incidental occurrences in the dataset (Hawkins, 2004).

Acciani, Fucilli, & Sardaro (2011) compare two separate approaches in building property sale price prediction models to improve identification of niche market segments; decision trees and regression. The dataset used in this study comprises various attributes of a type of farm unique to the south east of Italy, called "*trulli*". These attributes include distance to the nearest town, the size of the farm, the existence of nearby attractions such as tourist centres or archaeological sites, any recent renovations carried out on the farm/buildings and the availability of services such as water and telephone.

The dependent variable is numeric - the sale price per hectare – and therefore the decision tree must take the form of a regression tree. In this instance, the M5P algorithm is employed to generate the tree, which calculates a linear equation at each leaf node. The final pruned regression tree shows that recent renovations (RENOV) and farm size (LDSIZE = size of farm + trullo; ANINDEX = size of farm and any annexes) are the key variables in determining sale price.

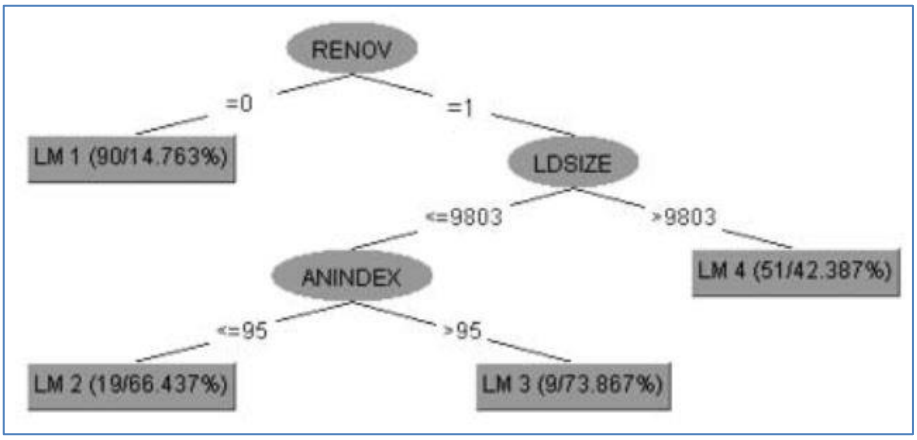


Figure 2.12 - Pruned decision tree (Acciani et al., 2011)

When compared to the results using a standard regression algorithm, the decision trees are acknowledged as successful in identifying previously undetected sub-markets, even with small datasets. Decision trees would therefore appear to be a useful candidate for modelling the quantity of applications lodged, by identifying the factors that exert the most influence over intake.

2.5.3. Random Forests

If overfitting is identified as an issue during modelling of decision trees, then a 'random forest' can be utilised – this involves building a series or 'ensemble' of short stubby regression trees. These trees are individually poor, but when examined as a group, the forest uses averaging to find a natural balance and thus can be quite accurate.

Random forests have the potential to be useful so long as the dataset is relatively small, since large datasets can be slow to analyse from a computational perspective (Hamner, 2010). They also have the advantage of making use of the "wisdom of the crowd" phenomenon; an ensemble of stubby decision trees are created which are individually poor but which together are useful and quite accurate (Ho, 1995).

Hamner (2010) uses random regression forests to predict future quantities of traffic based on current and recent automated traffic recorder readings, with the aim of improving traffic flows. A series of models is built, each with slightly different parameters, and the root mean square error (RMSE)[9] for each is compared. The results are encouraging with up to 40% of variance accounted for by the model, but the author stresses the need to include other external features in order to improve this score further. Also, for application to a larger dataset it may be necessary to experiment with boosting, feature selection or other techniques in order to facilitate faster processing.

2.5.4. Linear Regression

In comparison to decision trees and their information based approach, regression algorithms take an 'error-based' method to training models. The concept of a regression algorithm is to create a prediction model with a set of random parameters. It is assessed to judge how well it performs. Based on this error rate, the parameters are iteratively adjusted to fine-tune the model until it has reached optimum accuracy (Goldberger, 1962).

A linear regression model assumes that there is a linear relationship between the different features and that this relationship can be captured and used to make predictions. If a relationship does exist, then it can be quantified using the formula for the equation of a line ($y=mx + b$).

[9] The Root Mean Square Error translates to "*the square root of the mean/average of the square of all of the error*." The use of RMSE as an error metric is quite common, as it is suitable for numerical predictions. Kaggle. (2016, Jan 18). Root Mean Squared Error (RMSE). Retrieved Kaggle: https://www.kaggle.com/wiki/RootMeanSquaredError

As an example, a sample dataset might show office space rental information. Size is a descriptive feature and rental price is the target feature that we want to predict. When the data is plotted on a scatter plot, there appears to be a linear relationship; for every increase of 1 sq. metre in size, office rental prices increase. Using linear regression, this relationship can be quantified and the data can be used to calculate the expected rental price of future office spaces.

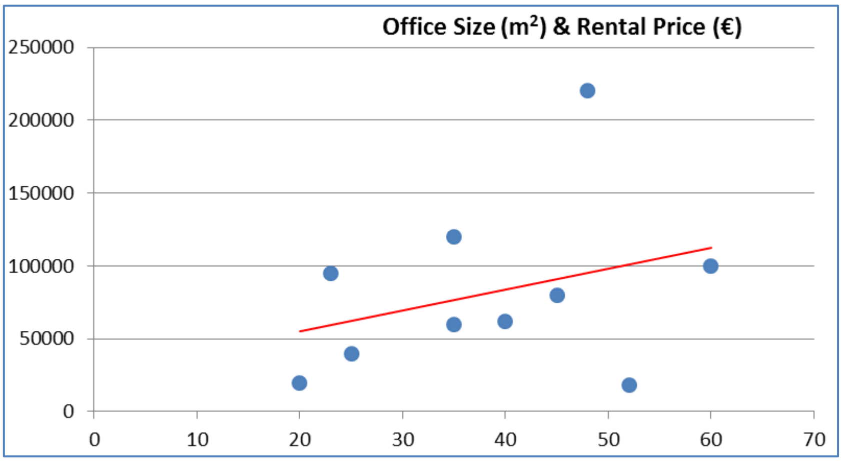


Figure 2.13 - Scatter plot showing linear relationship

More complex problems with multiple descriptive features can be handled using a combination of multivariate linear regression and gradient descent.

The key to getting results from linear regression is to determine the optimal values of the weights to apply to different features, in order to best capture the relationship between the target feature and the other descriptive features.

In attempting to predict house prices or quantity of house sales, most analysts begin with regression analysis, since it is simple, easy to interpret and can be used in conjunction with other techniques. Furthermore, the standard method of predicting future house prices or demand is the use of hedonic pricing equations, which effectively constitute a form of regression analysis.

As already mentioned, a hedonic pricing model classes the various attributes of a house as either positive and negative and analyses the extent to which these features impact on price. (Cebula, 2009) defines a sample model as:

$$ln(RSALESPRj) = f(Ij, Ej, SCj, Oj)$$

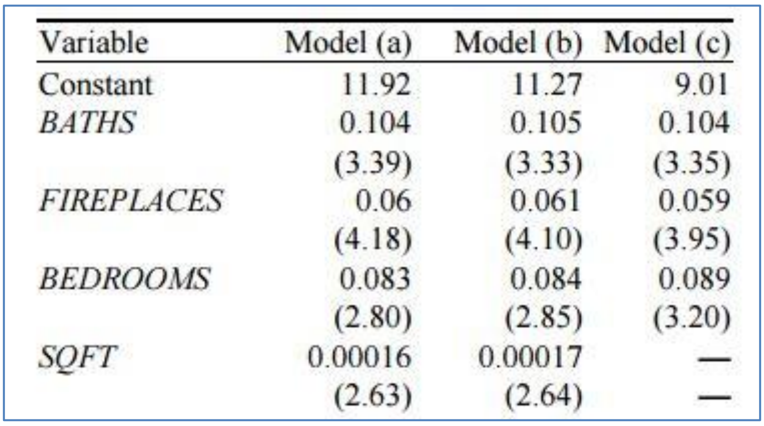
where:

- ln(RSALESPRj) = the natural log of the real price of house j in 2005 dollars

- Ij = a vector of interior physical characteristics for house
- j; Ej = a vector of external physical characteristics for house
- j; SCj = a vector of spatial control variables for house j
- Oj = a vector of other factors associated with house j.

A series of 25 attributes are identified which may be important factors. Internal features include the number of bedrooms and bathrooms, the presence of a fireplace and the square footage of the property. External characteristics might comprise the inclusion of a deck, garage or pool, and the structure of the house – for instance if it is built from brick. The spatial characteristics concern the location of the property – if it is on a corner, for instance, or near a park, in a cul-de-sac or near apartment blocks. Finally, 'other factors' would include important features that do not fall into the categories above. In the case of Cebula's study in Savannah, Georgia this category includes properties designated as houses of historical importance.

Three models were built using regression analysis, each including slightly different combinations of attributes. The first model includes all 24 attributes, with 15 of these found to be significant at a level of 0.01, the second model excludes two attributes, and so on. The regression algorithm applies weights to each variable, which are collated in a table. An extract from this table is shown below:



Variable	Model (a)	Model (b)	Model (c)
Constant	11.92	11.27	9.01
BATHS	0.104	0.105	0.104
	(3.39)	(3.33)	(3.35)
FIREPLACES	0.06	0.061	0.059
	(4.18)	(4.10)	(3.95)
BEDROOMS	0.083	0.084	0.089
	(2.80)	(2.85)	(3.20)
SQFT	0.00016	0.00017	—
	(2.63)	(2.64)	—

Figure 2.14 - extract from hedonic pricing model (Cebula, 2009)

The final conclusions show that houses in this area of Georgia are positively influenced by many of the features explored, including the number of stories, bedrooms and bathrooms, the house being new or second hand and situated in a cul-de-sac or near a park or river. Negative influences were identified as vicinity to apartment blocks or streets with high traffic.

There are many other instances of linear regression being used as a method of analysing housing market fluctuations. For example, Rousová & van den Noord (2011), Valadez (2011), Bin (2004), Demary (2010) and Ge & Lam (2001) all use regression algorithms to build models that analyse the factors that influence the frequent peaks and troughs in the housing market.

While some of the factors identified as significant in these experiments may seem obvious – for example square footage, or number of bedrooms - what is extremely useful is the calculation of the weighted extent to which each feature affects house prices. In the same way, if the dataset the subject of this dissertation is found to have a linear profile, then regression can potentially be used to estimate the extent to which various factors influence quantity of house sales.

2.5.5. Artificial Neural Networks

Another potential approach is the use of artificial neural network (ANN) to generate predictions (McCulloch & Pitts, 1943).

An artificial neural network (ANN) is a collection of input/output units which are connected, with each connection having an associated weight (Han *et al.*, 2012). As the model is trained, the weights applied to each connection are adjusted systematically and "tuned" to correctly predict the correct classification for new instances.

The disadvantages of neural networks lie in the time it takes to train the models, and in their difficulty of interpretation. However, they can also be more tolerant of noisy data than other algorithms, and are widely used in data mining to reveal hitherto unknown relationships between data (Han *et al.*, 2012).

There are many different neural network algorithms, but one of the most popular is "backpropagation". The backpropagation algorithm uses a multi-layer feed-forward structure; it consists of an input layer, one or more hidden layers, and an output layer.

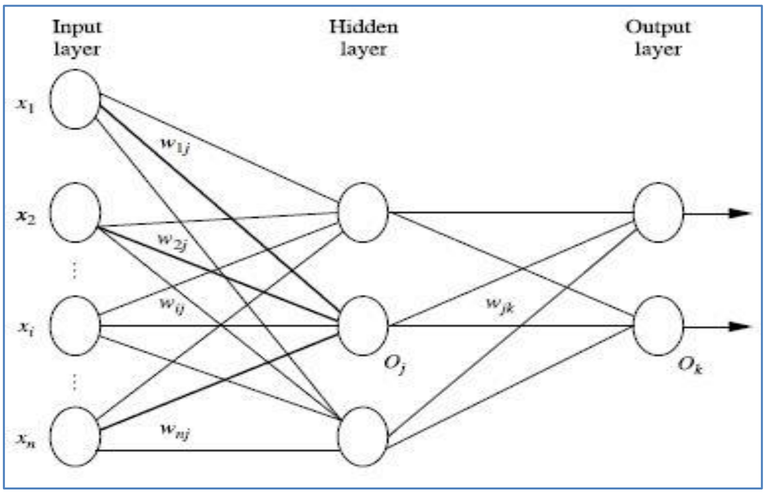


Figure 2.15 - Illustration of a multilayer feed-forward neural network (Han et al., 2012)

The input layer takes the raw data, calculates the appropriate weighted values and forwards the results to the next 'hidden' layer (Leahy, 2014). There can be only one hidden layer, or multiple layers, but each one applies a function to the data to iteratively adjust the weighting. Eventually, the final data is sent to the output layer, which provides the network's prediction for the data provided (Han *et al.*, 2012).

In the context of the housing market, several studies have used neural networks to identify the key determinants of property prices and use this information to predict future pricing trends. For example, Selim (2009) focuses on the Turkish housing market, and promotes the use of neural networks in place of traditional regression models. The dataset in this instance is extracted from household budget survey data for the year 2004 and 46 variables are identified as potentially influential. Selim highlights the risk implied in relying on generic hedonic regression, since the data may not follow a linear structure. In order to test this theory, two models are built – one using traditional hedonic techniques and the other a feed forward neural network model.

When performance of both models is compared to the actual prices achieved, it is clear that the artificial neural network outperforms the hedonic model quite significantly. The Root Mean Squared Error (RMSE) for the neural network is 0.66, compared to 1.57 for the hedonic model.

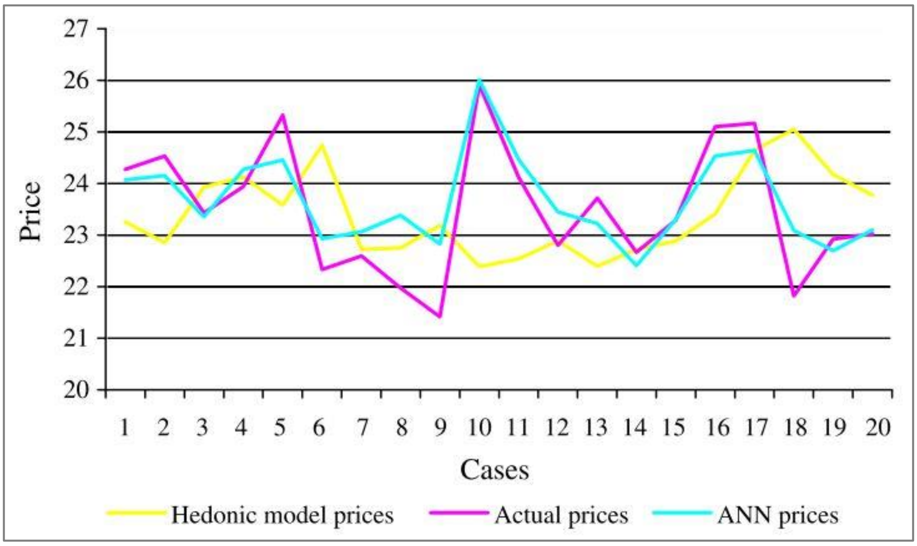


Figure 2.16 - Actual prices compared to hedonic & neural network model (Selim, 2009)

In another application of neural networks, Khalafallah (2008) builds models to gain insight into the current and historic housing market in the United States, and then to predict the ratio between a property's asking price and eventual sale price. After experimenting with various calibrations, the optimum model is identified and applied to unseen examples to test its success.

The model is found to be quite accurate, with an error rate between -2 and +2%. Khalafallah notes that that this success is likely to be restricted to predictions of less than six months, and that further work is required to extend the accuracy of long term predictions. However, this paper shows the potential of neural networks to be applied to the prediction of housing markets, and may also be relevant in attempting to predict future house sales in Ireland, and therefore also prediction of application intake to the PRA.

2.6. Key Findings

A wide range and breadth of literature has been explored in the course of this chapter. The key findings are outlined below.

2.6.1. The Irish Housing Market

The Irish housing market is cyclical in nature, with a number of peaks and troughs evident in both our near and distant past. The market has traditionally been intertwined with the wider economy – when the economy booms, houses are built, prices rise and borrowing increases. Conversely, when a recession sets in the housing market stagnates and many properties fall into negative equity. What is not entirely clear, however, is the extent to which this relationship is causal; whether strong economic growth stimulates the housing market, or a healthy housing market encourages the economy, or if both are subject to other influences.

The Irish housing market is also characterised by a marked urban/rural divide, with increasing urbanisation and rural depopulation. Prices are higher in urban regions, but these areas also appear to be more susceptible to economic booms and busts. Finally, the Irish housing market is marked by high rates of freehold tenure when compared to other European countries, with a high proportion of Irish homes being owner occupied.

2.6.2. Factors that Influence Property Markets

The factors identified as significant in the literature reviewed are as listed in the following table:

Factor	Notes
Gross Domestic Product	GDP, as a measure of the health of an economy, appears to have some correlation with the housing market, although the extent to which this is true varies from one academic paper to another.

Inflation Rates	Increases in inflation rates have been shown to have a negative impact on the market.
Interest Rates	Lower interest rates tend to stimulate the market, while higher rates depress it.
Population Demographics	As more and more people flock to towns and cities, the more rural counties are becoming depopulated. This appears to have a negative impact on the housing market in these areas, while prices tend to increase in more densely populated regions.
Unemployment	High rates of unemployment are associated with a stagnant property market – without a reliable income prospective buyers have less disposable income and less access to loans and credit.
Location	From the analysis of literature, location would appear to be the single most important factor in property transactions.
Property Attributes	The type, size and age of buildings is relevant, as well as decorative order and extras such as gardens, garages or pools.
Season	Month appears to be a contributing factor, with more sales in the summer months than in winter.

Table 2.1 - Factors that influence the housing market

2.6.3. Modelling Approaches

The modelling approaches discussed in this chapter have included a range of supervised machine learning techniques: decision trees and random forests, linear regression and artificial neural networks.

Each modelling technique has its advantages and disadvantages. For example, decision trees are relatively straightforward to train and to interpret and are relatively resistant to outliers, and yet can be a powerful tool in performing analysis on multiple variables. While usually used for classification purposes, they can be adapted to also perform with a continuous target variable.

Regression is a powerful tool for generating predictive output; however, it does require an assumption of linearity. It is also the method behind the model commonly employed in house price prediction, in the form of hedonic pricing equations.

Finally, while neural network models can be difficult and time-consuming to train, they are powerful tools in revealing unseen relationships, and offer a useful alternative to regression when elements of a dataset are non-linear.

2.7. *Conclusions*

A wide range of literature has been presented in the course of this chapter, covering cycles and trends in the property market both in Ireland and abroad. The factors that influence house prices and sales have been identified, and various approaches to modelling evaluated. The next chapter will focus on the Property Registration Authority; its functions, goals and the challenges faced in the context of the current economic climate.

3. THE PROPERTY REGISTRATION AUTHORITY

3.1. Introduction

The aim of this chapter is to set the context for the study, providing the reader with an insight into the functions and responsibilities of the Property Registration Authority, its aims and future goals, and the obstacles that must be overcome in order to achieve these targets.

3.2. Overview of Property Registration in Ireland

There are two systems of land registration in Ireland which operate concurrently; the Land Registry and the Registry of Deeds.

The Land Registry operates a registration of title system and maintains a 'Land Register' recording the ownership of land and other rights through a combination of a textual document (the 'folio') and a map recording location. The Land Registry holds title information dating from the year 1892[10].

The Registry of Deeds operates a registration of deeds system, maintaining a combination of manual, micro-film and electronic documents that record the existence of deeds affecting property. Registry of Deeds records date from 1708[11].

These two systems operate concurrently in Ireland. However, since 2011 compulsory first registration has been extended to all counties, and thus it is envisaged that eventually the vast majority of property will be registered in the Land Registry.

Since 2006, both the Land Registry and the Registry of Deeds have formed part of the Property Registration Authority, or 'PRA'. The PRA is "the State organisation responsible for the registration of property transactions in Ireland"[12] and operates under the jurisdiction of the Department of Justice and Equality.

3.3. First Registration

If a plot of land is not registered with the Land Registry (and therefore classed as 'unregistered'), the owner or their representatives can lodge an application for first registration

[10] Property Registration Authority. (2015). Land Registry Services. Retrieved March 28, 2016, from Property Registration Authority: http://www.prai.ie/land-registry-services

[11] Property Registration Authority. (2015). Registry of Deeds Services. Retrieved March 28, 2016, from Property Registration Authority: http://www.prai.ie/registry-of-deeds-services

[12] Property Registration Authority. (2015d). The Property Registration Authority. Retrieved March 28, 2016, from Property Registration Authority: http://www.prai.ie/the-property-registration-authority

at any time. However, with the introduction of compulsory registration for purchased lands in 2011, most applications for first registration are now lodged as part of a change of ownership.

The advantages of registering a title with the Land Registry include:

- the protection of a state guarantee with regard to ownership
- straightforward transfers of property without repetitive investigations of title
- the benefit of clearly marked plots using the most up-to-date Ordnance Survey maps[13].

Within the PRA, there are four main application types, each of which serves a different purpose:

- A **Form 1** application is applicable where the property is a freehold title, which has never been registered in the Land Registry. This registration is based on documentary evidence – the applicant must supply all relevant deeds and documents for inspection by PRA officers.
- **Form 2** applications are similar to Form 1 applications, but concern leasehold titles rather than freehold. Also within the Form 2 category, a **Form 2 Examiners** designation is applied in certain specific situations – for example where the freehold is unregistered and the application is for absolute title of the leasehold title.
- A **Form 3** application can be a freehold or leasehold title, and is also based on documentary title, but the title is certified by the legal professional lodging the application, and is therefore not investigated to the same extent by the PRA. There are some additional constraints – for example the value of the property must be less than €1 million. In certain circumstances, a Form 3 application may be categorised as a **Form 3 Examiners** application – for example if the deed executed is five years old, and therefore requires special attention. A **Form 3 MSB** application concerns 'Multi-Storey Building' applications – for example the registration of a single apartment within a block. MSB applications have alternative mapping requirements compared to standard applications.
- A **Form 5** application is based on adverse possession – the applicant does not have clear documentary title and is therefore considered in 'adverse possession'. Ownership is therefore claimed based on long possession (a minimum of 12 years), and the applicant must show that they have not acknowledged anyone else's ownership or paid any rents in this time.

[13] Property Registration Authority. (2015e). Certification of Title – Benefits of Registration of Title. Retrieved March 28, 2016, from Property Registration Authority: http://www.prai.ie/certification-of-title-benefits-of-registration-of-title

3.4. *Focus of Change*

Since 1999 the PRA has been undertaking a series of measures to modernise, streamline and generally improve the services offered to the customer. These on-going projects fall into three main categories; Completion of the Register, Exploitation of Data, and Digitisation and Online Services.

3.4.1. Completion of the Register

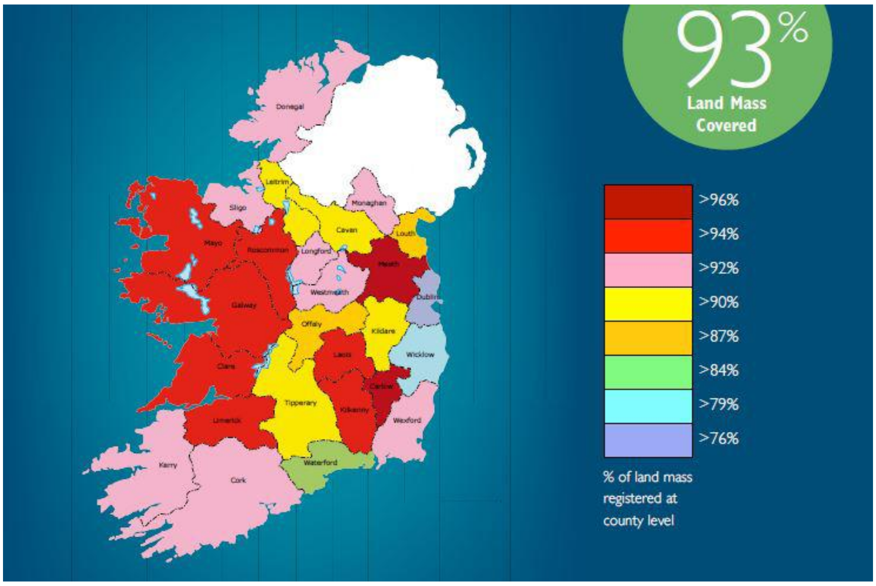


Figure 3.1 - Extent of land registration in Ireland (The Property Registration Authority of Ireland, 2014)

Currently, approximately 93% of the land in Ireland is registered. On a county level, this percentage is higher in rural areas than urban ones (for example over 96% of land in Meath or Carlow is registered, as compared to 76% in Dublin). This can be attributed in part to incentivised schemes such as the Congested District Board since the 1890s, which facilitated the work of the Land Commission in vesting land in tenant farmers. The gradual stages in which compulsory registration was gradually extended across the country are also significant – as can be seen from the table below, the urban centres of Galway, Limerick, Cork and Dublin were among the last areas to be subject to this legislation.

Regions	Dates
Carlow, Meath, Laois	1st January 1970
Longford, Westmeath, Roscommon	1st April 2006
Clare, Kilkenny, Louth, Sligo, Wexford, Wicklow	1st October 2008
Cavan, Donegal, Galway, Kerry, Kildare, Leitrim, Limerick, Mayo, Monaghan, Offaly, Tipperary, Waterford	1st January 2010
Cork, Dublin	1st June 2011

Table 3.1 - Dates for roll out of compulsory registration across Ireland

The fact that the Register is incomplete is a major obstacle in the conveyancing process. Legal professionals often find that the property in question is not registered, and must therefore refer to the Registry of Deeds system, conduct lengthy time-consuming searches and checks, and potentially deal with the problem of missing or damaged deeds. Therefore, completing the register is one of the key aims of the PRA for the near to mid future, as stated in the 2014 Annual Report:

> *"The Authority has a statutory remit to complete the Irish land register leading to a single system of title registration supportive of Government policies on land administration, including electronic registration, electronic conveyancing and the emerging Irish Spatial Data Infrastructure."* (The Property Registration Authority, 2015)

3.4.2. Exploitation of Data

The PRA processes and stores large amounts of data on a daily basis. This data includes:

- **Geographic data** regarding the land registered, displayed to the user in the form of a digital map. The data behind the map is stored in a geodatabase, which records the location and boundaries of the plot of land, as well as the location of the 'seedpoint' for the plot – this is a locator mark that links the land to the relevant digital folio.
- **Textual data** in the form of a 'folio' document that describes the property. An Oracle® database is employed to store this information. Each folio consists of a unique folio number, a description of the land (including plan numbers, townland names and the county) and any previous folio numbers. The ownership part of the folio records the current and previous owners' name and address and the date of registration. The folio may also include descriptions of any rights appurtenant or servient to the property – for example, rights of way, turbary or seaweed rights, and if there are charges pending on the lands in question.
- **Transactional data -** the Oracle® database also stores transactional data relating to every application lodged, using a unique application number as an identifier. For each application, data is stored regarding which folios are concerned, the type of application, the status of the application, if queries have been raised, the target data for completion and which members of staff have been involved in processing the application in any way.

There is massive potential to exploit these datasets for the potential benefit of the PRA. This opportunity has been recognised by the Senior Management Team, and is outlined in the Strategic Plan for 2016-2018:

> *"The Authority is aware of the benefits in facilitating even greater accessibility to and exploitation of the information held within the Register and continues to work towards positioning its property ownership data as a critical component in the move to improve decision making at all levels of Government and in the support of evidence-based policy making and is proactive in this regard."* (The Property Registration Authority of Ireland, 2016)

Some work has already taken place in this regard – for example, the Spatial Data Unit was set up recently to respond to spatial data queries from organisations like county councils and the National Roads Authority. These customers tend to have different requirements to legal professionals – for example, they may need to identify all of the registered owners along the proposed route of a new motorway. The Spatial Data Unit use GIS software to extract this information from the database and package it in a matter of minutes, thus replacing a procedure that historically would have taken several law searchers many months of laborious research.

The opportunities for building a valuable data source will be further enhanced with the upcoming merger of the PRA, Ordnance Survey Ireland and the Valuations Office into a single organisation entitled 'Tailte Ireland'. Although a final date has not yet been set for the merger to take place, work is on-going in determining how the new organisation will exist and function. The potential for consolidating and sharing data and resources between the three organisations is enormous. The Valuations Office is the state organisation with responsibility for valuing commercial property, while the OSi is Ireland's national mapping agency. Ultimately, it may be possible to combine the services provided by the Valuations Office and the OSi with the PRA's system of registration, for the benefit of the public. This could take the form of a combined digital map with different layers revealing historic, geographic, registration and valuation information.

However, the potential actualisation of such a service is still many years away. Instead, this experiment will use the historical data stored by the PRA in the course of day to day transactions to gain insight into the trends and features that influence intake.

3.4.3. Digitisation and Online Services

In the year 1999 the PRA was operating in a purely paper-based environment, with hard-copy folios and large paper maps on which boundaries were drafted by hand. Since then, the PRA has undertaken a series of projects to modernise and digitise essential processes. 1.75 million paper folios have been converted to an electronic format, a seamless digital map has replaced paper sheets, and various online services have been introduced in a series of phased projects under the umbrella of ‘eConveyancing’, or electronic conveyancing (The Property Registration Authority, 2005).

Today, all customers can conduct name and location searches online, order maps and copies of documents, track applications and create and lodge online forms via the website www.landdirect.ie. The eRegistration, eDischarges and eCharging Order services offer further enhancements, allowing registered professional users to draft deeds of transfer, circulate legal documents in a secure environment, pay fees by direct debit and lodge applications to create or cancel charges against property in certain circumstances.

Further additions are planned for the near future, and it is hoped that eventually the entire conveyancing process can be carried out electronically.

3.5. Challenges Faced

In terms of the near future, the Senior Management Team is pursuing the completion of the register as their principal, most urgent objective.

However, the organisation has encountered challenges in achieving this objective from various perspectives, including the volume of applications, staffing shortages and slow turnaround times.

3.5.1. Volume of Applications

As already stated, the introduction of compulsory first registration for sold lands for every county under Statutory Instrument 516 of 2010 has led to greatly increased volumes of applications for first registration since 2010. The following graph demonstrates the scale of the increase, over the course of only six years.

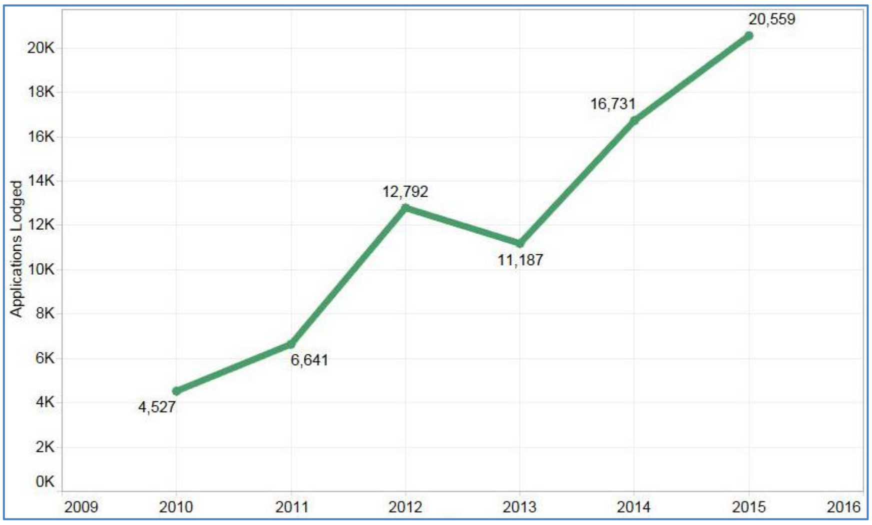


Figure 3.2 - Increase in intake of applications for first registration, 2010-2015

One consequence of this large intake has been the creation of a backlog in dealings. The next graph shows the increase in the arrear of first registration applications in the first four months of 2016. The total arrear has increased by 16% and the number of applications awaiting inspection has doubled.

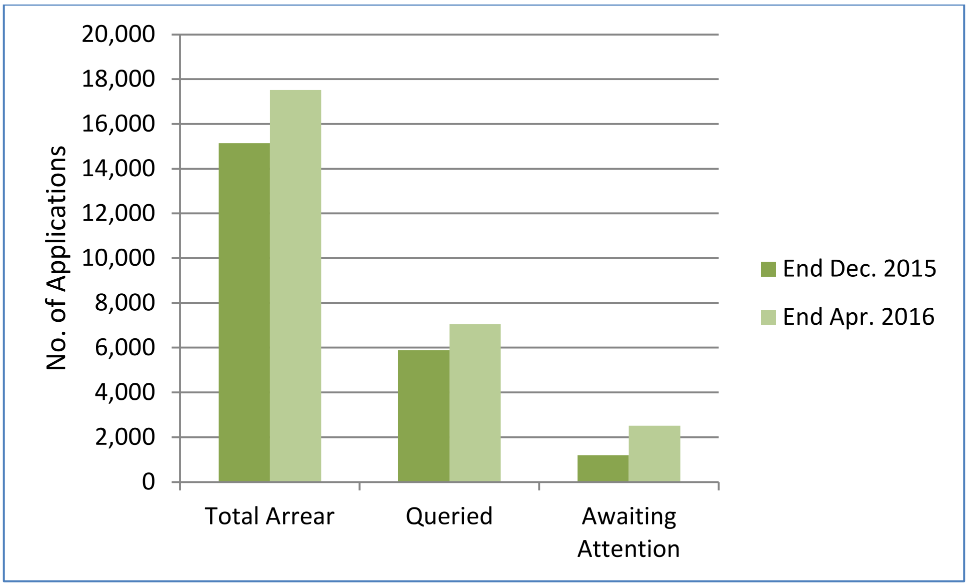


Figure 3.3 - Application arrear, 2016

If no new cases were lodged, it would still take many months or even years at current staffing levels to clear this backlog. Thus, if intake is to continue at current rates, some action will need to be taken to help staff to manage this workload pressure.

3.5.2. Staffing Shortages

The upwards trend in application volume comes at a time when budgets and staffing levels across the civil service have suffered significant cuts due to recessionary cuts and public sector reform initiatives. The PRA has not escaped these cuts, and also suffers from a relatively old staff demographic. The organisation carried out huge recruitment in the 1970s, and many of these staff are now reaching retirement age. All of these factors combined have resulted in a large staff exodus, and staffing levels have fallen from 705 employees at the end of 2008 to only 502 staff members in December 2014 (The Property Registration Authority, 2009, 2015).

The first registrations area has been particularly badly hit. Traditionally, all first registration applications have required extensive examination of documentary evidence. Therefore, these applications could only be processed by staff of a certain grade; ‘Examiner of Title’ or above. Examiner of Title is a professional grade that requires legal knowledge in order to make rulings and decisions on complex registrations. In order to qualify for this grade a staff member must hold a law degree and be qualified to practice as a solicitor or barrister.

As of May 2016, there are only nine officers at Examiner of Title or Chief Examiner of Title remaining in the PRA, with two scheduled to retire in coming months and another two at retirement age. The situation, therefore, is critical. In response, a three pronged-approach has been followed over the last five to six years.

The first tactic has been to invest heavily in training replacement staff. Unfortunately, this is not a fast process - it takes a minimum of three years for an officer to graduate from Kings Inn's (the sponsored programme of study is two years part time and one year full time). Even if a member of staff has qualified, any competition for a new appointment must be sanctioned by the Department of Justice and Equality and the Department of Public Expenditure and Reform, and the officer must also be successful in demonstrating that they have all of the relevant competencies and experience to fulfil the role. Therefore, simply training more staff is not a sufficient solution on its own.

Another approach has been the establishment of several Examiners Cross Functional (ECF) Teams in 2011. These teams include specially trained staff of both professional and generalist grades, who have permission to process certain first registration applications. Before the ECF Units, only an Examiner of Titles of above could 'settle' (complete) a first registration application. Now, staff of Assistant Principal, Higher Executive Officer and even Executive Officer can settle certain types of applications, under the supervision of an Examiner of Titles. This has eased the pressure on the Examiners of Title to a large extent, but the remit of ECF staff is still limited.

Finally, the third tactic has been the promotion of the Form 3 over the Form 1 or 2 application. Since Form 3 applications are certified by the solicitor, they generally do not require inspection by an Examiner of Title. This means that they have a much shorter turnaround time than Form 1 or 2 applications. In order to promote the use of the Form 3, the lodgement fee for a Form 3 application is €130, as compared to €500 for a Form 1 or Form 2. However, some solicitors are still reluctant to use the Form 3 since it places more responsibility on their shoulders. A dedicated study will shortly be carried out by another PRA staff member, investigating the potential advantages and possible repercussions of making self-certification in the form of Form 3 applications compulsory.

3.5.3. Slow Turnaround Times

As already discussed, the backlog in applications and shortages in qualified staff have contributed to slow turnaround times for some first registration applications. However, another significant reason for this is the number of queries that result from the lodgement of poorly structured applications – for example, applications lodged without essential documents such as affidavits or death certificates, documents not signed or witnessed correctly, and missing or incorrectly marked maps. If any of these flaws occur, the dealing must be queried with the lodging party.

It can take many months to receive replies to queries, and often follow-up letters are required. Furthermore, the Law Society insists that all correspondence from the PRA to solicitors be carried out by letter, rather than email, thus slowing the process down further.

An internal study was carried out in May 2016, in order to examine the most frequent reasons for issuing queries. All officers dealing with first registration applications were asked to record each query letter issued over the month, and the subject of the queries contained within. The results showed that the most common queries and reasons for rejection related to inadequately structured submissions for registration – essential documents not lodged, no maps or incorrectly marked maps provided, and errors in the forms submitted. In response to this, a programme of information sessions is being considered, to encourage legal professionals to improve the standard and comprehensiveness of the documentation being lodged for registration.

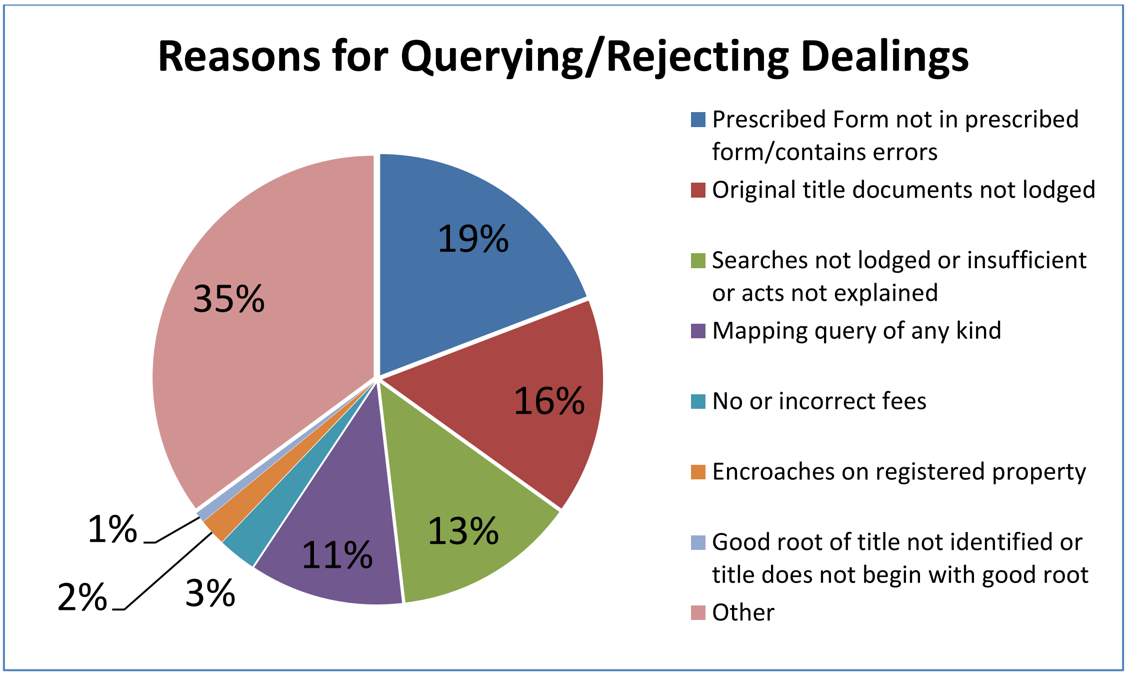


Figure 3.4 - Reasons for FR Applications to be Queried/Rejected

3.6. *Conclusions*

The PRA has faced significant challenges since the turn of the century in terms of modernising and embracing new technology, and has responded to these challenges in a positive, proactive manner. A series of new trials now face the organisation as many essential staff approach retirement age, intake increases and customers demand improved services and increased productivity. The PRA must continue to evolve and adapt in order to meet these new challenges, and data analysis can potentially assist in this process. Projects such as this one can facilitate management in planning for future intake, allowing the Authority to ensure that staffing levels are sufficient and that turnaround times are kept to a minimum. The next chapter will introduce the dataset assembled – the sources of data, the processing performed and the final variables selected for analysis.

4. THE DATASET

4.1. Introduction

The success of the proposed experiment hinges on the collation of relevant, useful, reliable data. In order to build a comprehensive dataset with as many initial features as possible, data was gathered from three sources; the PRA, the Property Services Regulatory Authority and the Central Statistics Office. The downloaded datasets were processed and collated to extract the relevant data in a series of data preparation exercises outlined below.

4.2. Preparing the Dataset

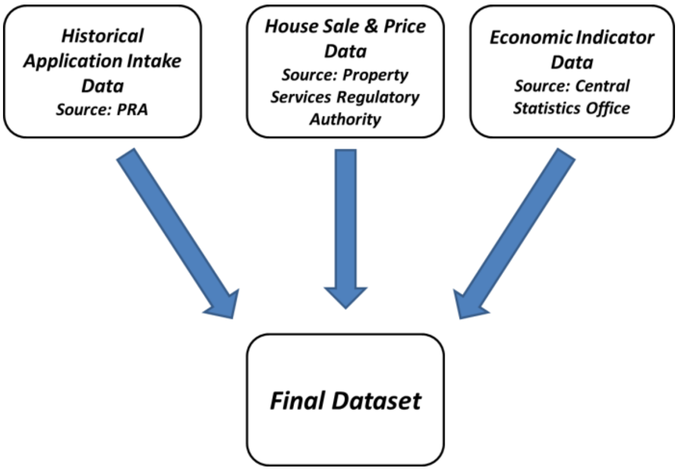

Historical Application Intake Data
Source: PRA
House Sale & Price Data
Source: Property Services Regulatory Authority
Economic Indicator Data
Source: Central Statistics Office
Final Dataset

Figure 4.1 - Sources of raw data

A range of datasets are available for certain aspects of this research – for example economic indicator data such as rates of GDP or unemployment can be accessed from various online sources. However, in the interests of building the most complete and reliable dataset possible, it was decided to only extract economic data from one official source – the Central Statistics Office. Other elements of the final dataset – the intake data and the house sale and price data – were only available directly from source; the PRA and PSRA respectively.

The process of extracting the datasets from the various sources is discussed in more detail in sections 4.3-4.5.

4.3. Historical application intake data (source: PRA)

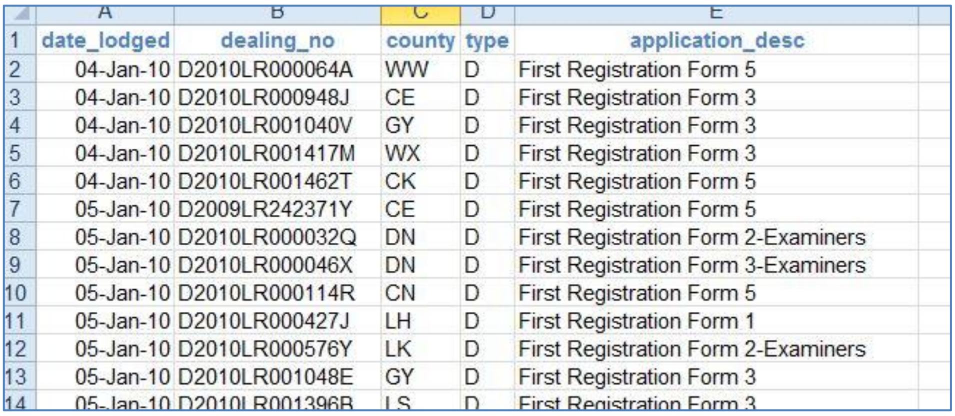
Applications for registration lodged with the PRA are called 'dealings'. An application for first registration takes the form of a bundle of hardcopy documents – an application form, deeds, affidavits and other documents. Upon lodgement, each dealing is set up on the internal system and assigned a unique ‘dealing number’. Thus, every application lodged is recorded in an Oracle® relational database.

In order to obtain data for this experiment, the following SQL query was executed to extract data directly from the database:

```
SELECT a.lodg_date, a.appl_num as "Application Number", a.county_id as "County", a.type as "Type",
c.application_type_desc as "Appl. Type"
FROM ITR_APPLICATIONS a
INNER JOIN ITR_APPLICATION_TYPE_LINES b
ON a.appl_num = b.appl_num
INNER JOIN REF_APPLICATION_TYPE_CODES c
ON b.application_type_code = c.application_type_code
WHERE lodg_date between '01-jan-10' and '31-dec-15'
and c.application_type_desc like 'First %'
ORDER BY a.lodg_date,c.application_type_desc,a.county_id;
```

This query returns the dealing number, county, application type and application description for every first registration application lodged between 2010 and 2015 inclusively, resulting in a single large excel sheet of 72,437 rows. Each row represents an instance of an application type. It must be noted that a single application can have more than one application type associated with in. For example, an application could be lodged by a customer who has recently purchased a property and has taken out a mortgage. In this instance an application would be set up as a single unique dealing number with two application types: ‘Transfer (Sale)’ and ‘Charge’. Thus, the 72,437 rows returned in relation to the first registration query represent 71,432 unique dealings.

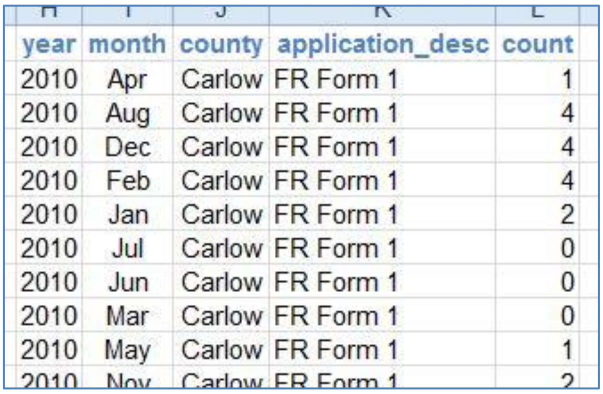
The structure of the initial dataset can be seen in the image below.



	A	B	C	D	E
1	date_lodged	dealing_no	county	type	application_desc
2	04-Jan-10	D2010LR000064A	WW	D	First Registration Form 5
3	04-Jan-10	D2010LR000948J	CE	D	First Registration Form 3
4	04-Jan-10	D2010LR001040V	GY	D	First Registration Form 3
5	04-Jan-10	D2010LR001417M	WX	D	First Registration Form 3
6	04-Jan-10	D2010LR001462T	CK	D	First Registration Form 5
7	05-Jan-10	D2009LR242371Y	CE	D	First Registration Form 5
8	05-Jan-10	D2010LR000032Q	DN	D	First Registration Form 2-Examiners
9	05-Jan-10	D2010LR000046X	DN	D	First Registration Form 3-Examiners
10	05-Jan-10	D2010LR000114R	CN	D	First Registration Form 5
11	05-Jan-10	D2010LR000427J	LH	D	First Registration Form 1
12	05-Jan-10	D2010LR000576Y	LK	D	First Registration Form 2-Examiners
13	05-Jan-10	D2010LR001048E	GY	D	First Registration Form 3
14	05-Jan-10	D2010LR001396B	LS	D	First Registration Form 3

Figure 4.2 - Format of raw data from SQL query

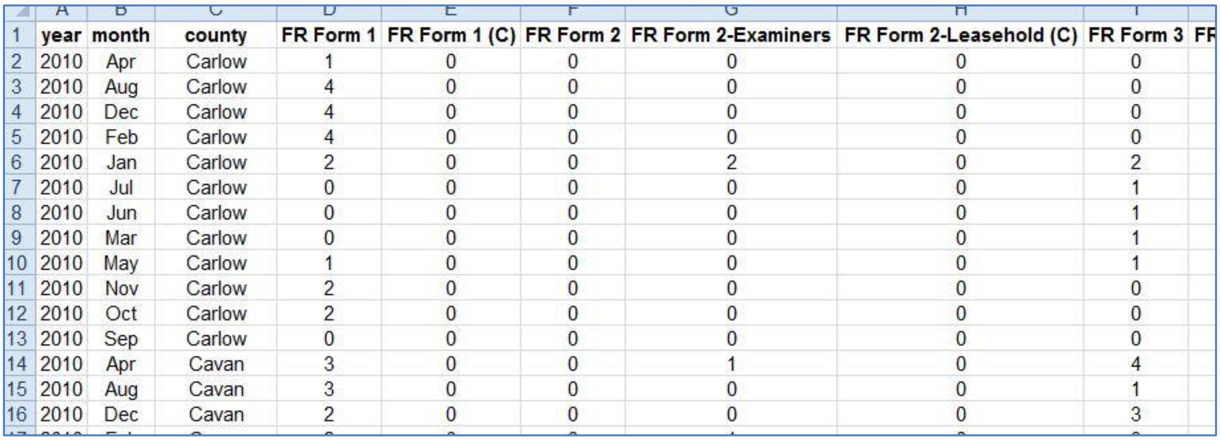
In order to organise the dataset into a usable format, the data was sorted by application type. The county codes were replaced with the textual county name, and the date was split out into year and month. A COUNTIFS formula was used to count the number of applications lodged per county per month.



year	month	county	application_desc	count
2010	Apr	Carlow	FR Form 1	1
2010	Aug	Carlow	FR Form 1	4
2010	Dec	Carlow	FR Form 1	4
2010	Feb	Carlow	FR Form 1	4
2010	Jan	Carlow	FR Form 1	2
2010	Jul	Carlow	FR Form 1	0
2010	Jun	Carlow	FR Form 1	0
2010	Mar	Carlow	FR Form 1	0
2010	May	Carlow	FR Form 1	1
2010	Nov	Carlow	FR Form 1	2

Figure 4.3 - Land Registry dataset, count of intake by type, per month, per county

The final Land Registry dataset appears as follows:



1	year	month	county	FR Form 1	FR Form 1 (C)	FR Form 2	FR Form 2-Examiners	FR Form 2-Leasehold (C)	FR Form 3	FR
2	2010	Apr	Carlow	1	0	0	0	0	0	
3	2010	Aug	Carlow	4	0	0	0	0	0	
4	2010	Dec	Carlow	4	0	0	0	0	0	
5	2010	Feb	Carlow	4	0	0	0	0	0	
6	2010	Jan	Carlow	2	0	0	2	0	2	
7	2010	Jul	Carlow	0	0	0	0	0	1	
8	2010	Jun	Carlow	0	0	0	0	0	1	
9	2010	Mar	Carlow	0	0	0	0	0	1	
10	2010	May	Carlow	1	0	0	0	0	1	
11	2010	Nov	Carlow	2	0	0	0	0	0	
12	2010	Oct	Carlow	2	0	0	0	0	0	
13	2010	Sep	Carlow	0	0	0	0	0	0	
14	2010	Apr	Cavan	3	0	0	1	0	4	
15	2010	Aug	Cavan	3	0	0	0	0	1	
16	2010	Dec	Cavan	2	0	0	0	0	3	

Figure 4.4 - Final Land Registry dataset

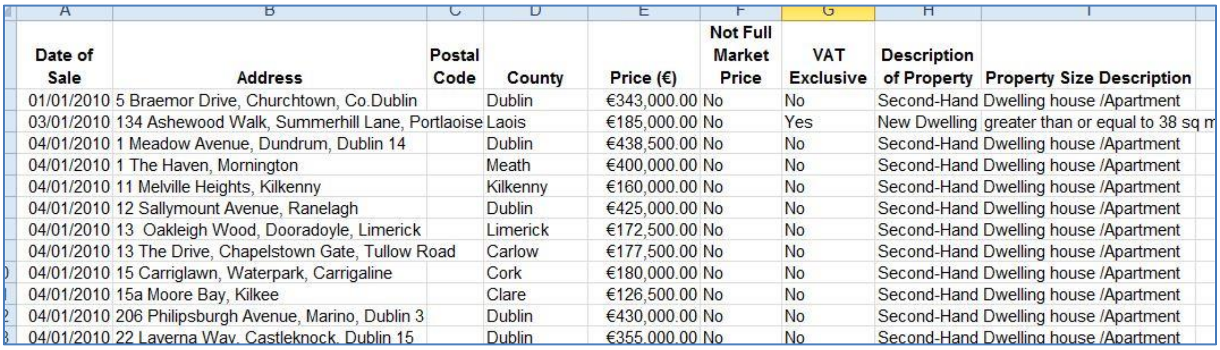
4.4. *House sale and price data (source: PSRA)*

The Property Services Regulatory Authority (PSRA) maintains a ‘Residential Property Price Register’, recording details of each residential property sold in the state since the beginning of 2010[14]. The data is gathered from stamp duty returns declared to the Revenue Commissioners, with stamp duty being payable on all residential property transactions where ownership is changed.

[14] Property Services Regulatory Authority. (2012). Registers. Retrieved April 13, 2016, from Property Services Regulatory Authority: http://www.psr.ie/website/npsra/npsraweb.nsf/page/publicregister-en

For each year of the chosen timeframe, a .csv file was extracted from the ‘Download’ section of the PSRA website. These files contain information on the date the property was sold, the address, county and postal code, the sale price and if this was full market price or not, a description of the property (new or second hand) and in some cases an indication as to size.

The six spread-sheets downloaded from the PSRA website were collated into a single sheet for ease of processing.



A	B	C	D	E	F	G	H	I
Date of Sale	Address	Postal Code	County	Price (€)	Not Full Market Price	VAT Exclusive	Description of Property	Property Size Description
01/01/2010	5 Braemor Drive, Churchtown, Co.Dublin		Dublin	€343,000.00	No	No	Second-Hand	Dwelling house /Apartment
03/01/2010	134 Ashewood Walk, Summerhill Lane, Portlaoise		Laois	€185,000.00	No	Yes	New Dwelling	greater than or equal to 38 sq m
04/01/2010	1 Meadow Avenue, Dundrum, Dublin 14		Dublin	€438,500.00	No	No	Second-Hand	Dwelling house /Apartment
04/01/2010	1 The Haven, Mornington		Meath	€400,000.00	No	No	Second-Hand	Dwelling house /Apartment
04/01/2010	11 Melville Heights, Kilkenny		Kilkenny	€160,000.00	No	No	Second-Hand	Dwelling house /Apartment
04/01/2010	12 Sallymount Avenue, Ranelagh		Dublin	€425,000.00	No	No	Second-Hand	Dwelling house /Apartment
04/01/2010	13 Oakleigh Wood, Dooradoyle, Limerick		Limerick	€172,500.00	No	No	Second-Hand	Dwelling house /Apartment
04/01/2010	13 The Drive, Chapelstown Gate, Tullow Road		Carlow	€177,500.00	No	No	Second-Hand	Dwelling house /Apartment
04/01/2010	15 Carriglawn, Waterpark, Carrigaline		Cork	€180,000.00	No	No	Second-Hand	Dwelling house /Apartment
04/01/2010	15a Moore Bay, Kilkee		Clare	€126,500.00	No	No	Second-Hand	Dwelling house /Apartment
04/01/2010	206 Philipsburgh Avenue, Marino, Dublin 3		Dublin	€430,000.00	No	No	Second-Hand	Dwelling house /Apartment
04/01/2010	22 Laverna Way, Castleknock, Dublin 15		Dublin	€355,000.00	No	No	Second-Hand	Dwelling house /Apartment

Figure 4.5 - Raw PSRA dataset

Certain columns were deleted as irrelevant to this query, or containing incomplete data: the ‘Address’, ‘Postal Code’, and ‘Property Size Description’ columns.

The “VAT Exclusive” label applies to all properties marked as “New Dwellings” in Ireland; VAT at 13.5% is applied to all new houses purchased. Given that this is a tax, and not part of the core value of the property, this column was excluded also.

The “Not Full Market Price” column affects certain properties. This designation can be due to various reasons, as explained by the Property Services Regulatory Authority:

> “*the price declared may reflect the retention of an interest in the property by the previous owner, or the fact that a part or fraction only of the property is being purchased; alternatively, the property may have been purchased at a reduced price under the Affordable Homes Scheme. In addition, in a very small number of cases, properties may be declared as purchased in exchange for other property, stocks and shares, etc.*”[15]

The fact that full and accurate price information is not available for these properties constitutes a problem. Given that there is no way of accurately calculating the full market price, and that

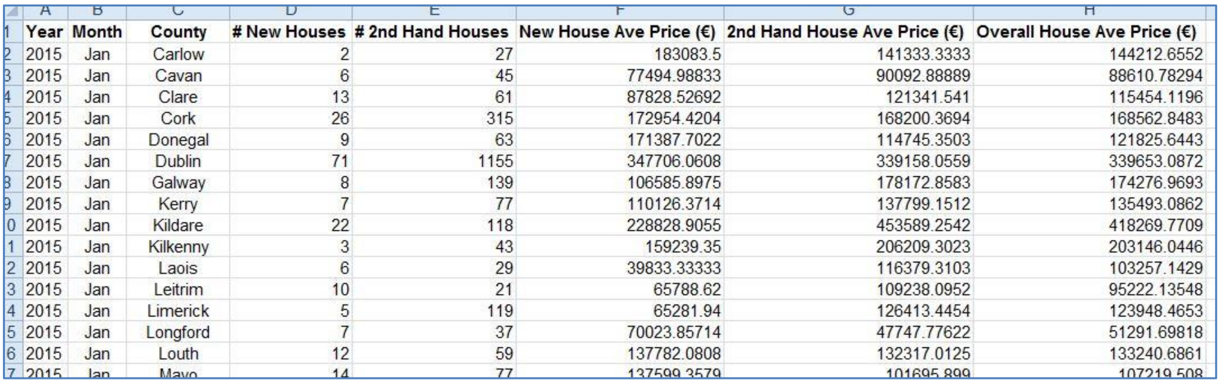
[15] Property Services Regulatory Authority. (2016). Prices Given Not Full Market Price. Retrieved May 4, 2016, from Property Price Register Ireland: https://www.propertypriceregister.ie/Website/npsra/pprweb.nsf/page/ppr-price-en

these lower values could introduce skew to the dataset, any row labelled 'Yes' for Not Full Market Price was removed from the dataset. This constitutes a total of 8,118 rows over the 6 years, or 4% of the original raw dataset.

Some of the 'Description of Property' fields were in Irish; to ensure consistency, "Teach/Árasán Cónaithe Atháimhe" was replaced with "Second-Hand Dwelling house /Apartment" and "Teach/Árasán Cónaithe Nua" was replaced with "New Dwelling house /Apartment".

The TEXT formula was used to split the date into Year and Month, and COUNTIFS and AVERAGEIFS to count the number of properties sold per month per county and to calculate the average sale prices.

The final Property Price Register dataset appears as follows:



Year	Month	County	# New Houses	# 2nd Hand Houses	New House Ave Price (€)	2nd Hand House Ave Price (€)	Overall House Ave Price (€)
2015	Jan	Carlow	2	27	183083.5	141333.3333	144212.6552
2015	Jan	Cavan	6	45	77494.98833	90092.88889	88610.78294
2015	Jan	Clare	13	61	87828.52692	121341.541	115454.1196
2015	Jan	Cork	26	315	172954.4204	168200.3694	168562.8483
2015	Jan	Donegal	9	63	171387.7022	114745.3503	121825.6443
2015	Jan	Dublin	71	1155	347706.0608	339158.0559	339653.0872
2015	Jan	Galway	8	139	106585.8975	178172.8583	174276.9693
2015	Jan	Kerry	7	77	110126.3714	137799.1512	135493.0862
2015	Jan	Kildare	22	118	228828.9055	453589.2542	418269.7709
2015	Jan	Kilkenny	3	43	159239.35	206209.3023	203146.0446
2015	Jan	Laois	6	29	39833.33333	116379.3103	103257.1429
2015	Jan	Leitrim	10	21	65788.62	109238.0952	95222.13548
2015	Jan	Limerick	5	119	65281.94	126413.4454	123948.4653
2015	Jan	Longford	7	37	70023.85714	47747.77622	51291.69818
2015	Jan	Louth	12	59	137782.0808	132317.0125	133240.6861
2015	Jan	Mayo	14	77	137599.3579	101695.899	107219.508

Figure 4.6 – Final PSRA dataset

4.5. *Economic indicator data (source: CSO)*

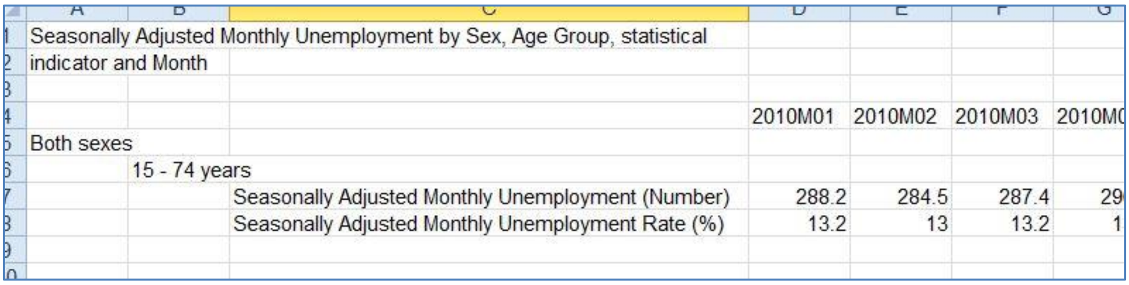
The Central Statistics Office (CSO) is the state organisation responsible for the "*collection, compilation, extraction and dissemination for statistical purposes of information relating to economic, social and general activities and conditions in the State*"[16].

The 'StatBank' section of the CSO website comprises a publicly available searchable database and includes current and historical data on a range of topics. In this instance, the sections concerning population, rates of unemployment, GDP and the consumer price index were accessed.

[16] Central Statistics Office. (2014). About Us. Retrieved April 13, 2016, from Central Statistics Office: http://www.cso.ie/en/aboutus/

Five datasets were downloaded from the CSO StatBank repository - Population per county, Unemployment returns, Gross Domestic Product, Consumer Price Index (All) data and Consumer Price Index (Housing) data.

The population dataset is based on the 2011 Census. For each county, the dataset outlines the number of persons, of all ages, resident in the area. The figures have a wide range from only 31,798 people in Leitrim up to 1,273,069 in Dublin alone.



	A	B	C	D	E	F	G
1	Seasonally Adjusted Monthly Unemployment by Sex, Age Group, statistical						
2	indicator and Month						
3							
4				2010M01	2010M02	2010M03	2010M0
5	Both sexes						
6		15 - 74 years					
7			Seasonally Adjusted Monthly Unemployment (Number)	288.2	284.5	287.4	29
8			Seasonally Adjusted Monthly Unemployment Rate (%)	13.2	13	13.2	1
9							
0							

Figure 4.7 - Unemployment dataset from StatBank

The Unemployment and GDP datasets downloaded cover the period 2010-2015.

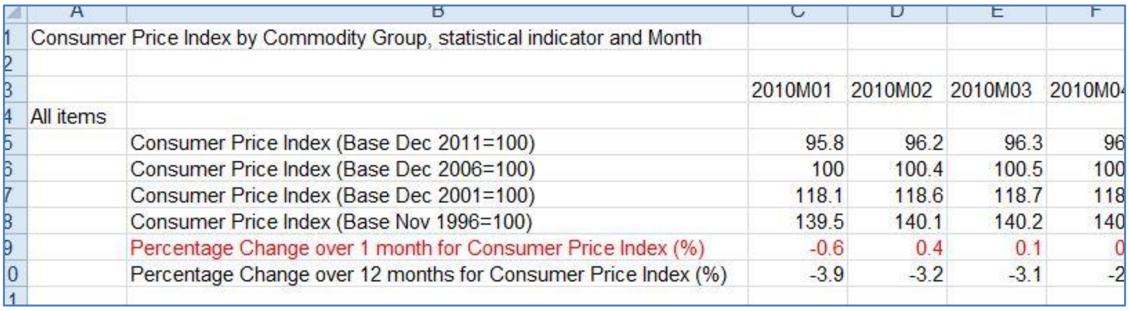
- The Unemployment dataset shows seasonally adjusted rates of unemployment for both sexes in the 15-74 age-group – this demographic can be viewed as the potential workforce.
- The GDP dataset is measured in millions of euro, and is based on ‘Constant Factor Cost’; in other words, it has been adjusted to allow for inflation. Since inflation affects the value of currency over time, constant price GDP calculates the value of goods and services in relation to a base year – in this case 2013. This allows changes in GDP to be accurately analysed over a time period[17].

Neither dataset required extensive processing; the Transpose formula was used to flip the columns and rows, and the relevant data was then copied into the final dataset.

The Consumer Price Index (CPI) datasets were in a similar format, and were processed in the same manner. Two separate indices were chosen, as both may potentially be relevant in modelling the factors that influence house sales. The CPI monitors changes in average prices across a range of goods and services in Ireland, and thus constitutes a measure of economic inflation. The CPI (All) dataset covers 12 different categories of products and services, including health, transport, food, clothing and communications. The CPI (Housing) dataset restricts calculations to housing, water, electricity and fuel. The CSO sheets include various

[17] Constant-Price GDP. (n.d.). Retrieved from Investing Answers: http://www.investinganswers.com/financial-dictionary/economics/constant-price-gdp-5673

categories of indices, but in this instance the "percentage change over 1 month" is of most relevance.



	A	B	C	D	E	F
1	Consumer Price Index by Commodity Group, statistical indicator and Month					
2						
3			2010M01	2010M02	2010M03	2010M0
4	All items					
5		Consumer Price Index (Base Dec 2011=100)	95.8	96.2	96.3	96
6		Consumer Price Index (Base Dec 2006=100)	100	100.4	100.5	100
7		Consumer Price Index (Base Dec 2001=100)	118.1	118.6	118.7	118
8		Consumer Price Index (Base Nov 1996=100)	139.5	140.1	140.2	140
9		Percentage Change over 1 month for Consumer Price Index (%)	-0.6	0.4	0.1	0
0		Percentage Change over 12 months for Consumer Price Index (%)	-3.9	-3.2	-3.1	-2
1						

Figure 4.8 - CPI (All) dataset

It should be noted at this stage that based on the literature explored in Chapter 2, interest rates would appear to be another variable with potential to impact on property prices and sales. Unfortunately, it has not been possible to identify a consistent representation of interest rate change over the course of the chosen timeframe from 2010-2015.

The CSO includes datasets in relation to interest rates in the StatBank section of their website; this data is sourced from the Central Bank of Ireland. The interest rate of relevance to this study is the standard variable rate (SVR) available for home purchases. However, the SVR data available from the CSO site only runs up to the end of 2013.

Contacting the Central Bank directly confirmed that the organisation changed their method of recording the SVR for 2014 onwards – whereas the older rates were an average of advertised rates, the new data represents a weighted average of drawn down mortgage loans. Therefore, the two rates are not comparable, and cannot be reliably included in an analysis such as this.

4.6. The Final Dataset

The final task was to collate the datasets from the three original sources into a single final dataset. Some additional processing was also carried out:

- To allow the list to sort correctly, the month column was changed from a textual description to a numeric format (Jan became 1, etc.).
- In the 'Ave New House Price' column, some cells were labelled as an Excel error. In these instances, since no new houses were sold in the county in a particular month, it is not possible to calculate the average price. These 30 cells were replaced with 'N/A'.

At the beginning of the project, it was envisaged that specific software such as Pentaho Kettle might need to be used to integrate and transform the datasets. However, in practice Excel was found to be more than sufficient for the purpose.

The final variables are as follows:

	Variable Name	Format	Notes
	Year	Numeric	2010-2015
	Month	Numeric	1-12
	County	Text	26 counties of the Republic of Ireland
	Population	Numeric	Leitrim is lowest, Dublin the highest
	New_Houses (no. of new houses sold/month)	Numeric	Ranges from 0 to 414
	Sec_Hand_Houses (no. of second hand houses sold/month)	Numeric	Ranges from 2 to 1,908
	New_Price (avg. price of new house)	Currency	Ranges from €20,000 to € 2,069,016
	Sec_Hand_Price (avg. price of 2nd hand house)	Currency	Ranges from €35,395 to €455,293
	All_Price (avg. price of either type of house)	Currency	Ranges from €38,053 to €584,049
	CPI_All	%	The % monthly change
	CPI_Housing	%	The % monthly change
	GDP	Numeric	The GDP per capita (€ million)
	Unemployment	%	The % monthly change
Application Type	Form 1	Numeric	The number of applications lodged per county per month
	Form 2		
	Form 2 – Examiners		
	Form 3		
	Form 3 – MSB		
	Form 3 – Examiners		
	Form 5		
	Count	Numeric	(Target Feature) This is the total sum of applications lodged for any month, by county. Ranges from 0 – 1,714.

Table 4.1 - Variables in final prepared dataset

4.7. Conclusions

The final dataset represents a summary of the various factors that could potentially influence house sales, and therefore indirectly also influence intake of applications. The dataset is complete, with few missing values, and should provide a solid basis with which to build models and gain insight. The following chapter will perform some exploratory statistics and data profiling to gain a deeper understanding of the data and to identify the biggest trends in the Irish property market.

5. EXPLORATORY STATISTICS & PROFILE OF DATA

5.1. Introduction

Exploratory statistics are a useful tool in gaining an initial understanding of the profile of a dataset; its principal characteristics and scope. Visualisations can also assist the analyst by revealing trends and by highlighting outliers and other inconsistencies as areas for further investigation.

The final combined dataset in this case covers an observation period of six years from 01/01/2010 – 31/12/2015 and comprises 1,872 rows. The following sections will explore the principal attributes of this dataset in detail. The analysis will begin with population, and then examine house sales and price, rates of unemployment, the consumer price index and finally application intake data.

5.2. Software Tools

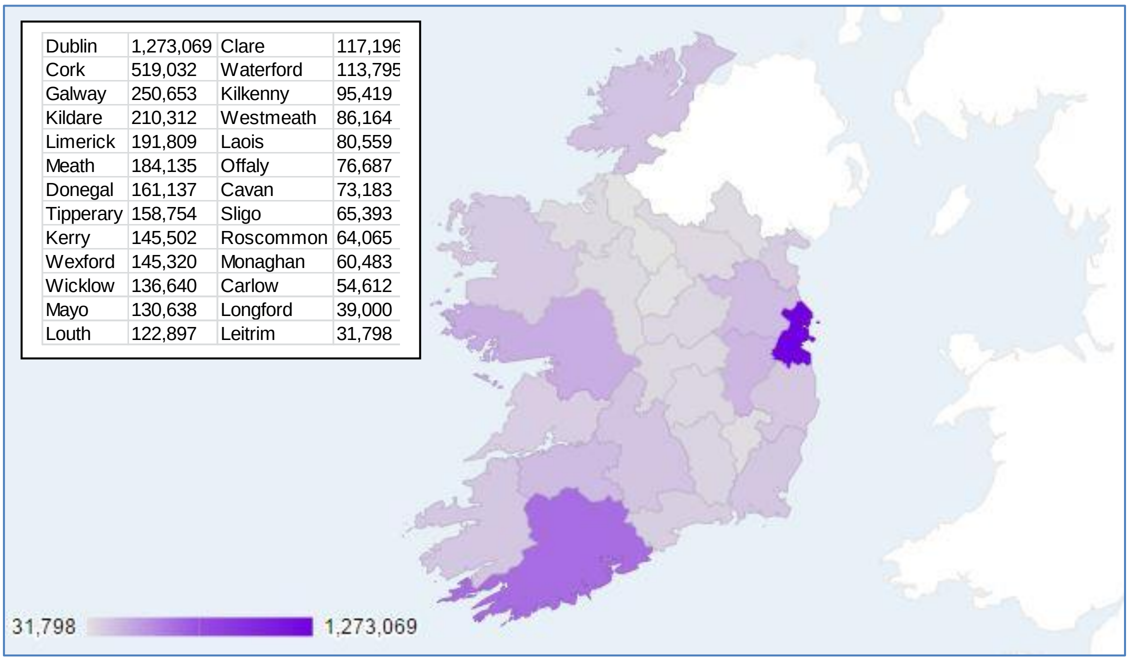
Two software tools were used to create these visualisations. The majority of graphical output was created using Tableau, proprietary software specifically designed to create exploratory and explanatory visualisations. The advantages of Tableau are in the speed of generation and in the quality and appearance of the output. Data can be uploaded in a variety of formats and is automatically classified into one of two roles: ‘dimensions’ (discrete variables) or ‘measures’ (continuous variables); multiple sources can also be accessed concurrently. For building visualisations, Tableau uses a drag and drop methodology, so that each feature can be added or removed to the graph with ease. Depending on the dimensions and measures selected, the programme then displays the possible visualisation types available. Thus, Tableau has many advantages, not least that the graphics created are aesthetically pleasing and communicate trends and patterns effectively.

However, one problem was encountered while creating choropleths to show the distribution of various features by county across Ireland – for example population, unemployment etc.. The data extracted from the PRA is categorised by county, of which there are 26 in the Republic of Ireland. However, the shapefile employed by Tableau specifies the administrative regions of Ireland instead, and therefore Dublin is divided into four areas (Fingal, Dublin City, South Dublin and Dún Laoghaire-Rathdown) and Tipperary is shown as Tipperary North and Tipperary South. Upon contacting Tableau, it appears this error is to be corrected with the launch of Version 10, but this will only be introduced after this project is completed.

Therefore, it was not possible to use Tableau for the creation of choropleths. As an alternative, a WordPress plugin called Visualizer was employed. This free software is currently used within the PRA to create statistical maps for reports, and is straightforward to implement. It uses the Google 'GeoChart' API to generate the choropleth and other graphs for publishing online. To create a choropleth, the dataset is loaded as a .csv file and the map is automatically shaded according to the values. The colours can be adjusted to create different colour ramps and labels applied as required. In the usual scheme of things the finished map is then published to the web – in this instance a screenshot was taken for inclusion in this document.

5.3. Population

The population of Ireland is unequally distributed by county, as the choropleth below shows. Cities like Dublin, Galway and Cork and their neighbouring counties have the highest populations, while midland regions and the North-West have become scarcely populated.



Dublin	1,273,069	Clare	117,196
Cork	519,032	Waterford	113,795
Galway	250,653	Kilkenny	95,419
Kildare	210,312	Westmeath	86,164
Limerick	191,809	Laois	80,559
Meath	184,135	Offaly	76,687
Donegal	161,137	Cavan	73,183
Tipperary	158,754	Sligo	65,393
Kerry	145,502	Roscommon	64,065
Wexford	145,320	Monaghan	60,483
Wicklow	136,640	Carlow	54,612
Mayo	130,638	Longford	39,000
Louth	122,897	Leitrim	31,798

Figure 5.1 - Population by County, 2011 Census

The 2011 Census returned Ireland's population as just over 4.5 million people. It also showed that despite being Ireland's third smallest county in area, Dublin has 28% of the overall population. The three most densely populated urban areas were all suburbs in Fingal; Kinsealy-Drinan, Lusk and Swords respectively (Central Statistics Office, 2012). The Kinsealy-Drinan area was found to contain 5,101 persons per km^2 as compared to the least densely populated area of Bunbeg-Derrybeg in Co. Donegal with only 196 persons per kilometre squared[18].

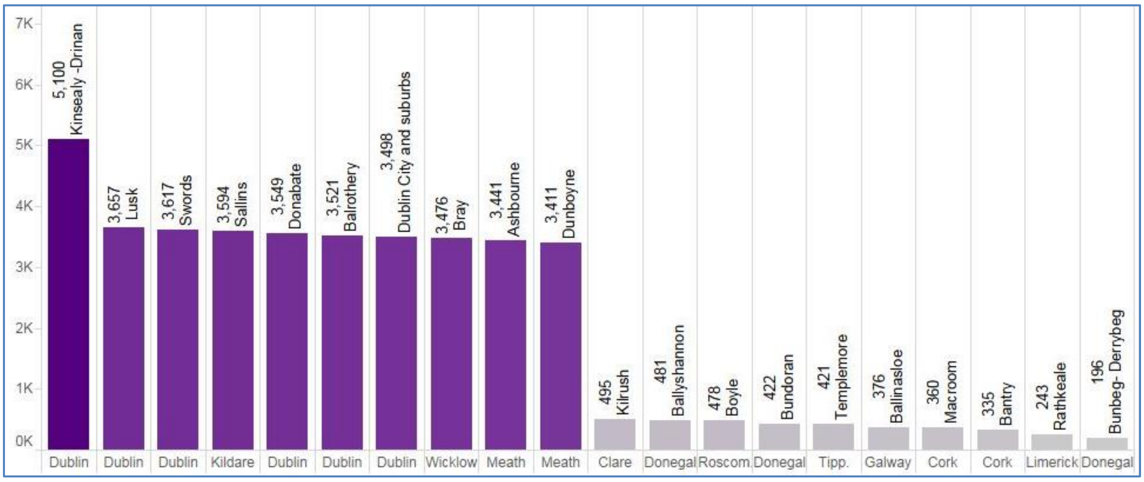


Figure 5.2 – Areas of Highest and Lowest Population Density; persons per sq. km

These figures also demonstrate the overall trend towards urbanisation that has been taking place in Ireland over the last 50 years. In 1961, only 46% of Irish people lived in urban areas. This has increased to 62% in 2011, with the largest swell towards urban living occurring in the 15 years between 1996 and 2011.

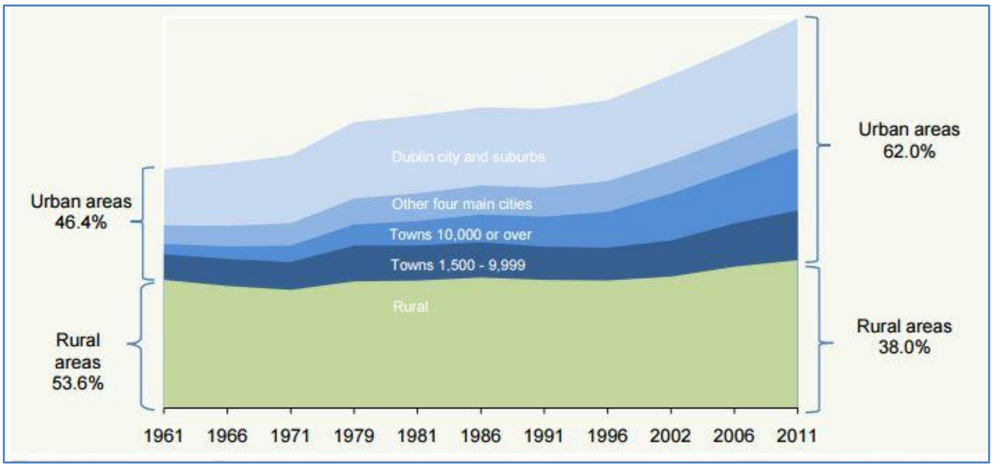


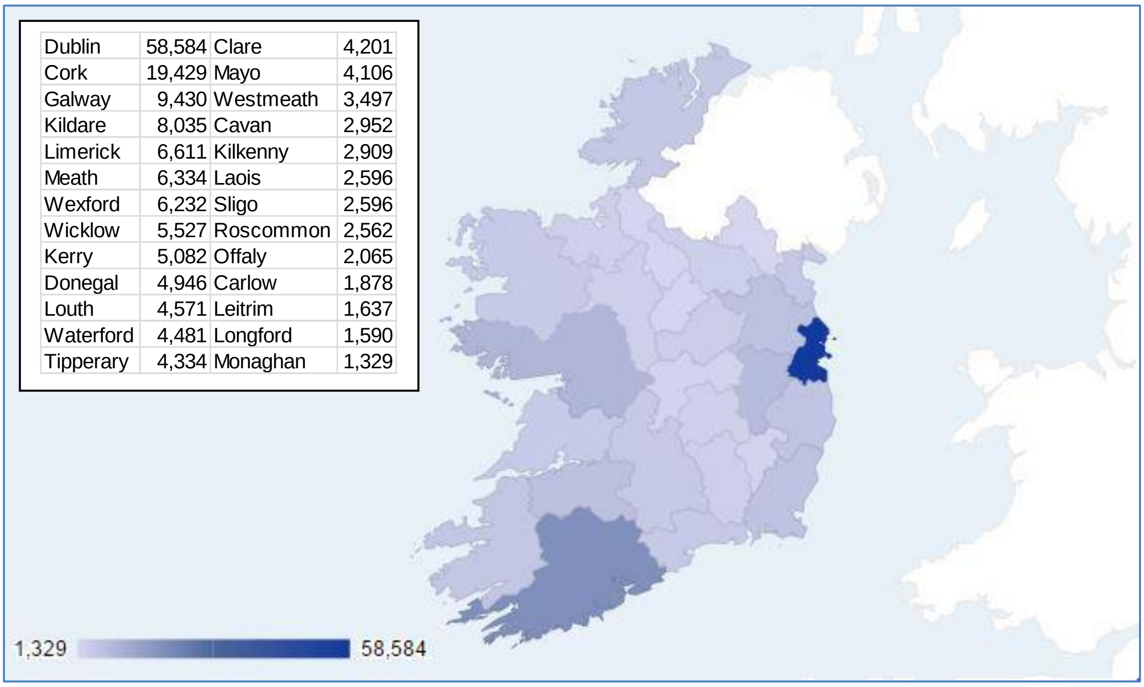
Figure 5.3 - 50 years of Urbanisation in Ireland (Central Statistics Office, 2012)

5.4. *House Sales*

As people increasingly move towards urban centres, it therefore stands to reason that the largest number of residential property transactions will also take place in these areas. At first glance, this appears to be the case. The choropleth below shows the aggregated number of

[18] Central Statistics Office. (2015, Oct 16). 2011 - Population density (persons per sq. km) (Number) by Towns by Size. Retrieved from Central Statistics Office StatBank: http://www.cso.ie/px/pxeirestat/Statire/SelectVarVal/Define.asp?Maintable=CD116&Planguage=0

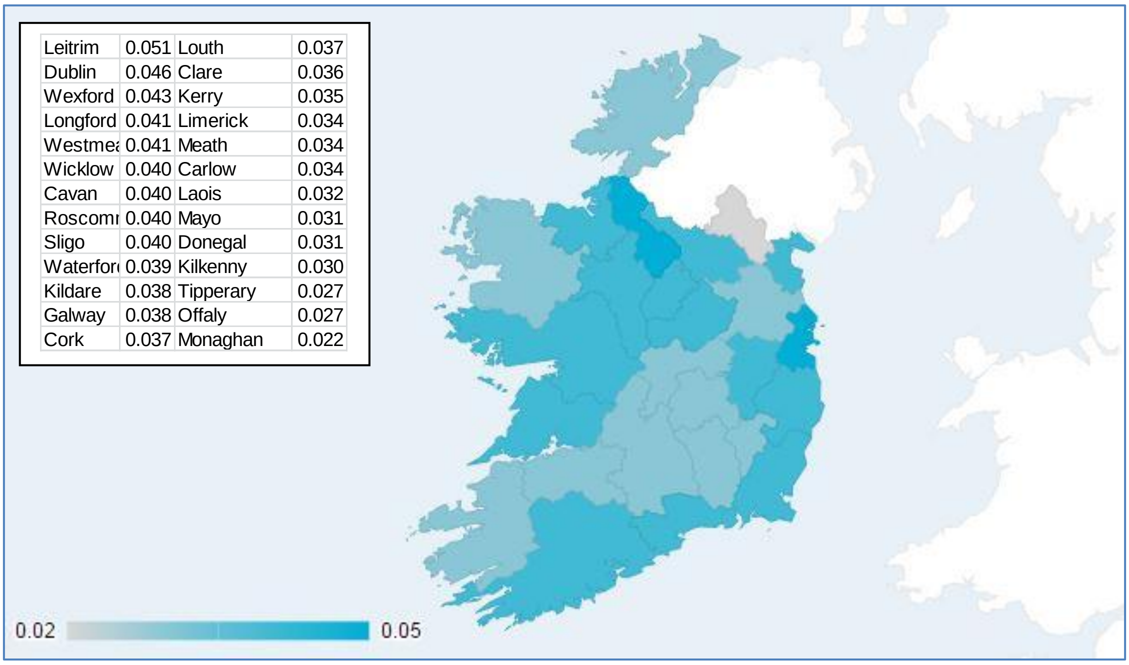
houses sold per county between 2010 and 2015, as deduced from the Property Price Register dataset. The highest numbers of property sales are in the largest cities of Dublin, Cork and Galway, and in commuter belt counties like Kildare, Limerick and Meath.



Dublin	58,584	Clare	4,201
Cork	19,429	Mayo	4,106
Galway	9,430	Westmeath	3,497
Kildare	8,035	Cavan	2,952
Limerick	6,611	Kilkenny	2,909
Meath	6,334	Laois	2,596
Wexford	6,232	Sligo	2,596
Wicklow	5,527	Roscommon	2,562
Kerry	5,082	Offaly	2,065
Donegal	4,946	Carlow	1,878
Louth	4,571	Leitrim	1,637
Waterford	4,481	Longford	1,590
Tipperary	4,334	Monaghan	1,329

Figure 5.4 - Number of houses sold per county, 2010-15

However, considering the two factors of population and house sales concurrently, a different picture emerges. While the major urban centres still score highly, the distribution of sales per capita of population shows a much more even distribution pattern around the country.



Leitrim	0.051	Louth	0.037
Dublin	0.046	Clare	0.036
Wexford	0.043	Kerry	0.035
Longford	0.041	Limerick	0.034
Westmea	0.041	Meath	0.034
Wicklow	0.040	Carlow	0.034
Cavan	0.040	Laois	0.032
Roscomn	0.040	Mayo	0.031
Sligo	0.040	Donegal	0.031
Waterford	0.039	Kilkenny	0.030
Kildare	0.038	Tipperary	0.027
Galway	0.038	Offaly	0.027
Cork	0.037	Monaghan	0.022

Figure 5.5 - Houses sold per capita of population, 2010-15

The county with the highest number of residential property transactions per head of population is in fact Leitrim, the least densely populated county of all. There were 1,637 houses sold over the 6 year period for a population of 31,798, which equates approximately to 1 house for every 20 Leitrim residents. In nearby Monaghan, only 1 house was sold for every 45 inhabitants. Of course, not all property in a given county is necessarily purchased by residents of that county – nevertheless, this anomaly will require some further attention.

The partial recovery of the property market after the recent economic crash can clearly be seen in an analysis of the number of houses sold per year of the dataset:

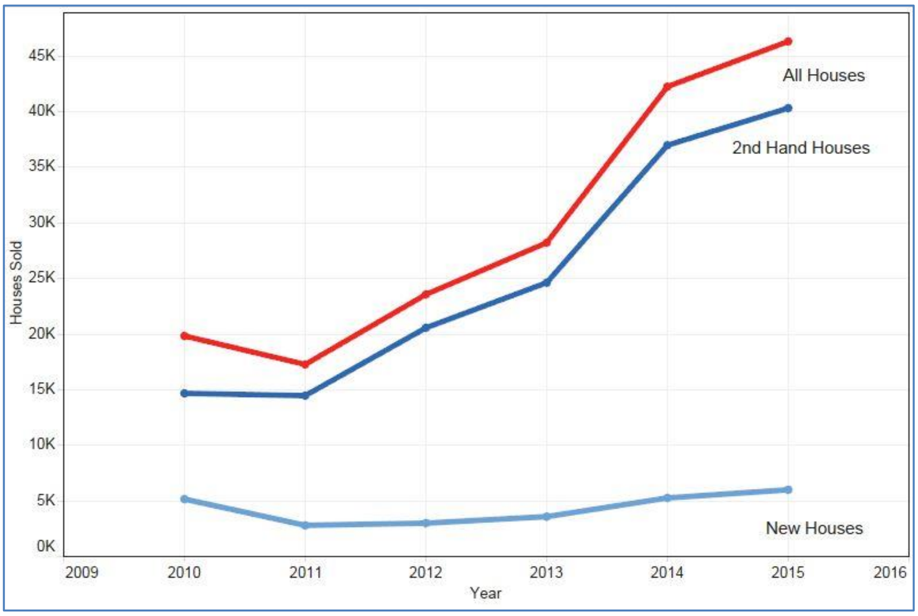


Figure 5.6 - House Sales 2010-15

The number of new houses sold has remained consistently static across the dataset – the only recovery has been in the second hand housing market, although this has been impressive. Less than 15,000 second hand homes were sold in 2010, compared to over 40,000 in 2015.

In terms of seasonality, the highest average property sales take place in December, followed by July. The months with the least number of sales recorded are January and February. In general, sales appear to increase as the year progresses, with the largest number of sales in the last six months of each year.

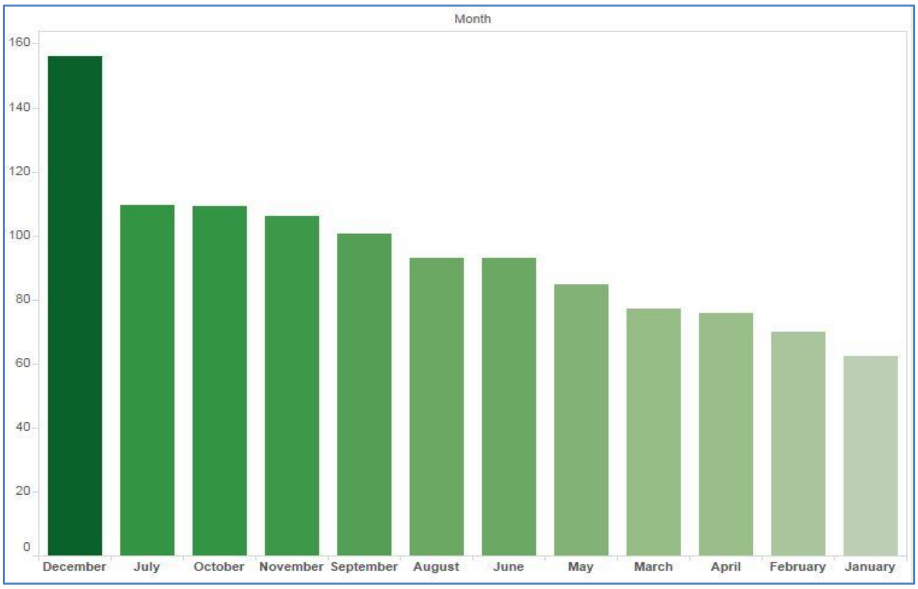


Figure 5.7 - Average number of property sales by month 2010-15

5.5. *Average Sale Price*

In terms of house sales, the average sale price must also be considered. If house prices are low, does this stimulate the market and therefore increase sales? Or are house prices low because properties are difficult to sell? Scatter plots reveal that there does appear to be a relationship between house price and quantity of sales:

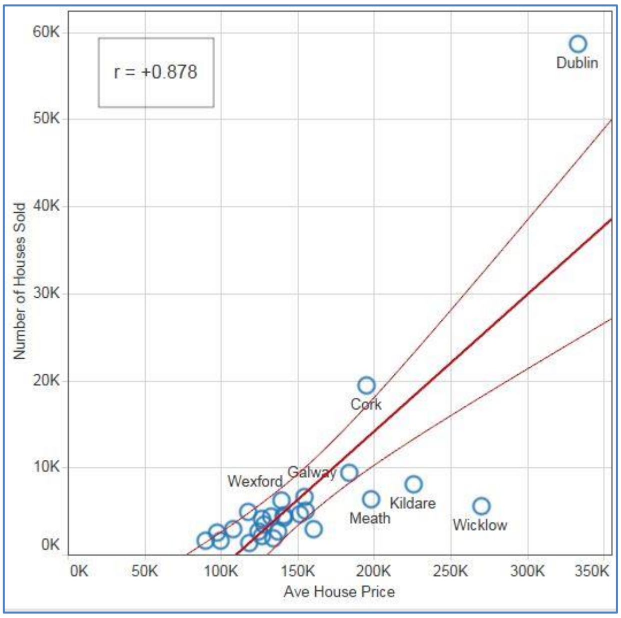


Figure 5.8 - Correlation between House Prices & Quantity of Sales 2010-15

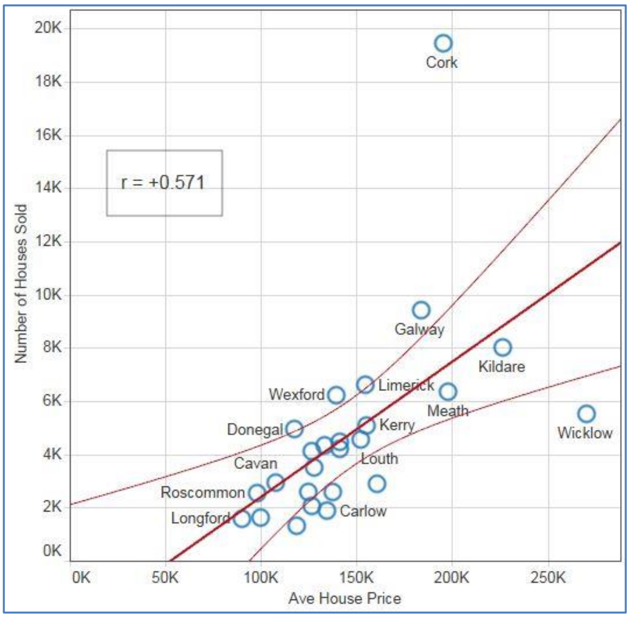


Figure 5.9 - Correlation between House Prices & Quantity of Sales 2010-15, excl. Dublin

The first scatter plot shows a very strong positive correlation of 0.8 between house price and number of sales. Dublin is the county with the highest average house price and the largest number of sold properties. However, Dublin represents a clear outlier and is undoubtedly skewing the result for the other counties.

The second plot is the same correlation but excluding Dublin – this still results in a positive correlation, but it is weaker than before. On this second plot, this time Cork is seen to be an outlier.   This result indicates a relationship between house price and number of sales, but that this may be more relevant in more urbanised counties than in predominately rural ones.

As another hypothesis, it might be reasonably assumed that the higher the population density in an area, the higher the demand for housing and therefore the higher that houses prices will be. However, this does not appear to be the case. On the first scatter plot below, the positive value of R indicates that there is some correlation between population density and house prices, but it is relatively weak at 0.685. However, once again Dublin represents a clear outlier. Removing Dublin from the equation still results in a positive correlation, but it is weaker. The conclusion must be that other factors have a greater effect on house prices than population density.

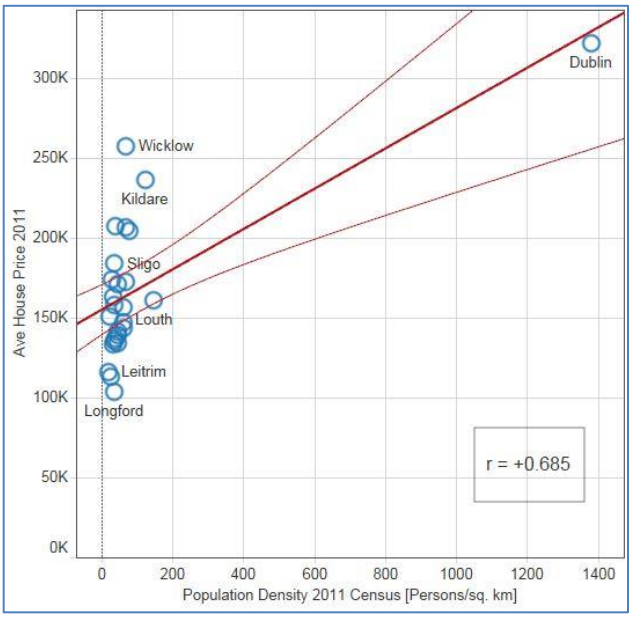


Figure 5.10 - Correlation between Pop. Density and House Prices 2011

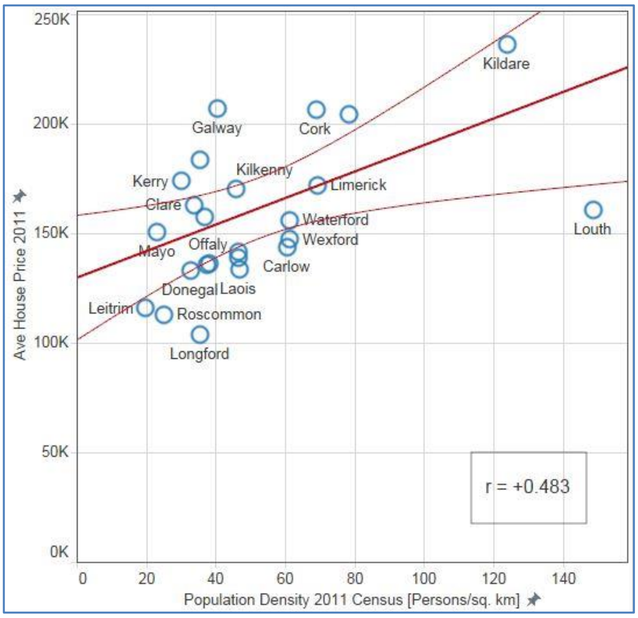


Figure 5.11 - Correlation between Pop. Density and House Prices 2011, excluding Dublin

The effect of the economic crash and subsequent recovery can also be seen in the dataset. Nationally, house prices fell from an average price of over €190,000 in 2010 to a low of €127,000 in 2013, before recovering slightly by the end of 2015. However, house prices in 2015 were still only 75% of the 2010 average. The graph also shows that while the average cost of a new or 2nd hand property was similar in 2010, the average prices of new houses at first held up more strongly during the recession, before crashing to a lower rate in 2013, and are now staging a recovery.

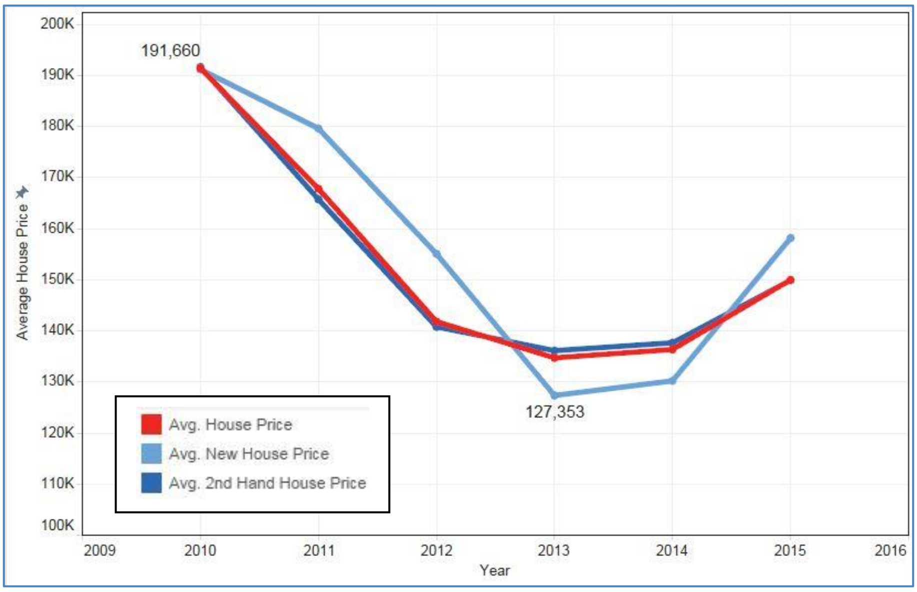


Figure 5.12 - Collapse of House Prices 2010-15

By county, the rates of collapse and recovery vary considerably. The counties with the biggest yearly percentage drops in prices were all in the North West, while the biggest increases were in midland counties like Laois, Cavan and Westmeath – and all in 2015. This may suggest that property prices in counties closer to urban centres were less affected by the economic turmoil.

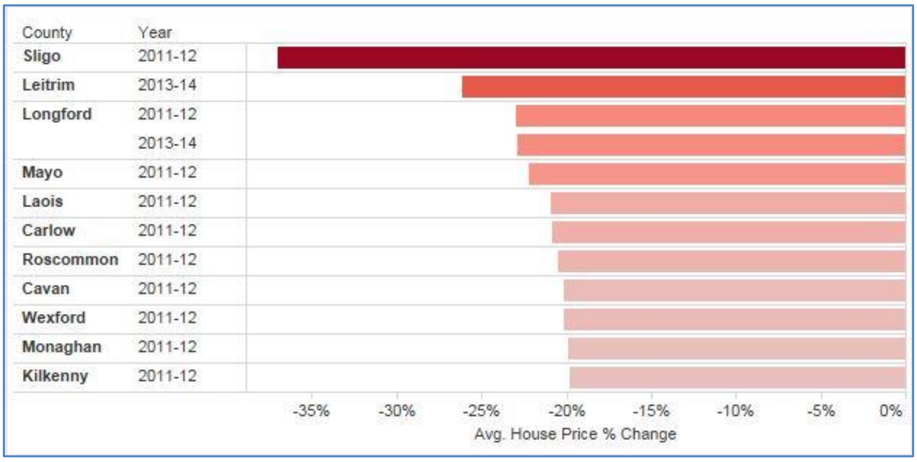


Figure 5.13 - Largest Price Decreases 2010-15

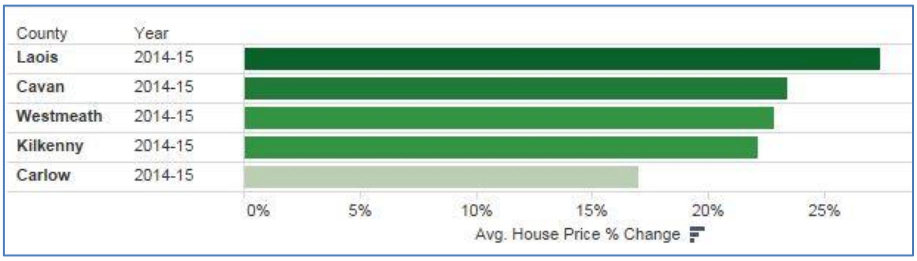


Figure 5.14 - Largest Price Increases 2010-15

5.6. Unemployment Rates

The rate of unemployment in Ireland rose gradually between 2010 and 2012, but has fallen off consistently since then to a low of 9.4% in 2015.

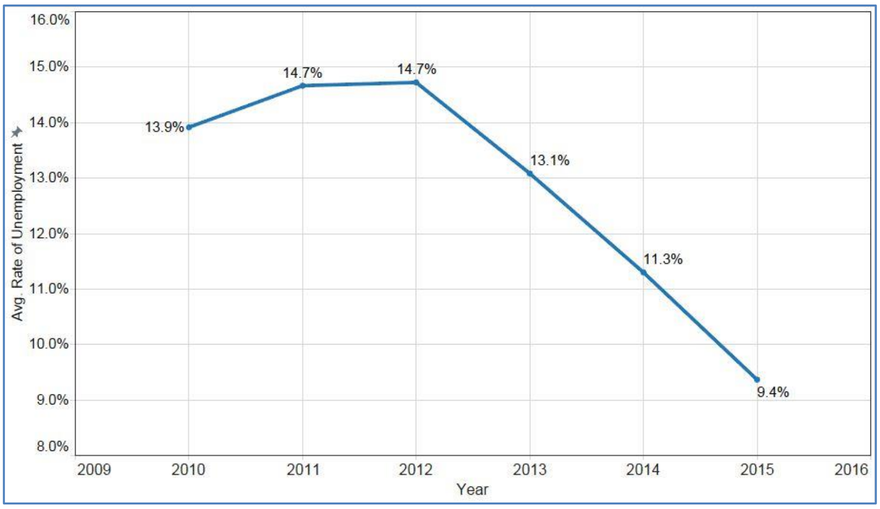


Figure 5.15 - Unemployment Rate 2010-15

5.7. Consumer Price Index

The CPI (All) dataset covers 12 different categories of products and services, including health, transport, food, clothing and communications, while the CPI (Housing) dataset restricts calculations to housing, water, electricity and fuel. Plotting the fluctuations in both indices over the six years, it is clear that the CPI (Housing) data slows much greater variation than the general CPI rate. While the CPI (All) percentage varies between 0.21% and -0.3%, the CPI (Housing) rate ranges from 0.79% to -0.16%. This variance demonstrates the devastating crash experienced in the property industry, and how the effects in this sector were far more severe than in others.

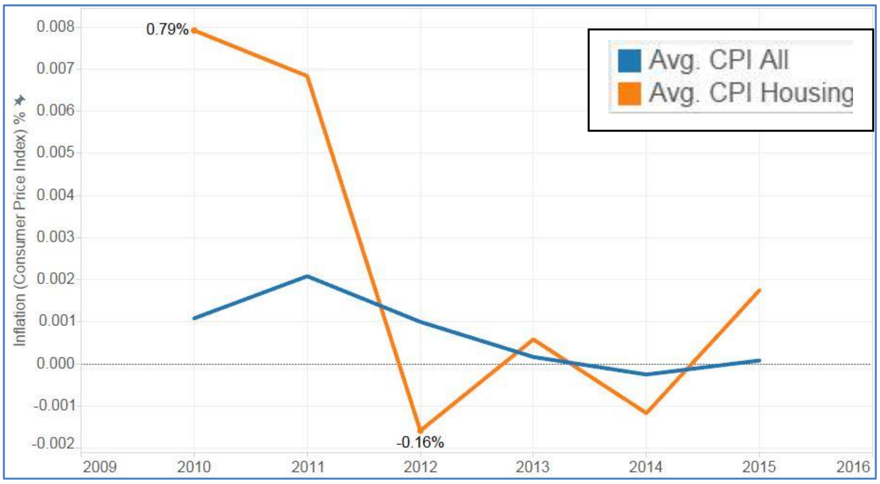


Figure 5.16 - Fluctuation in CPI, 2010-15

5.8. Applications for First Registration

The target feature in this dataset is Count; the total sum of applications lodged for any month per county. The bar plot below shows the greater than 400% increase in applications for first registration over the course of the dataset, despite a minor fall in intake from 2012-2013.

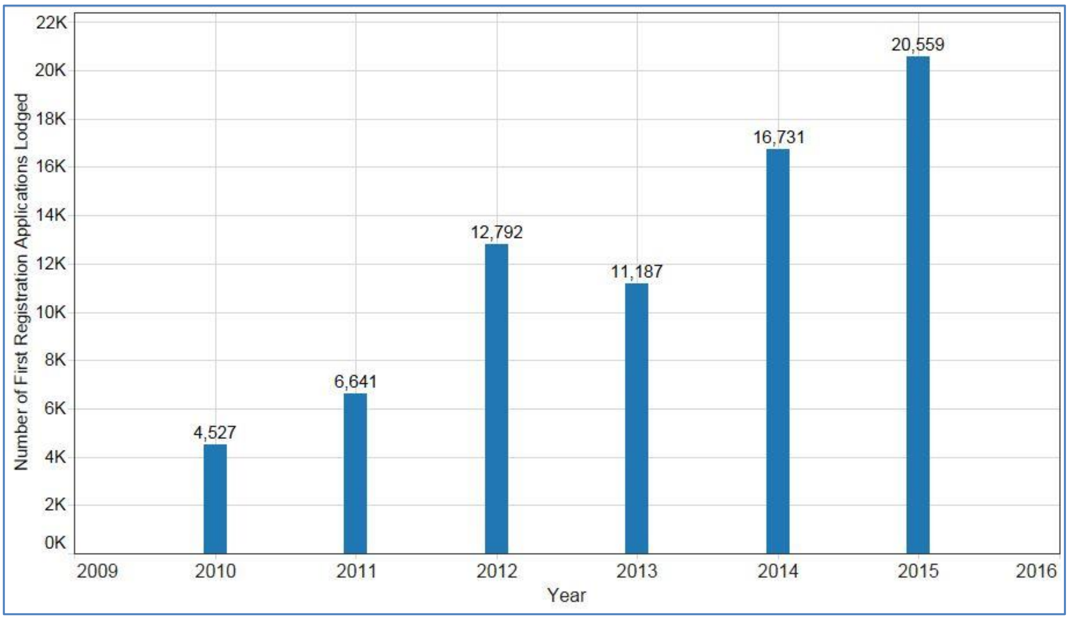


Figure 5.17 - Increase in Applications 2010-15

When analysed by type, all four categories have increased over the six years. In particular, intake of Form 3 applications has soared by 700%, while Form 5 cases have shown only a minor increase. The fluctuations in Form 1 and Form 2 intake between 2011 and 2013 mirror each other, and intake has increased by 250% and nearly 400% respectively.

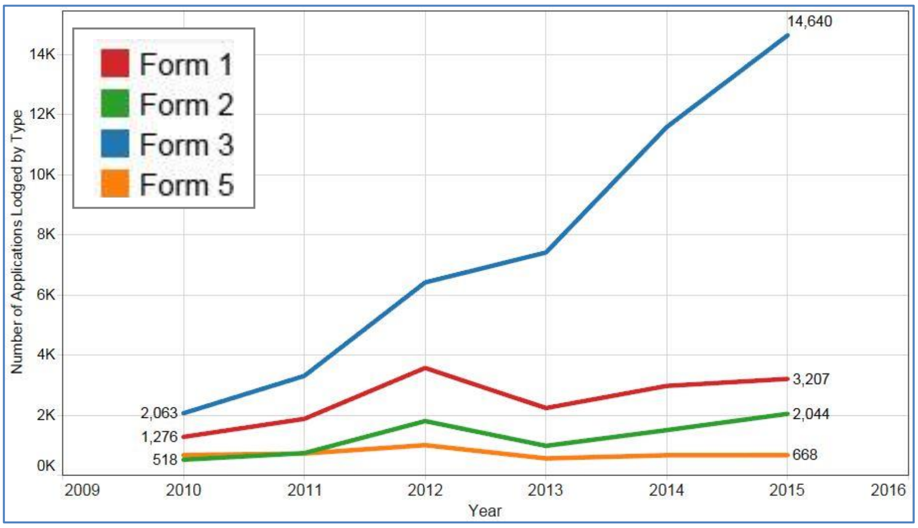
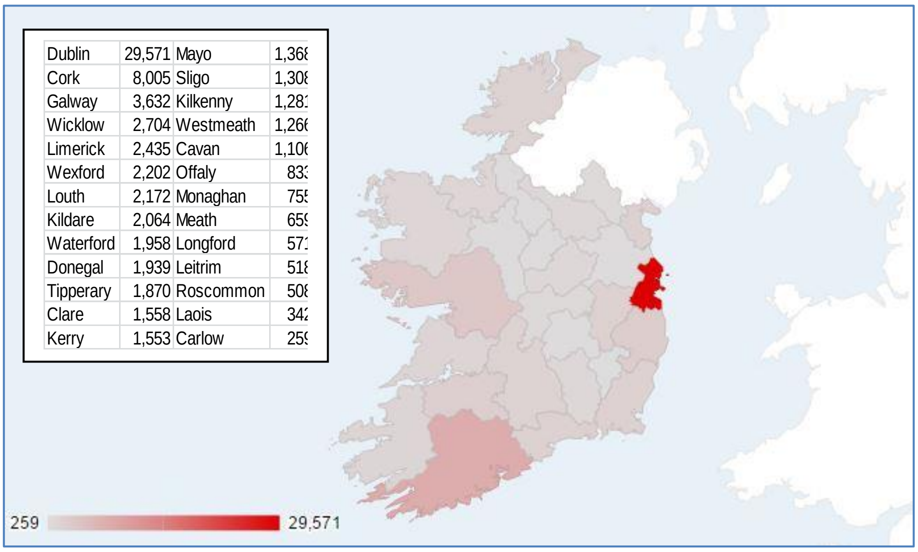


Figure 5.18 - Intake by Application Type, 2010-15

Given that registration of applications is now compulsory across Ireland, it would appear reasonable to assume that there is a correlation between the number of houses sold in a given county, and the number of applications lodged pertaining to that county. However, on inspection the relationship is not quite so clear cut.

An initial geographic analysis shows that Dublin has by far the highest number of applications – 29,571 as compared to the next highest figure in Cork at 8,005 applications. This is unsurprising, given that Dublin has the greatest proportion of unregistered land at up to 24% (see Figure 3.1, Section 3.4.1) and that it was also one of the last counties to be included in legislation for compulsory registration in 2011 (see Table 3.1, Section 3.4.1).



Dublin	29,571	Mayo	1,368
Cork	8,005	Sligo	1,308
Galway	3,632	Kilkenny	1,281
Wicklow	2,704	Westmeath	1,266
Limerick	2,435	Cavan	1,106
Wexford	2,202	Offaly	833
Louth	2,172	Monaghan	755
Kildare	2,064	Meath	659
Waterford	1,958	Longford	571
Donegal	1,939	Leitrim	518
Tipperary	1,870	Roscommon	508
Clare	1,558	Laois	342
Kerry	1,553	Carlow	259

Figure 5.19 - No of first registration applications lodged per county, 2010-15

By creating a scatter plot showing the number of houses sold versus the number of applications lodged, a strong relationship is revealed. The coefficient value of 0.84 demonstrates that as more houses are sold, more applications are lodged.

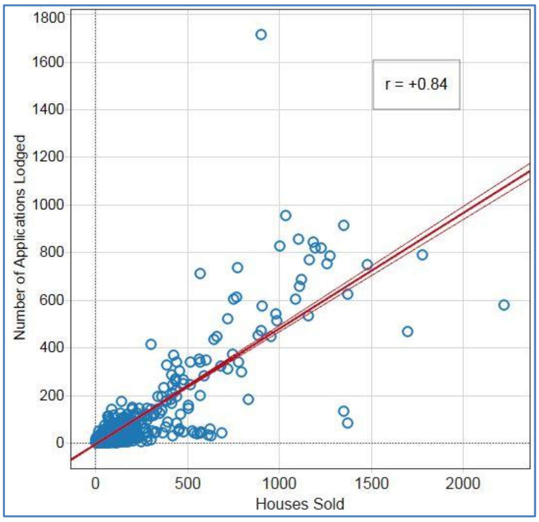


Figure 5.20 - Correlation between House Sales and Application Intake 2010-15

However, by filtering the results by county, a different picture emerges. If the counties selected are restricted to only Dublin, Wicklow and Wexford – the counties with the lowest proportion of registered land – then the coefficient value increases to 0.9, indicating a very strong relationship between the number of houses sold and applications lodged. Conversely, if the counties are filtered to include only the most highly registered counties of Carlow and Meath, then the correlation coefficient drops dramatically to 0.37, showing a much weaker analysis.

This indicates that two additional factors must also be included in the modelling stage – the date at which compulsory registration was rolled out, and the percentage of land registered in a given county.

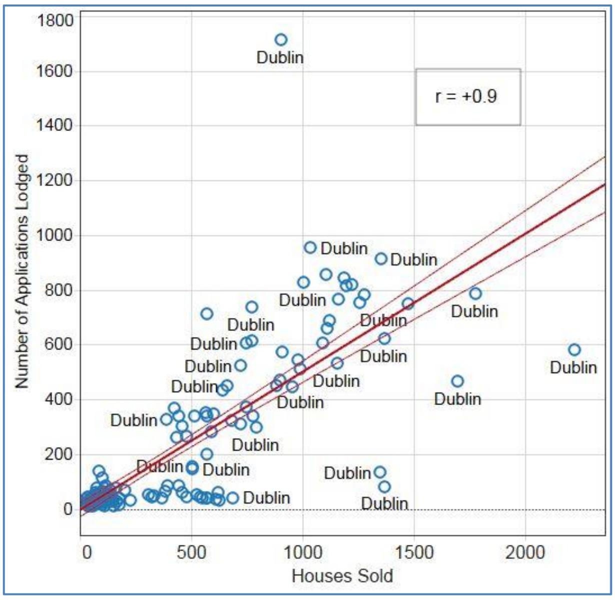


Figure 5.21 - Correlation between House Sales & Application Intake – Dublin, Wicklow, Wexford only

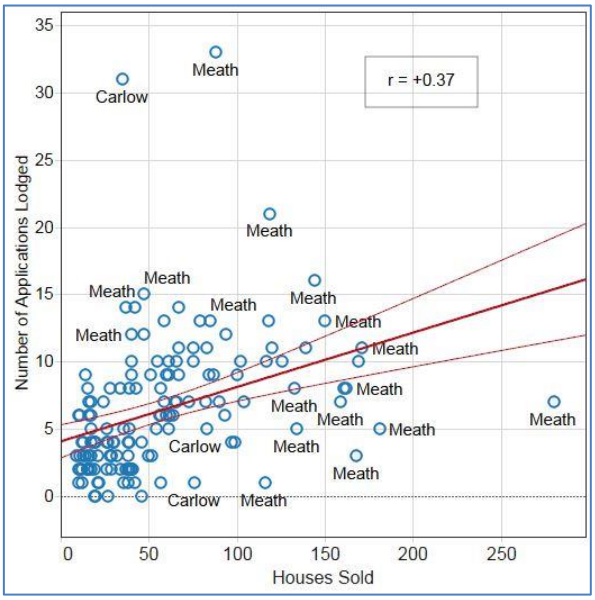


Figure 5.22 - Correlation between House Prices & Quantity of Sales – Carlow, Meath only

5.9. *Key Findings*

The key findings from this research are as follows:

- **Population -** Population in Ireland is concentrated in urban areas. There has been a steady migration from rural to urban living over the last fifty years.
- **House Sales -** The largest number of house sales takes place in densely populated urban areas, but the largest number of sales per head of population is in the region of the north west. The number of new houses sold has not recovered from the recession to the same extent as second hand properties – however, this could be due to a low

number of new houses coming on the market as the construction industry struggles to recover.

- **House Prices -** There is a strong correlation between house price and number of sales in the biggest cities of Dublin and Cork, but this is less pronounced in other counties. Population density does not appear to be highly correlated with house prices, and while average sale prices are starting to recover after the recession, they are still well below the height of the peak.
- **Unemployment Rates** – Rates have been falling consistently since 2012.
- **The Consumer Price Index -** the CPI (Housing) data slows much greater variation than the CPI (All) rate. This variance demonstrates that the effects of the crash on the property sector were far more severe than in others areas of the economy.
- **Intake of Applications for First Registration** – Overall intake has risen by over 400% between 2010 and 2015, most significantly for Form 3 applications. Dublin has the highest intake, and there appears to be a strong correlation between the number of houses sold and applications lodged in Dublin also. This trend is weaker in counties where registration has been compulsory for longer.

The findings also show that the date at which compulsory registration was rolled out and the percentage of land already registered in a given county are significant factors – therefore these two additional factors must be included in the final dataset for modelling. When added, the 'Compulsory Registration' dates range from 01/01/1970 to 01/06/2011 and the '% Registered' column holds values between 76-96%.

5.10. Conclusions

The visualisations presented in this chapter provide insight into the patterns and trends implicit in the dataset. The next chapter will build on these findings, using them to guide the process of setting parameters and training each model.

6. MODELLING & RESULTS

6.1. Introduction

This chapter will outline the design, implementation and results of the predictive modelling experiments performed. The dataset has been prepared for processing and will be divided into a training and test set. Based on the literature reviewed in Chapter 2 and the dataset assembled, the following models have been chosen for the predictive experimentation part of this project: Decision Trees, Random Forests, Multivariate Linear Regression and Artificial Neural Networks. These four distinct model types will be trained and tested using R Studio in a series of experiments. Finally, the results will be recorded and presented for comparison and evaluation in Chapter 7.

6.2. Splitting the Dataset

At the Evaluation stage, it will be necessary to analyse the various models produced and evaluate the comparative performance of each. Thus, it is necessary to split the dataset into a training set and a test set to facilitate the comparison of the predictions outputted by the training set against the test set to determine which performs the best.

There are various methods of splitting a dataset. The most common approach is the standard hold-out test set method, whereby the available data is split to create a Training Set and a Test Set (Han et al., 2012). The ratio for splitting can vary, but a 60:40 split is commonly applied. The 60% cut is used to train a series of models, while the 40% cut is held in reserve as the test set until the models are complete, and is then used to evaluate performance.

Another approach is cross validation. This method partitions the data into N partitions or ‘folds’ – typically N is set to 10, and the method becomes ‘10-fold cross validation’. 10 separate evaluation experiments are then performed on the data. The advantage of cross validation is that it makes maximum use of the dataset provided, although it can be slower to train than other methods.

In this instance, the hold-out test set method was chosen as the optimal method of splitting the dataset. After splitting, the Training Set includes 1105 rows, and the Test Set 737 rows – giving a total dataset of 1,842 records, across 24 variables.

6.3. Experiment Structure

There are several steps in building any experiment with R Studio.

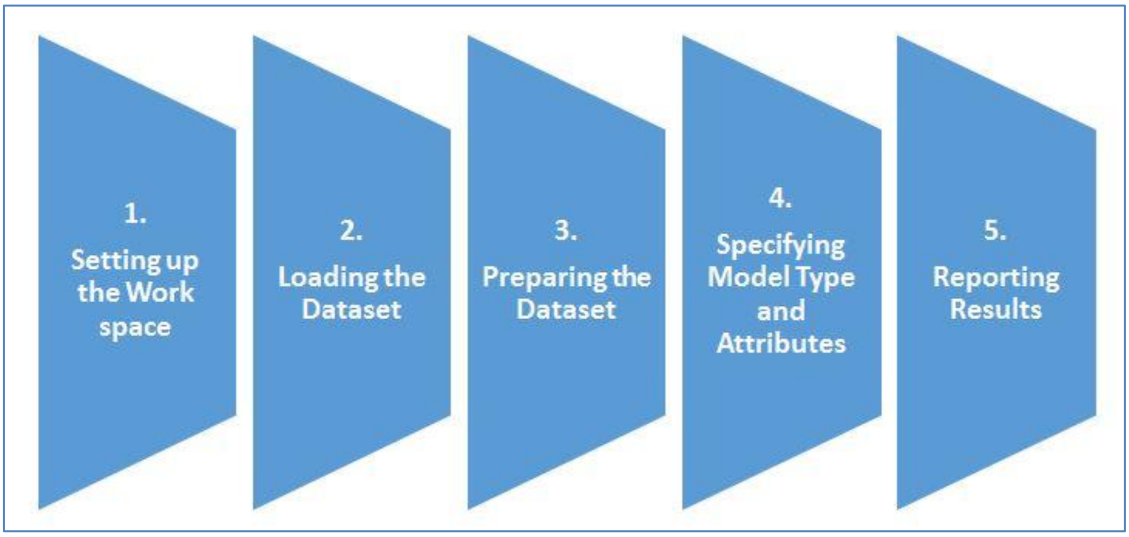


Figure 6.1 - Stages in model development

1. **Setting up the Workspace -** The first task is to install any required packages and open libraries as necessary. One of the principal advantages of using R Studio is the availability of free 'packages' - libraries of functions written by R users and made available online for free download from the CRAN repository[19]. For instance, the 'ggplot' package is a visualisations package, which enhances the inbuilt plotting functions of R and simplifies the process of outputting graphs and charts[20].
2. **Loading the Dataset** – There are various methods for importing data into R Studio, but the simplest is simply to load the training and testing sets as .csv files using the *read.csv()* command.
3. **Preparing/Manipulating the Dataset** – Numerous R Studio packages exist to assist in the preparation of data – for instance the 'stringr' package for string manipulation or the 'lubridate' package for standardising date formats.
4. **Specifying Model Type and Attributes** – In order to build a model, an appropriate algorithm and the necessary variables are specified in code. For example, a very basic linear regression could be performed using:

   lm(formula = ROLL ~ UNEM, data = datavar)

[19] Grolemund, G. (2016, April 16). Quick list of useful R packages. Retrieved from R Studio Support: https://support.rstudio.com/hc/en-us/articles/201057987-Quick-list-of-useful-R-packages

[20] Wickham, H. (2013). ggplot2. Retrieved from ggplot2: http://ggplot2.org/

In this instance '*lm*' specifies a linear regression function, *ROLL* is the attribute to be predicted (being future enrolment on a course) and it is to be analysed by reference to the unemployment rate, or *UNEM*. The *data* = *datavar* references the name of the dataset.

The experiments outlined in this chapter will use the *lm* function for regression, and also the 'rpart' (regression tree), 'randomForest*'* and 'neuralnet*'* packages.

5. **Reporting Results** – The final step in modelling in R Studio is to interpret and communicate the results. This is achieved in various ways – for example, by plotting models graphically, or by producing performance indicators such as p-values and R^2 coefficients for evaluation.

6.4. Basic Statistics

After setting up the workspace and loading the dataset as described above, a series of descriptive statistics were performed. The results can be seen in the table below.

Variable	Min	Max	Median	Mean	Standard Deviation
Population	31,798	1,273,069	120,047	176,471	238,725.83
New_Houses ***(new houses sold/month)***	0	414	7	13.8	26.03
Sec_Hand_Houses ***(second hand houses sold/month)***	2	1,908	40	81.03	151.99
New_Price ***(avg. price new house €)***	20,000	2,069,016	144,836	156,938	96,226
Sec_Hand_Price ***(avg. price 2nd hand house €)***	35,395	455,293	136,499	153,649	64,864.97
All_Price ***(avg. price of either type of house €)***	38,053	584,049	138,444	153,688	62935.56
CPI_All	-0.8	1	0	0.07	0.39
CPI_Housing	-1.6	3.5	0.1	0.24	0.86
GDP	38,859	47,577	40,619	41,577	2,366.52
Unemployment	8.8	15.2	13.6	12.84	2.01
Percent_Reg	76	96	92	90.73	4.7
Form_1	0	498	3	8.09	23.14
Form_2	0	141	0	1.38	6.46

Form_2_EX	0	281	0	2.67	11.76
Form_3	0	758	9	22.05	57.15
Form_3_MSB	0	184	0	0.86	8.12
Form_3_EX	0	39	0	1.36	2.64
Form_5	0	58	2	2.29	2.95
Count	0	1,714	18	38.69	101.17

Table 6.1 – Descriptive statistics

Certain variables were excluded from this examination; Year, Month, County and Compulsory Registration Year. These variables are ordinal or categorical; therefore they have no ‘min’ or ‘max’ as numeric variables do.

The basic statistics – min, max, median and mean – give an overview of the range and some preliminary insight into the distribution of each variable:

- MIN & MAX – the smallest and largest figure per variable. Comparing the MIN to the MAX provides the range.
- MEDIAN – the ‘middle’ number in the dataset – the number for which half the observations are higher, and half are lower.
- MEAN – the average of all the observations for a given variable.
- STANDARD DEVIATION – standard deviation quantifies the extent to which elements of a dataset differ from the mean (Donnelly, 2007); in other words, how spread out it is from the central point. A normally distributed attribute is one where 99.7% of values are within 3 standard deviations of the mean.

6.4.1. Normality

It is important to analyse the distribution of the dataset, since some algorithms only function correctly with data of a particular profile – assuming a dataset follows a normal distribution when in fact it does not will result in inaccurate results.

There are several ways to test an attribute for normality. Firstly, data can be assessed in terms of skew, and kurtosis.

Skew quantifies the extent to which data is symmetrical around the mean – if it exhibits the same profile to the left and the right[21]. When data is normally distributed, the skew value is 0. Meanwhile, kurtosis is concerned with outliers – these are the single observations that differ significantly from others in the dataset. When plotted as a histogram, these figures are shown as ‘tails’ – kurtosis measures the extent of these tails[21]. When data is normally distributed the kurtosis value is 0 – any higher than 3 excludes the possibility of normality.

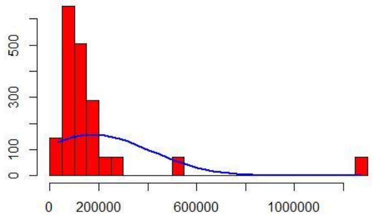
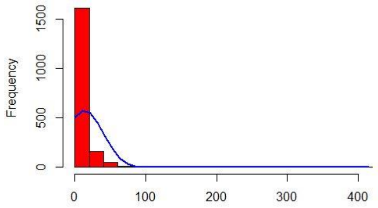
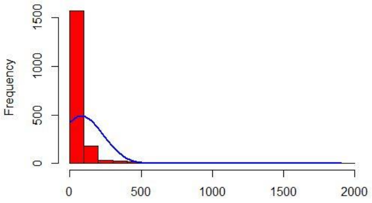
Most variables in the dataset show relatively high to extremely high levels of skew and kurtosis.

Other methods of testing data for normality include graphing the data and analysing the shape of the distribution by eye, or using a statistical test such as the Shapiro-Wilks test. This test evaluates for goodness of fit by analysing a sample of the data using the formula:

$$W = \frac{(\Sigma_{i=1}^{n} a_i x_{(i)})^2}{\Sigma_i^n (x_i - \overline{x}\ )^2}$$

When applied using R-Studio, the ‘*shapiro*’ function returns a p-value. Using the typical threshold of 0.05, p-values lower than this point can be deemed to deviate from normality.

The results of calculating the skew and kurtosis, plotting histograms and performing Shapiro-Wilks tests for each variable can be seen in the table of graphics below; it is clear that in this instance that none of the variables follow a strictly normal distribution, and some diverge from it severely. The closest to a normal distribution is the CPI_All variable.

Population Shapiro-Wilks - 2.761726e-59 Skew – 3.80 Kurtosis - 14.31 	**New_Houses** Shapiro-Wilks - 1.268482e-61 Skew – 7.46 Kurtosis – 79.58 	**Sec_H_Houses** Shapiro-Wilks - 5.377119e-61 Skew – 5.44 Kurtosis – 36.8 

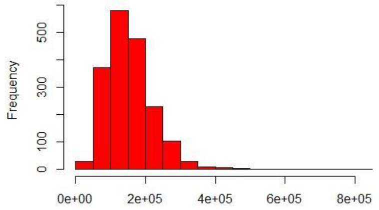
[21] National Institute of Standards and Technology. (2012, April). 1.3.5.11. Measures of Skewness and Kurtosis. Retrieved from Engineering Statistics Handbook: http://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm

New_Price

Shapiro-Wilks - 9.266512e-30

Skew – 9.26

Kurtosis - 162.12

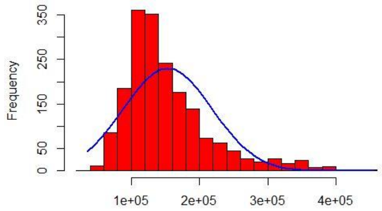


Sec_H_Price

Shapiro-Wilks - 3.237554e-36

Skew – 1.52

Kurtosis – 2.64

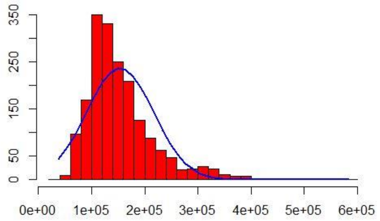


All_Price

Shapiro-Wilks - 8.177314e-35

Skew – 1.51

Kurtosis - 3.19

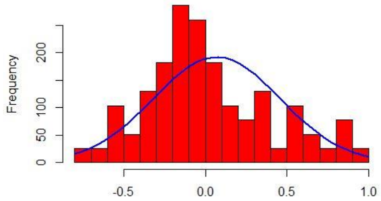


CPI_All

Shapiro-Wilks - 7.035167e-22

Skew – 0.47

Kurtosis - -0.17

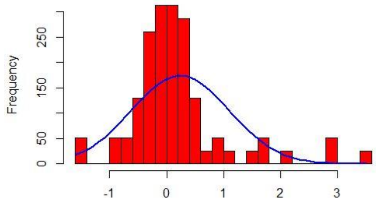


CPI_Housing

Shapiro-Wilks - 4.367645e-40

Skew – 1.60

Kurtosis – 3.82

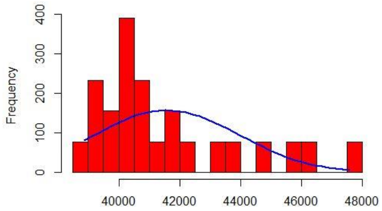


GDP

Shapiro-Wilks - 5.465786e-38

Skew – 1.10

Kurtosis – 0.17

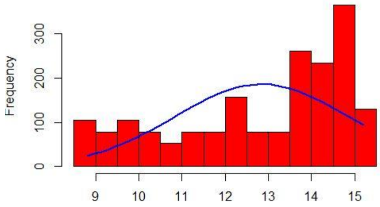


Unemployment

Shapiro-Wilks - 1.648824e-35

Skew – -0.68

Kurtosis - -0.89

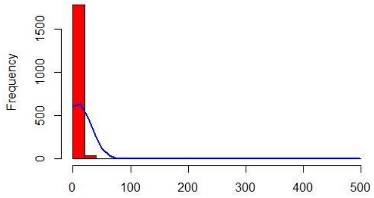


Form_1

Shapiro-Wilks - 4.856482e-65

Skew – 9.00

Kurtosis – 127.84

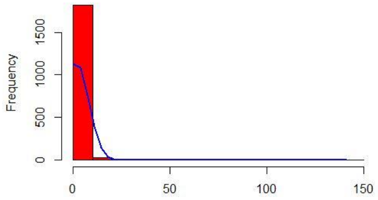


Form_2

Shapiro-Wilks - 5.92973e-67

Skew – 11.99

Kurtosis - 204.44

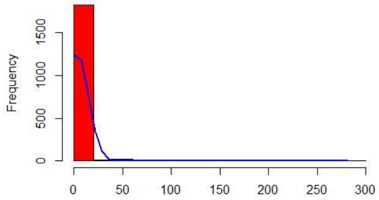


Form_2_EX

Shapiro-Wilks - 1.114058e-66

Skew – 11.01

Kurtosis - 191.64

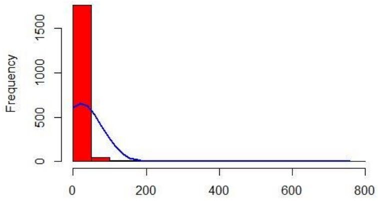


Form_3

Shapiro-Wilks - 9.514263e-64

Skew – 6.43

Kurtosis – 48.59

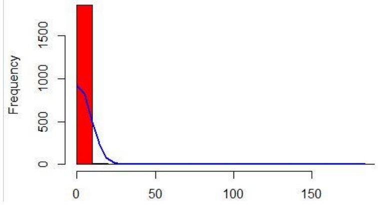


Form_3_MSB

Shapiro-Wilks - 9.168834e-70

Skew – 18.12

Kurtosis - 352.34



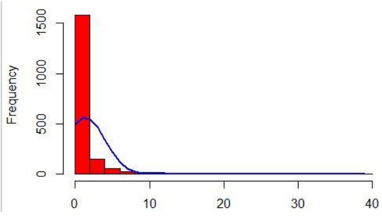
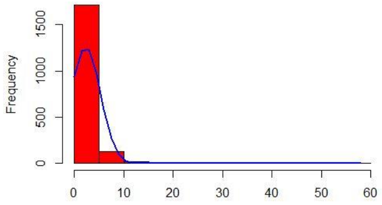
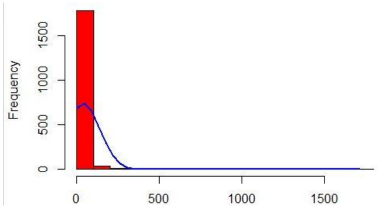
Form_3_EX	**Form_5**	**Count**
Shapiro-Wilks - 1.43826e-56 Skew – 4.28 Kurtosis – 31.63	Shapiro-Wilks - 7.995661e-53 Skew – 5.86 Kurtosis - 78.63	Shapiro-Wilks - 1.643571e-64 Skew – 7.37 Kurtosis – 70.88
		

Table 6.2 - Shapiro-Wilks test results and histograms for each variable

The implication of this analysis is that it cannot be verified that the data is normally distributed, at a confidence interval of 95%. This is not uncommon – many datasets based on real life events do not show a strong central tendency. However, it does mean that certain statistical tools cannot be used – for example analysis of variance (ANOVA), since it requires an assumption of normality to function correctly.

6.4.2. Outliers and their Treatment

It is important to analyse the reasons for such large numbers of outliers in nearly every variable. The statistical exploration in Chapter 5 suggested some reasons for this already – for example Dublin’s population is far greater than that in other counties, more houses tend to be sold in urban areas, and house prices tend to be higher in cities like Dublin and Cork and Galway. Therefore, removing instances such as these would at first glance appear to be a poor idea, in the interests of maintaining the accuracy and meaning of the dataset. However, it is useful to establish if the dataset is inherently non-normal in distribution, or if it tends towards normality once the most serious outliers are removed.

Thus, it is necessary to undertake a process of outlier detection. In order to simplify the process, given the high levels of skew in the application types, it is decided for the modelling stage to concentrate on the overall application type Count variable as the target feature. It might be possible as an extension of this project to attempt to predict individual application type intake in the future, but for the purposes of training the initial models, the overall Count variable is sufficiently challenging. Therefore the application types are removed from the modelling dataset. This leaves 10 independent variables and one dependent variable; Count.

To identify significant outliers, a univariate approach is taken. A boxplot is created for each variable in turn, with any observations that fall outside 1.5 times the interquartile range being identified as outliers and plotted in red.

On this basis, three variables return no outliers at all – CPI_All, GDP and Unemployment.

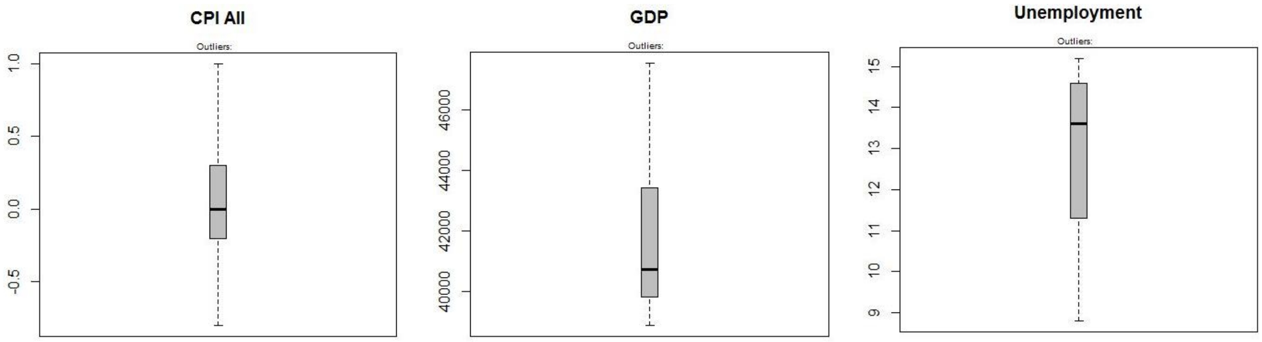


Figure 6.2 - Distribution of features without significant outliers

In the case of these variables, they have no significant outliers to be removed, and yet they do not follow a normal distribution. It is therefore not possible to transform these variables to follow a normal distribution through outlier manipulation.

The other variables show varying numbers of outliers – for example, the Population feature has only two outliers, representing the populations of Dublin and Cork, while the Count feature has many observations outside the threshold, one of which stands out as particularly distant from the boxplot whiskers.

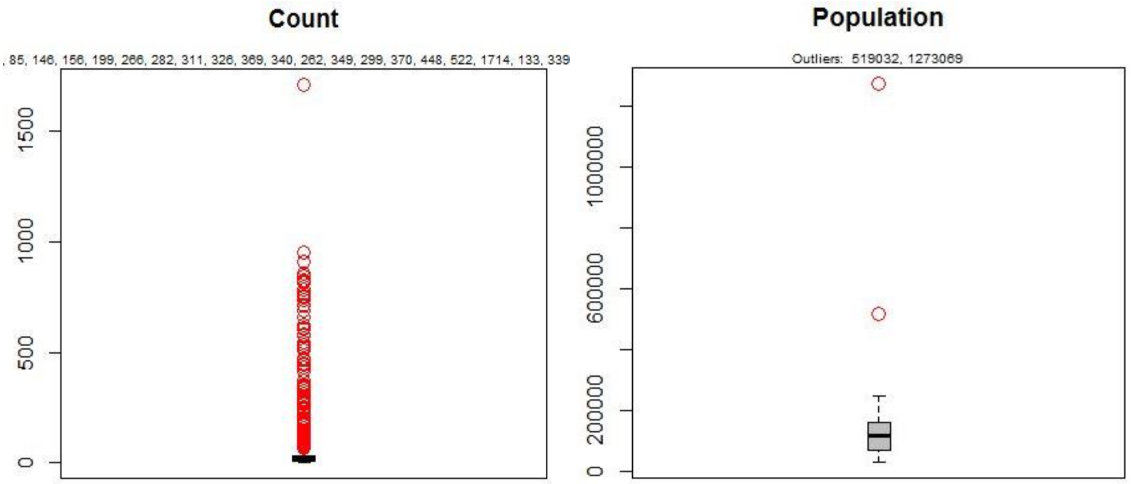
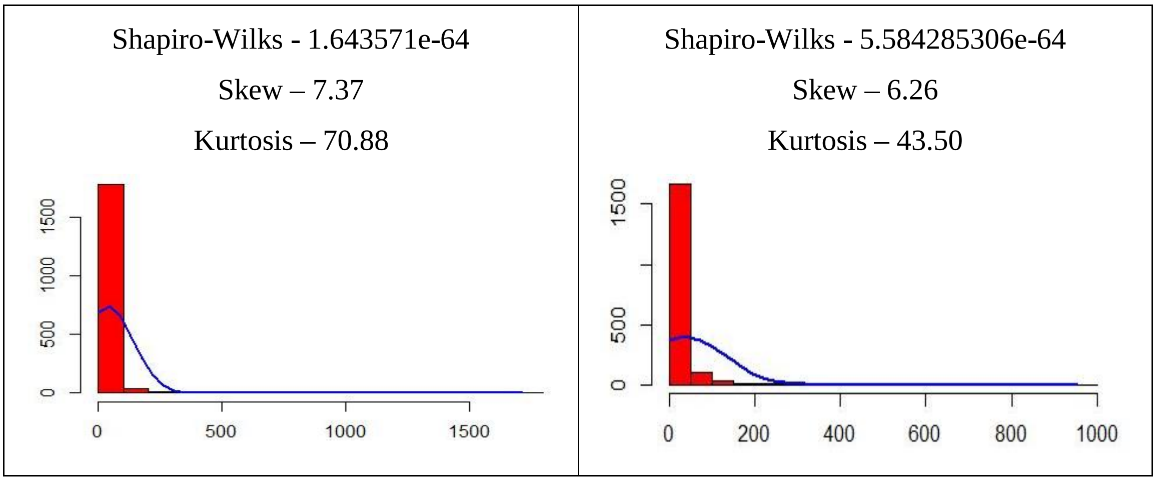


Figure 6.3 - Presence of outliers in Count & Population variables

The question, therefore, is how to handle outliers. If an observation is investigated and found to be due to an error, then it should be dropped – however, in this case all of the data points have been checked and are valid.

An outlier may be a unique observation that simply does not follow the pattern of the rest of the data. For instance in the Population variable, Dublin and Cork simply have larger populations than other counties. Likewise, the Count variable shows numerous outliers, one of which stands out as an instance of particularly high intake – 1,714 applications, lodged in

Dublin in November 2012, which is nearly twice as high as the next highest observation of 953 applications. The plots below show the impact of removing/replacing these outliers on distribution:



Shapiro-Wilks - 1.643571e-64

Skew – 7.37

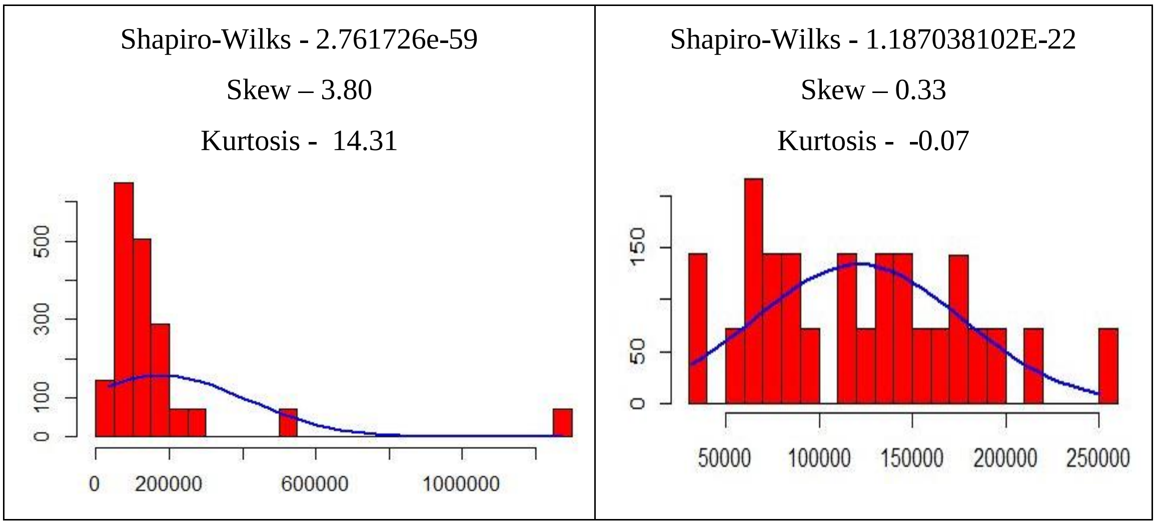
Kurtosis – 70.88

Shapiro-Wilks - 5.584285306e-64

Skew – 6.26

Kurtosis – 43.50

Figure 6.4 – Count: distribution before and after removing largest outlier



Shapiro-Wilks - 2.761726e-59

Skew – 3.80

Kurtosis - 14.31

Shapiro-Wilks - 1.187038102E-22

Skew – 0.33

Kurtosis - -0.07

Figure 6.5 – Population: before and after replacing largest outliers with the mean

Removing or replacing such observations is a controversial topic. The advantage of taking action is in reducing skew and helping the distribution to approach normality, as can be seen to some extent in the plots above. This opens up a wider range of potential modelling techniques and evaluation metrics. However, removal or replacement with the mean could equally undermine the veracity of the predictions made during the modelling process. Removing data reduces the information that can be used to train models, and replacing it with a value – for instance the mean – is bound to distort the findings. Therefore, for the purposes of this project it is decided not to remove any outliers from the modelling dataset.

6.4.3. Associations

In order to examine potential correlations between the target variable Count and the other variables, a scatterplot matrix was created:

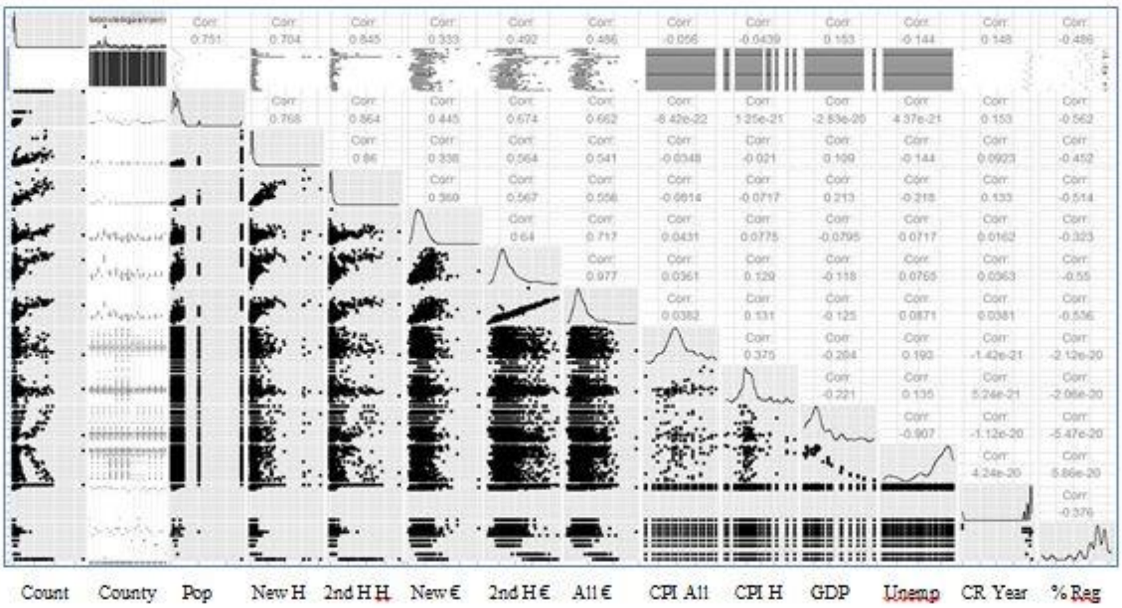


Figure 6.6 - Scatter plot of multiple variables

The number of variables makes this scatter plot difficult to read, but several variables stand out both in shape and in terms of correlation coefficient value. Values close to 1 indicate a positive correlation; close to -1 implies a strong negative correlation, and near 0 suggests a lack of correlation. The variables with a strong positive correlation with the target variable Count are shown in the smaller scatter plot shown below; Population, New Houses Sold and 2nd Hand Houses Sold.

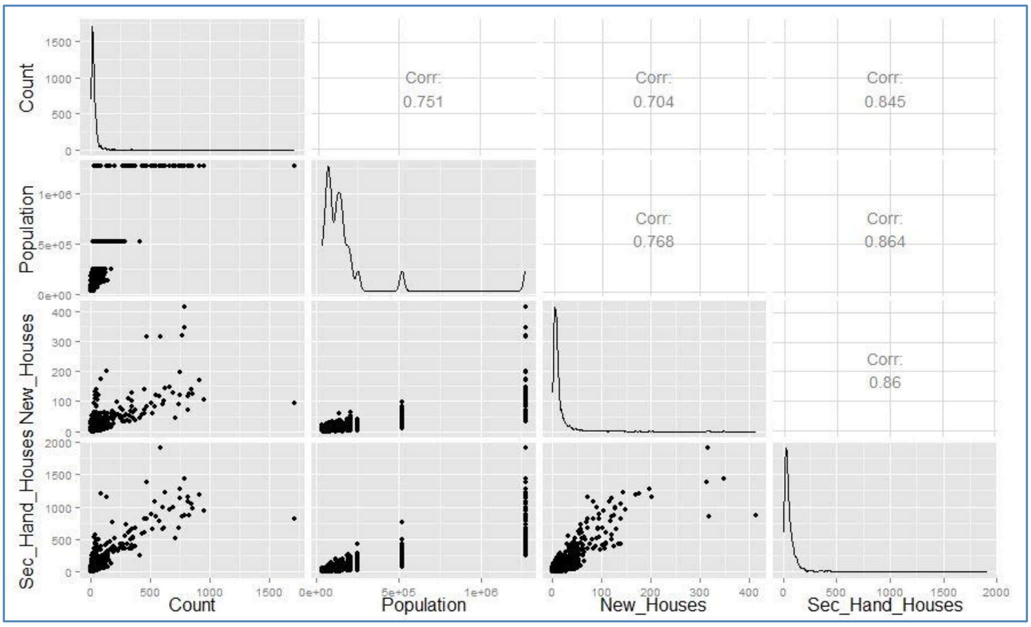


Figure 6.7 - Scatter Plot showing variables correlated to Count

Since the New Houses Sold and 2nd Hand Houses Sold variables have very similar profiles, a new column is created in R, representing an aggregate of both variables. This new column is called Houses_Sold.

6.5. *Predictive Models & Results*

This section will describe the building and training of models, and outputting of results in a format that can be evaluated in Chapter 7.

6.5.1. Regression Trees

The regression tree models are built using the ‘rpart’ package.

Model 1

All of the variables are included for consideration in the initial model, and the ‘cp’ (cost complexity factor) is set to an initial value of 0.0.

When modelled, this tree includes all variables and comprises 82 separate splits. The ‘plotcp’ function is used to model the splitting points, and a function applied to find the optimal value for cp – this is returned as a very low value; 0.00001695847. This value represents the complexity parameter associated with the minimum possible cross-validated error. Pruning the tree using this value for cp reduces the number of splits to 49 and the number of variables to 9; however the resultant tree is large and complex, and may be overfitted.

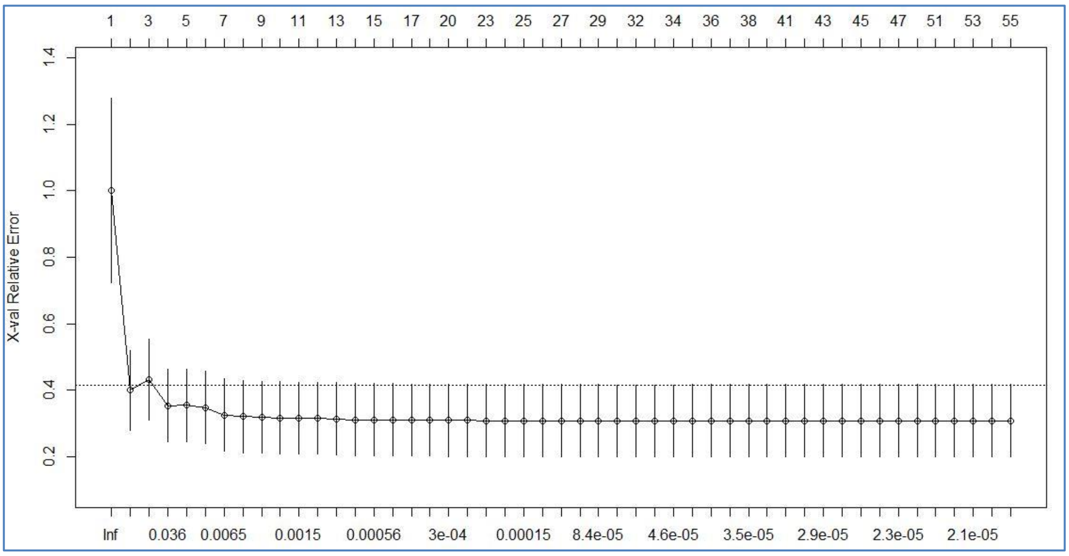


Figure 6.8 - Plot showing number of splits after pruning.

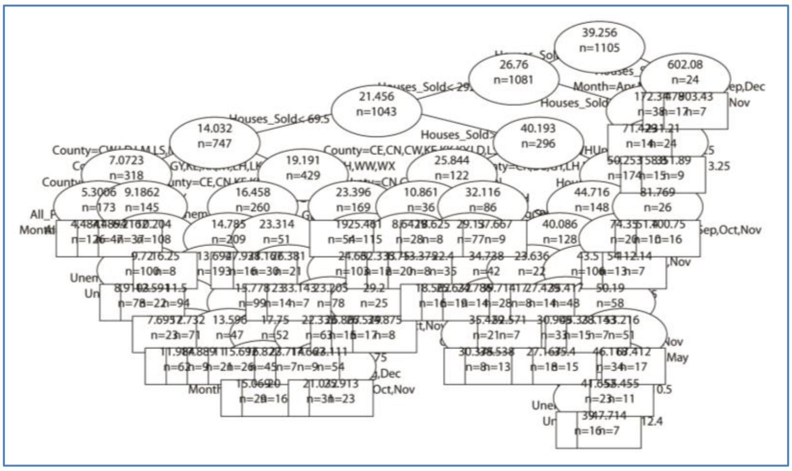


Figure 6.9 - Model 1, Regression Tree

Nonetheless, the model may be accurate - the final steps are to run the model against the test set and build a dataframe of the results for evaluation in the next chapter.

Model 2

Using the same initial variables, the model is rerun, but the parameter ‘minsplit=25’ is added as a parameter, to ensure that before a split is attempted, the minimum number of observations in a node is at least 25. This should reduce the number of small leaves based on only a few instances. The decision is also taken to prune the tree manually from the point after the 5th split, after which the cp values remain relatively static. This returns a tree with only two variables; Houses_Sold and Year.

The resultant tree is plotted, as can be seen below, compared to the test set and the results recorded for evaluation.

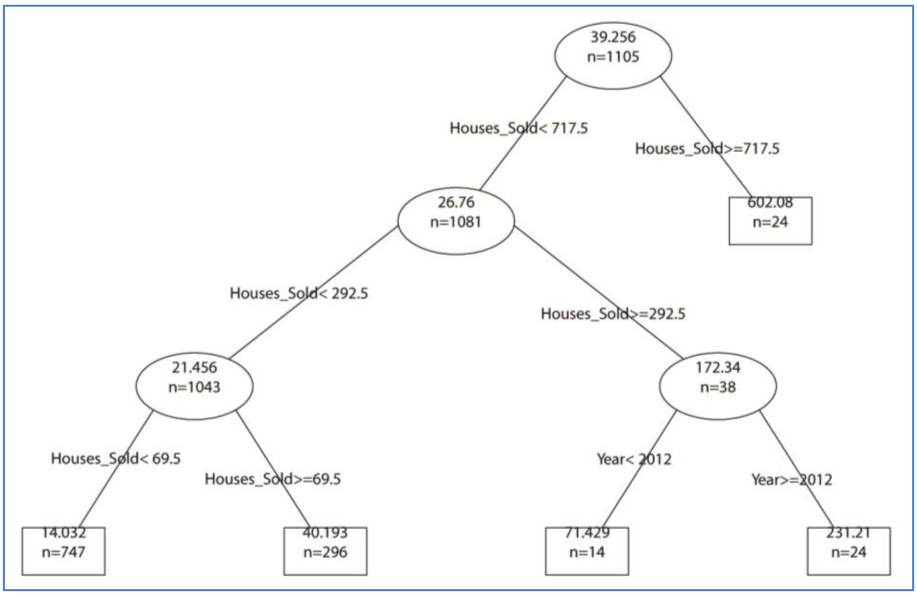


Figure 6.10 - Model 2, Regression Tree

6.5.2. Random Forests

Model 3

To create a random forest model, the 'randomForest' package is applied to the training set. This creates an ensemble of 500 trees, with four variables tried at each split. An analysis of each predictor used shows that Houses_Sold is the most important variable, followed by Percent_Reg, Population and County.

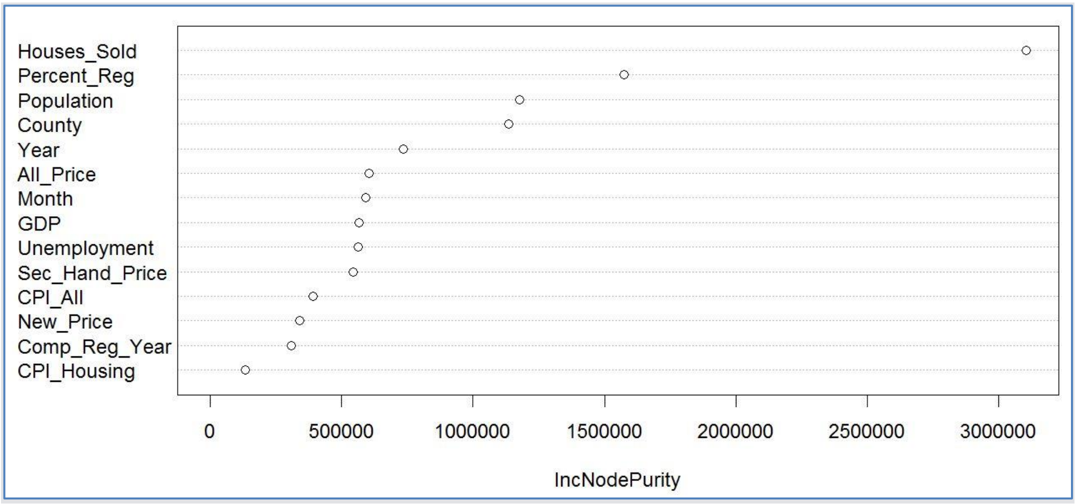


Figure 6.11 - Variables in order of importance

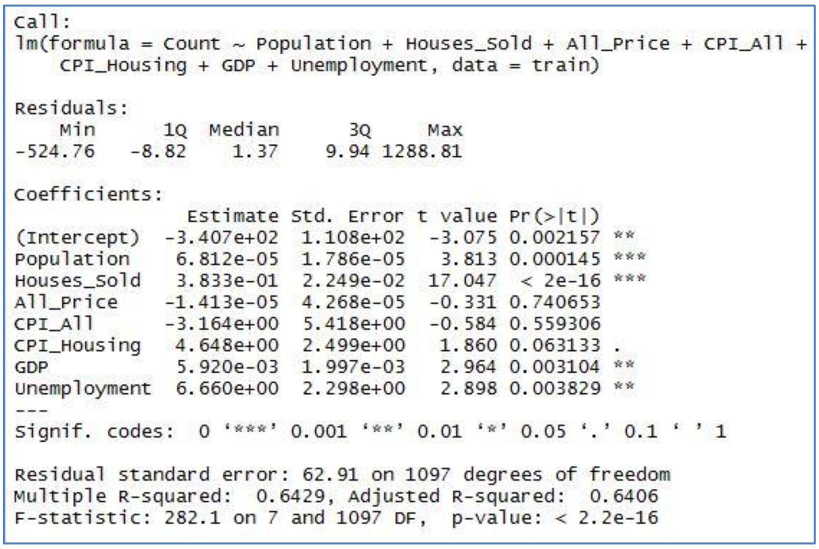
The final model explains just over 71% of variance in the dataset, and was run against the test set as before.

6.5.3. Multivariate Linear Regression

Multivariate regression was performed on the training set excluding the ordinal and categorical variables Year, Month, County, and Comp_Reg_Year, since these data types cannot be meaningfully interpreted in a regression model.

Model 4

The first regression model is built using seven variables – Population, Houses_Sold, All_Price, CPI_All, CPI_Housing, GDP and Unemployment The initial model is fitted and the results analysed.



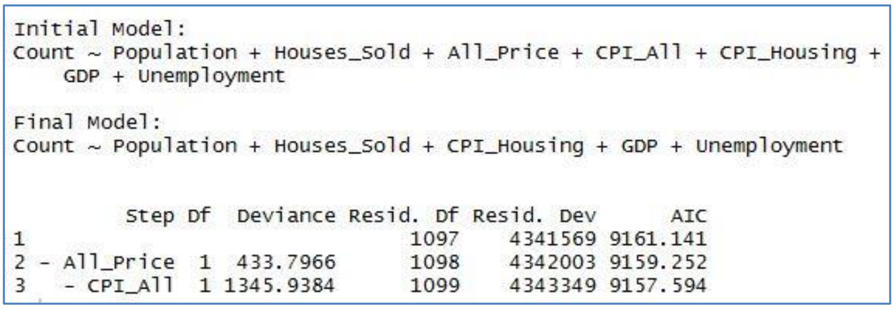
```
Call:
lm(formula = Count ~ Population + Houses_Sold + All_Price + CPI_All +
    CPI_Housing + GDP + Unemployment, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-524.76   -8.82    1.37    9.94 1288.81

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.407e+02  1.108e+02  -3.075 0.002157 **
Population    6.812e-05  1.786e-05   3.813 0.000145 ***
Houses_Sold   3.833e-01  2.249e-02  17.047  < 2e-16 ***
All_Price    -1.413e-05  4.268e-05  -0.331 0.740653
CPI_All      -3.164e+00  5.418e+00  -0.584 0.559306
CPI_Housing   4.648e+00  2.499e+00   1.860 0.063133 .
GDP           5.920e-03  1.997e-03   2.964 0.003104 **
Unemployment  6.660e+00  2.298e+00   2.898 0.003829 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 62.91 on 1097 degrees of freedom
Multiple R-squared:  0.6429, Adjusted R-squared:  0.6406
F-statistic: 282.1 on 7 and 1097 DF,  p-value: < 2.2e-16
```

Figure 6.12 - results of first regression model

As the R^2 value shows, this model explains up to 64% of the variance in the dataset, and the overall p-value is low. These are encouraging results. However, when analysing the variables individually, it is clear that some are having a greater impact than others on the model. Using the standard threshold of 0.05 (Donnelly, 2007), All_Price, CPI_All and CPI_Housing can be seen to be statistically insignificant. Houses_Sold appears to be the most influential variable, followed by Population.

To improve the model, stepwise regression is used to remove the weakest variables one by one (starting with All_Price), until the point where the R^2 value ceases to improve has been reached. This point is reached after two steps, leaving the final model with 5 remaining variables. All_Price and CPI_All are the variables removed.

In order to check this result, the ‘MASS’ package was used to conduct the stepwise regression automatically. This package performs both forward and backward regression and evaluates based on AIC (Akaike Information Criterion) rather than p-values. AIC measures goodness of fit by identifying the smaller residual errors in the model, and is therefore useful for avoiding overfitting. In this instance, the same two variables were identified as suitable for exclusion - All_Price and CPI_All.



```
Initial Model:
Count ~ Population + Houses_Sold + All_Price + CPI_All + CPI_Housing +
    GDP + Unemployment

Final Model:
Count ~ Population + Houses_Sold + CPI_Housing + GDP + Unemployment


         Step Df  Deviance Resid. Df Resid. Dev      AIC
1                               1097    4341569 9161.141
2 - All_Price  1  433.7966      1098    4342003 9159.252
3   - CPI_All  1 1345.9384      1099    4343349 9157.594
```

Figure 6.13 - Results of MASS stepwise regression

The optimum linear regression model identified is run against the test set and the results outputted for later analysis.

6.5.4. Neural Networks

The final model type to be implemented is a Neural Network, and for this the ‘neuralnet’ package is employed. This package only handles quantitative variables; hence for these first models the decision is taken to exclude the categorical variable County and the ordinal variable Month.

The next step is to normalise the dataset. Normalisation can be a controversial topic; however with neural network models it is difficult to extract accurate results without bringing data points within a common range. In this instance, the dataset was scaled by setting a minimum value of 0 and a maximum value of 1.

Models 5 & 6

In terms of setting parameters for the neural network model, one hidden layer is usually sufficient, as extra layers are not deemed to increase performance to any great degree. In order to test this theory, two version of this model are created; one with a single hidden layer and the other with three hidden layers. The two models created are as follows:

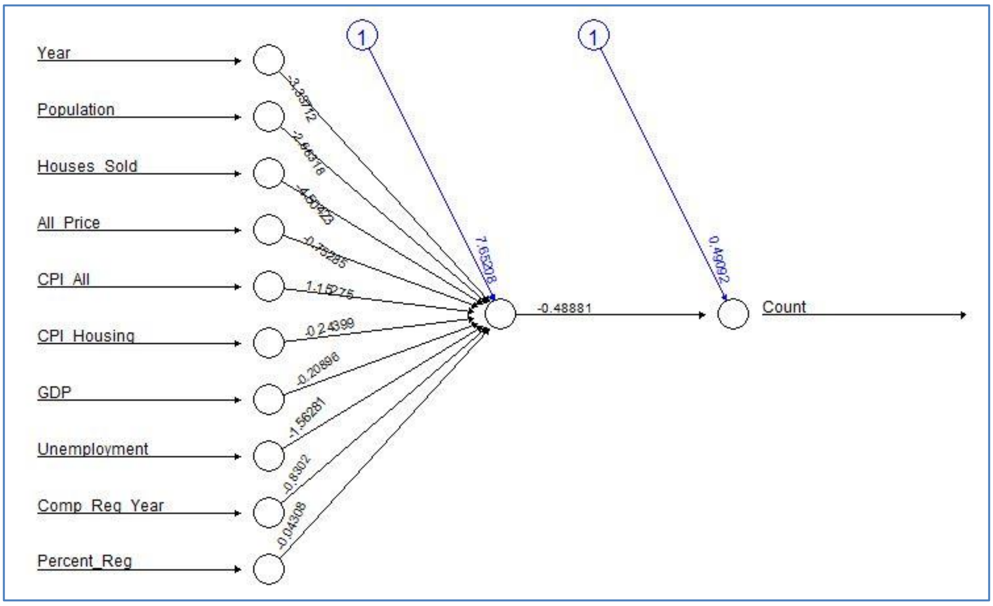


Figure 6.14 - Model 5; Neural Network with one hidden layer

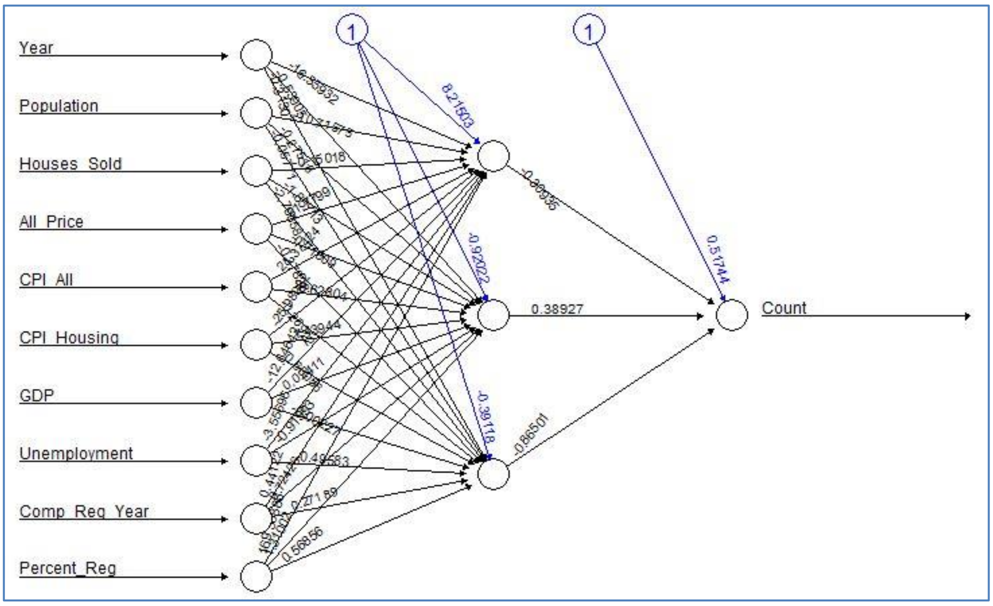


Figure 6.15 - Model 6; Neural Network with three hidden layers

The two models are run against the Count variable in the test set, and the results recorded.

Models 7 & 8

The final models generated include the variables excluded in the initial neural networks – County and Month. In order to include these variables, they must be processed into a quantifiable format. This is achieved by using the 'model.matrix' function to convert the data points into binary ("dummy") variables. Once this has been achieved, the data is normalised as before, and again two models are built using a single and three hidden layers.

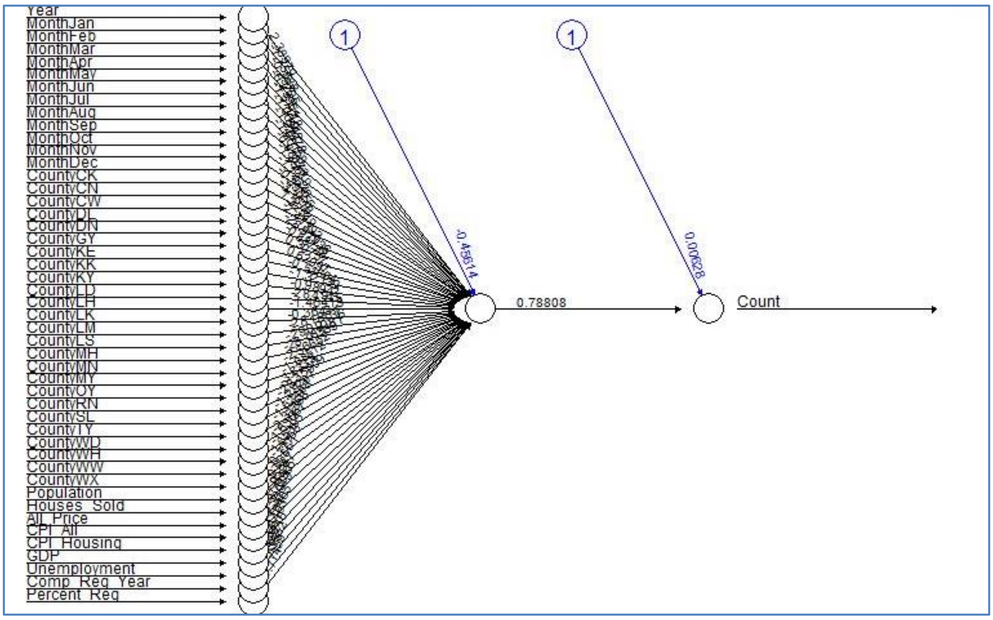


Figure 6.16 - Model 7; Neural Network with one hidden layer

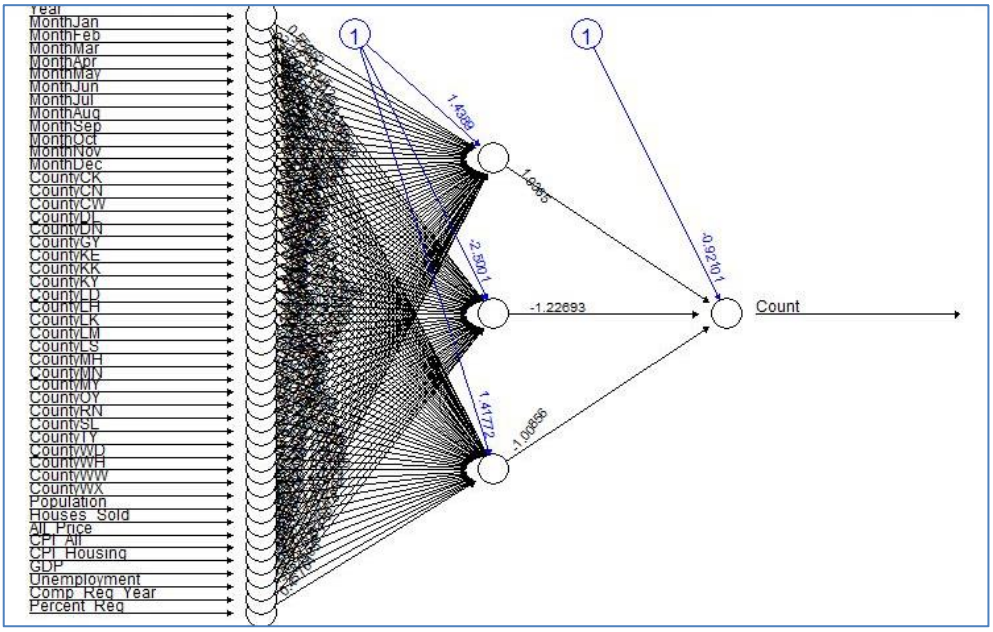


Figure 6.17 - Model 8; Neural Network with three hidden layers

6.6. Summary of Models Developed

Model #	Description
1	Regression Tree, no MinSplit, automatic pruning
2	Regression Tree, MinSplit = 25, manual pruning
3	Random Forest
4	Multivariate Linear Regression using stepwise regression
5	Neural Network – one hidden layer, only continuous inputs
6	Neural Network – three hidden layers, only continuous inputs
7	Neural Network – one hidden layer, categorical and continuous inputs
8	Neural Network – three hidden layers, categorical and continuous inputs

Table 6.3 - Summary of models developed

6.7. Conclusions

This chapter has explained the processes and principles used in creating a series of models based on a range of Machine Learning techniques. The concept of splitting a dataset has been explored, the structure of the modelling process has been described, and the dataset has been analysed in terms of distribution. The issue of outliers and the risks associated with removal or replacement has been discussed in full. The next chapter will evaluate and contrast the performance of these models, using a selection of methodologies.

7. EVALUATION & ASSESSMENT

7.1. Introduction

The purpose of this chapter is to evaluate the models designed in Chapter 6, using two error measures: Percentage Mean Absolute Error (%MAE) and Percentage Root Mean Squared Error (%RMSE). After ranking the models, ways in which the optimum model can be used to the benefit of the Property Registration Authority will be discussed, and opportunities for further enhancements will be identified.

7.2. Predictions

The model outputs for the eight trained models are as follows:

Model #	Description	Count (Actual)	Count (Prediction)
1	Regression Tree - automatic pruning	28,673	26,687
2	Regression Tree - manual pruning	28,673	27,636
3	Random Forest	28,673	28,214
4	Multivariate Linear Regression	28,673	27,077
5	Neural Network – one hidden layer, continuous inputs	27,841	28,770
6	Neural Network – three hidden layers, continuous inputs	27,841	28,545
7	Neural Network – one hidden layer, categorical and continuous inputs	27,841	28,635
8	Neural Network – three hidden layers, categorical and continuous inputs	27,841	30,428

Table 7.1 - Model outputs

As can be seen, all models correctly predicted Count within a range of 10% of the actual intake figure. The most successful was the Random Forest, which returned a figure less than 2% outside of the actual observed value, and the least successful was Model 8; the Neural Network model with three hidden layers and including continuous and categorical variables.

7.3. Statistical Analysis & Goodness of Fit

In order to train the models in Chapter 6, some small evaluation has already taken place – for example by reference to R^2 values or p-values. However, the presence of a high R^2 value is not a sufficient indicator of ‘goodness of fit’ on its own.

The term ‘Goodness of Fit’ refers to how well a model fits the data points. One of the principal challenges of tuning Machine Learning models is to avoid ‘underfitting’ or ‘overfitting’ the data. An overfitted model is overly complex and includes irrelevant data, while an underfitted model will be over-simplified and does not make use of potentially influential data points (Welling, 2011). Goodness of Fit, therefore, is a measure of how well a model fits the dataset, or how well it performs in predicting future occurrences.

It should be noted however, that prediction analysis is not an exact science. The statistician George E. P. Box wrote that:

> “*Essentially, all models are wrong, but some are useful”* (Box, 1987)

A statistical model is not a true depiction of reality; it is only a simplified limited representation of a particular set of data points. The function of a prediction model is to help the analyst to understand the structure and relationships of the observations, and to use mathematical formulas to project and extrapolate the data. No model, therefore, can be entirely accurate. Indeed, rather than attempting to identify the most precise model, evaluation measures actually try to find the model with the lowest errors; the one that is “least wrong”.

There are a range of potential evaluation measures for ascertaining Goodness of Fit, depending on the data being analysed and the choice of models employed. Since in this instance a wide range of exploratory models have been used, an error measure that can be applied to every model is required - %MAE and %RMSE fulfil these criteria.

7.4. Mean Absolute Error & Root Mean Squared Error

Mean Absolute Error is a measure of accuracy, and is calculated using the formula:

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n} |y_j - \hat{y}_j|$$

To calculate MAE, the extent to which each predicted value differs from the actual observation is measured, and the total of these errors is divided by the size of the dataset[22]. MAE is useful since it specifically measures the size of errors made in a prediction set –therefore, lower MAE

[22] The European Virtual Organisation for Meteorological Training. (n.d.). *Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)*. http://www.eumetcal.org/resources/ukmeteocal/verification/www/english/msg/ver_cont_var/uos3/uos3_ko1.htm

scores indicate a model with fewer inaccuracies. MAE is also less sensitive to large outliers than RMSE, since the errors are not squared in the formula[23].

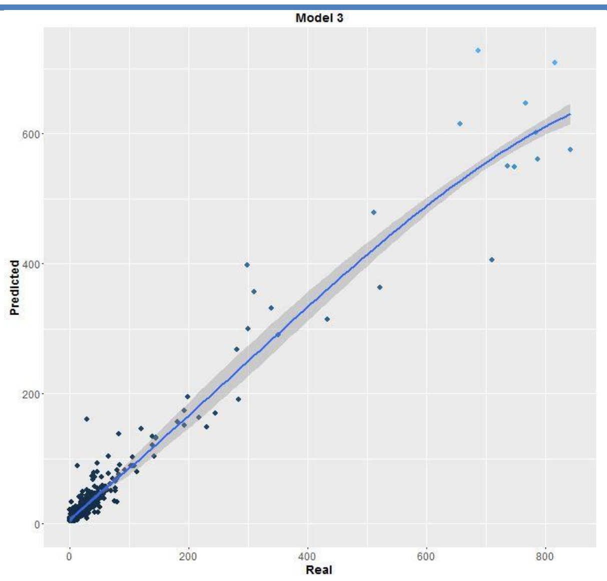
Root Mean Squared Error is a measure of the average magnitude of error[23]. To calculate RMSE, the differences between the prediction values and observation values are squared, and then the average of these figures is compiled. The mathematical formula is:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

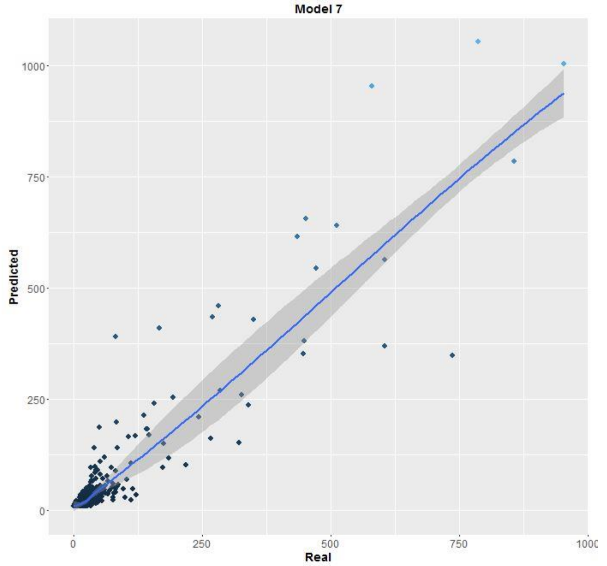
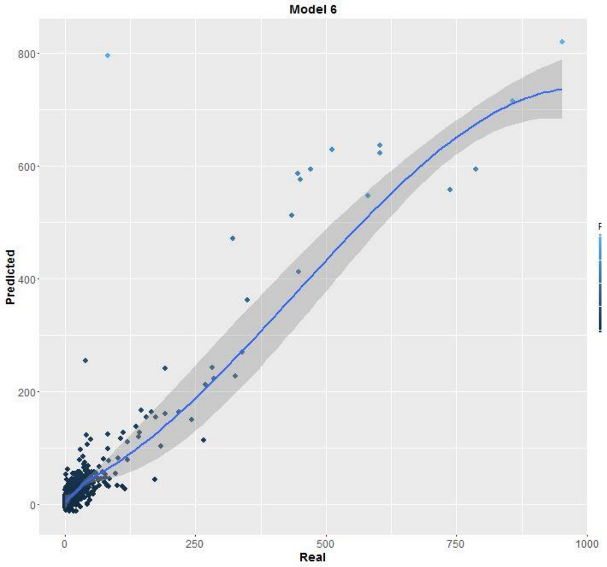
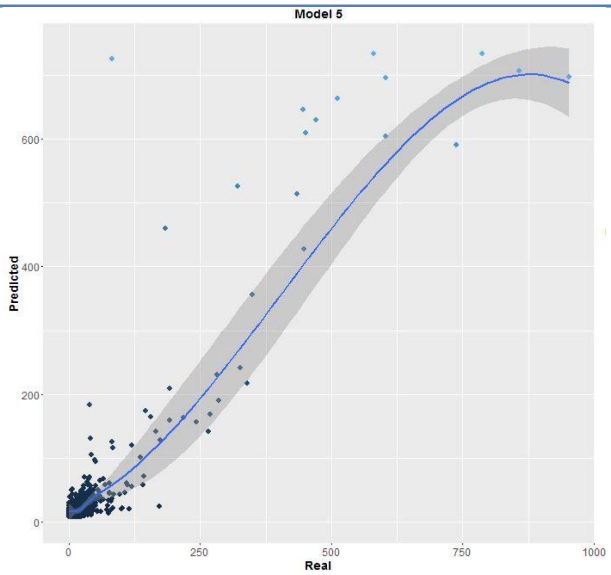
Thus, between them, MAE and RMSE provide a strong estimation of the variation of errors in a prediction model[23]. In terms of scale, both MAE and RMSE return a value based on the range of the dependent variable in the dataset. In order to view these values as percentages, for this analysis both measures have been normalised to a percentage of the dependent variable range.

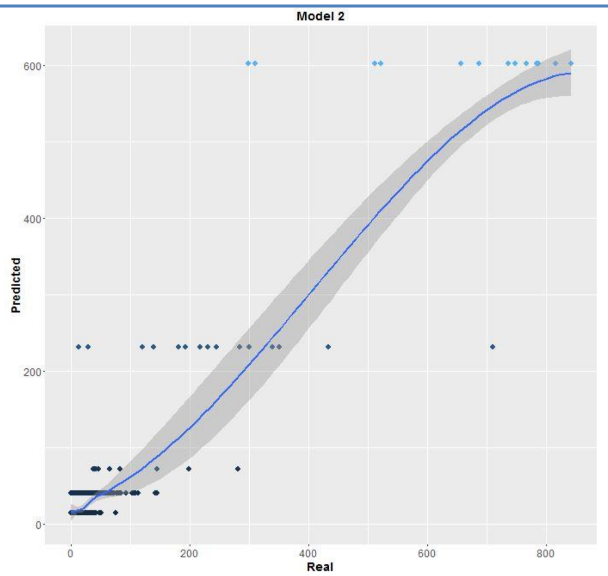
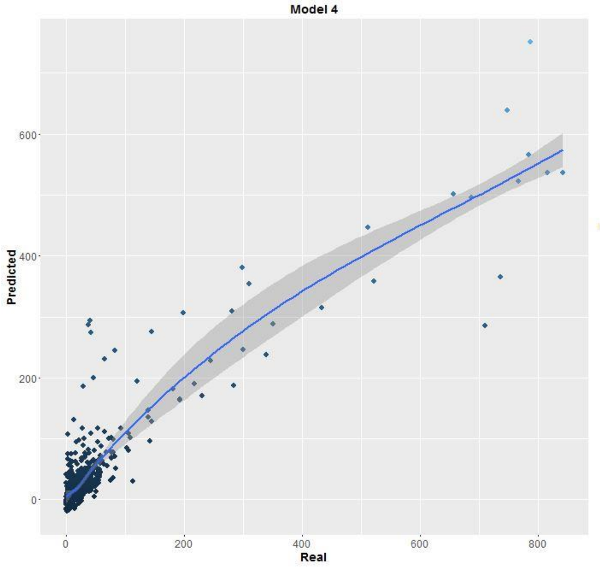
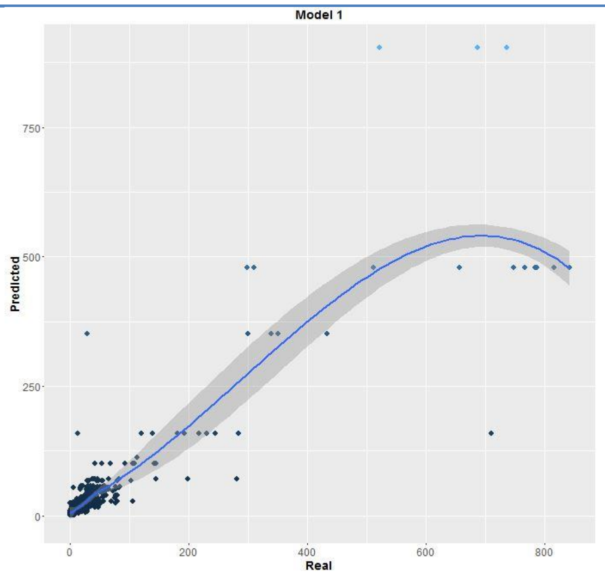
7.5. Model Comparison

The %MAE and %RMSE values for the eight models generated are presented over the next three pages, along with a plot for each showing the real vs predicted scores. The models are also ranked in order of %RMSE score.

Model #	Plot	%MAE	%RMSE
Model 3 **Random Forest**		1.2%	3.1%

[23] Nau, R. (2016). *What's the bottom line? How to compare models*. Retrieved from Statistical forecasting: notes on regression and time series analysis: http://people.duke.edu/~rnau/compare.htm

Model #	Plot	%MAE	%RMSE
Model 7 **Neural** **Network**		1.5%	3.9%
Model 6 **Neural** **Network**		1.5%	3.9%
Model 5 **Neural** **Network**		1.7%	4.0%

Model #	Plot	%MAE	%RMSE
Model 2 **Regression Tree**		2%	4.6%
Model 4 **Linear Regression**		2.1%	5%
Model 1 **Regression Tree**		1.7%	5.4%

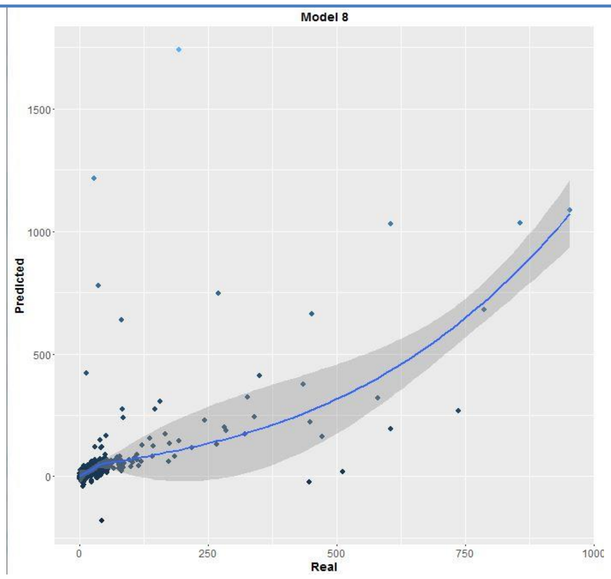
Model #	Plot	%MAE	%RMSE
Model 8 – **Neural Network**		2.5%	10.0%

Table 7.2 - Evaluation results

The model that performed the best in fitting the data and predicting accurate results was the ensemble Random Forest Model. The MAE value shows that the average difference between the real and predicted intake values was only 1.2%, and the RMSE of 3.1% demonstrates that the average squared error is also relatively small.

Random Forest Models have various advantages, including the capability to handle both categorical and continuous data; this makes them suitable for a dataset such as this with a mixture of data types. In addition, the easily interpretable output of variable ranking enables the analyst to quickly identify the most important features.

However, the biggest advantage of the Random Forest approach is that it does not require linearity in the data profile, and this may partly explain the success of Model 3. In comparison, the Linear Regression model - which does require an assumption of linearity - performed poorly and is ranked in 6th place. It has already been shown in Chapter 5 that not all features follow a linear profile, hence these results are to be expected. Linear Regression was included as a model type because it is a technique used very frequently in contemporary property price prediction models, as discussed extensively in Chapter 2; however based on the results of this experiment it is not to be recommended as an approach for modelling application intake.

In comparison to the Random Forest, the poorest performing model was the Neural Network model trained with three hidden layers and including the County and Month variables. The RMSE in this model is 10% - so over three times that of the best performing model – and the accuracy level is twice as bad at 2.5%. The large difference between the MAE and RMSE in this model is evidence of variation in the scale of errors. The fact that the RMSE has increased to such a large extent indicates the presence of some large errors, suggesting that certain predicted points are quite dramatically different to the actual observed values.

The poorest model may be a Neural Network model, but the other three models using this algorithm scored much higher, in particular Model 6 and Model 7. Neural Networks are more difficult to train and interpret than other model types, and identifying the parameters that increase or decrease the fit is difficult. Models 6 and 7 were fitted with different variables and different numbers of hidden layers, and yet the predictions outputted scored very similarly. Meanwhile, Models 7 and 8 have the same variables and only differ in the number of hidden layers, and yet the results are dramatically different. The potential predictive power of Neural Networks is undeniable, but it would take more time than is possible in this instance to gain a true understanding of their structure and capabilities.

Finally, the Regression Trees performed poorly over all, with evidence of large errors and poor overall accuracy.

7.6. Overall Evaluation

In evaluating the overall success of the modelling process, it is important to return to the original research question and ascertain the degree to which it has been answered:

> *"What are the factors that influence intake of applications for first registration to the Property Registration Authority, and can these features be used to build models to predict future intake?"*

7.6.1. Variable Importance

For each model type built, it has been possible to identify the most important features, and to rank them in order of importance. They are as follows:

Regression Trees (Model 1)	**Random Forest**	**Linear Regression**	**Neural Network (Model 5)**
1. Houses_Sold 2. Year 3. Month 4. County 5. Unemployment	1. Houses_Sold 2. Percent_Reg 3. Population 4. County 5. Year	1. Houses Sold 2. Population 3. GDP 4. Unemployment 5. CPI_Housing	1. Houses Sold 2. Population 3. Year 4. Percent_Reg 5. All_Price

Table 7.3 - Variable Importance

It is clear that every model type identified the total number of houses sold as the single most important factor in calculating intake.

Apart from Houses_Sold, there is a lot variation in the other variables labelled as important. Population features strongly, as does Year. Interestingly, the models that can handle categorical or ordinal variables were the ones that performed most strongly; for example the most successful model (the Random Forest) was able to take County and Year into consideration, which could explain the higher levels of predictive power.

Thus, based on the optimal model, it appears that the single largest influence on intake of applications to the PRA is the number of houses sold, followed by the percentage of a county already registered, the County (and its population), and finally the year.

7.6.2. Predictive Power

The second part of the research question asks if these factors can be harnessed to build prediction models. The modelling and evaluation conducted has shown that this is indeed possible. A range of models has been produced and one has been highlighted as most useful.

However, in terms of actually putting this into practice as a useful forecasting tool for the PRA, it will require further development before being applied as a reliable source of information.

At present, the model predicts the value of Count based on Houses_Sold, Percent_Reg, Population and so on. However, if trying to predict the Count of applications to be lodged in the future (for instance for Q4 of 2016) the model fails, since the data for one of the independent variables – the number of house sold – is not yet available for this time period. There may also be difficulty getting access to updated values for other features in a timely manner; for instance as of the end of June 2016 the CSO has yet to publish the GDP figures for Q1 of 2016.

It is not a total failure – the model has highlighted the key factors that influence intake, and has shown that these factors can be harnessed in such a way as to explain variation. However, it will be necessary to take a different approach to develop this project further. There are several options: for example, to use other published forecasts as input data – however locating sufficiently detailed datasets for this purpose may be problematic. Alternatively, it might be interesting to explore time series forecasting instead, and attempt to build univariate time series models.

7.7. Key Findings

The most successful model was the Random Forest Model 3, in terms of predicting Count and with the smallest values of %RMSE and %MAE. Random Forest models have various attributes that may make them more suitable for mixed datasets such as this one; they can handle categorical or continuous data, clearly rank the variables in order of importance and do not require an assumption of linearity.

Among the least successful model types in terms of prediction accuracy were one of the Neural Network models (trained with three hidden layers and including the County and Month variables) and the Regression Tree models. The larger RMSE values in these instances indicate the presence of some large errors, suggesting that certain predicted points are quite dramatically different to the actual observed values. Regression Trees and Random Forests are similar in construction in that both models use the tree structure; however Random Forests have the advantage of using bootstrap aggregation to only use a random sample of the training set at any time (Han et al., 2012) – this reduces overfitting and may explain the superior performance of the Random Forest over the Regression Trees in this instance.

The Linear Regression model performed poorly overall, and may have suffered from an absence of linearity in certain elements within the dataset. Another possibility is that the outliers identified in Chapter 6 have had an adverse impact, since multiple linear regression can be sensitive to outlier effects.

The number of levels within some of the variables in the dataset can contribute to this problem also; for example while the County variable in the modelling dataset only has 26 dimensions, every single observation in the All_Price variable represents a unique sale price, and therefore the variable comprises 1,872 dimensions. One of the problems associated with such 'high-dimensional' data is the 'Curse of Dimensionality', whereby in order to estimate a function with any degree of accuracy, several samples of each combination of values in a dataset is required. If there are a very high number of dimensions, then the quantity of training data required grows exponentially (Samet, 2006). In terms of outliers, the dimensionality curse introduces noise to the dataset, and therefore conventional methods of outlier detection cease to work as effectively (Han et al., 2012).

Finally, mixed returns were reported from the Neural Network models. Neural Networks are notoriously difficult to train and interpret; as Han et al., 2012 point out:

> *"A major disadvantage of neural networks lies in their knowledge representation."*

The human eye finds it difficult to interpret the combination of neurons connected by weighted links presented by a Neural Network model. It is therefore extremely challenging for the analyst to train a model in the standard way, by identifying the parameters that increase or decrease the fit. Therefore, tuning a Neural Network is a complex matter – Han et al. propose a process of using rule-extraction and sensitivity analysis to extract the knowledge from a neural network; however this is beyond the remit of a project such as this.

7.8. Conclusions

This chapter has covered the evaluation of the individual models, their successes and failures. Modelling is not an exact science, and at best is merely a simplified representation of the real world. The models constructed are not by any means perfect, but they have been trained in a structured way, adhering to methodology and with every effort made to avoid the entry of bias into the process. Of the models created, the Random Forest shows the most promise, and may form the basis for further development in the future.

8. CONCLUSIONS & FUTURE WORK

8.1. Introduction

This chapter summarises the findings and conclusions arrived at in the course of this study, which sets out to explore the possibilities for predicting intake of applications for first registration in the Property Registration Authority. It will present the aims and objectives of the project and assess how successful it has been in achieving these goals. Finally, ways in which the research conducted can contribute to the body of knowledge will be explored, and areas for future research will be discussed.

8.2. Problem Definition & Research Overview

The impetus for this project was rooted in a real business need. The PRA requires a reliable system for predicting future intake, to replace the current practice of basing the targets for the coming month on the number of applications lodged the previous month. Prior to this project, there has also been a lack of understanding of the factors that influence intake, and that go much of the way to explaining the peaks and troughs in intake levels that have been seen over recent years. Therefore, this dissertation has sought to identify the factors that influence intake of applications for first registration, and to ascertain if these features could be used to build models to predict future intake.

To answer the question, an exercise in data analytics has been designed and implemented, following the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology (Shearer, 2000). The main objective of the exercise was to test the hypothesis that intake of ‘first registration’ applications in the Property Registration Authority can be predicted to a set degree of accuracy through analysis of historical intake data and external factors.

Contemporary literature is reviewed in Chapter 2, on the subject of the Irish property market, the factors that influence the level of demand for registration, and modelling approaches applied to variable selection and predictive modelling. The nature of the business problem was elaborated in Chapter 3, which explains why the level of applications for registration is a challenge for the PRA.

Chapter 4 discusses the collation of data from several sources to form a dataset on which the experiment is to be based. The dataset is organised and prepared for use in the modelling process, and an extensive exploratory statistical analysis of the various elements is conducted in Chapter 5.

The modelling chapter discusses the construction of various predictive models and presents the results. As explained in Chapter 6, this process involves the application of several different approaches with the objective of identifying the most effective approach for this particular purpose. The evaluation and assessment of the models in Chapter 7 indicates that the Random Forest model is the most effective, and the combined outcome of all of the models indicates that the number of houses sold is the single most important factor in predicating volume of applications for registration.

8.3. Contributions to Body of Knowledge

This dissertation has made various contributions to the body of knowledge. Apart from the main objective of testing a specific hypothesis, the gathering and organisation of a body of data relevant to property registration in Ireland is in itself an important outcome of this study, and is of significant benefit to the PRA and to the general body of knowledge available on this subject. An extensive exploration of the literature relevant to the prediction of quantity of property sales focused on two main areas of commentary: housing markets and the factors that influence them; and the various approaches taken to modelling datasets in housing market or other similar trend research. Several significant insights have emerged:

8.3.1. Cyclical Market

The key characteristic of the Irish housing market, emerging from a synthesis of the conclusions of a range of commentators in the literature, is that it is cyclical, and that the cyclical variations in activity in the market are heavily influenced by regionality factors (including the gradual urbanisation of the population) and by economic fluctuation. The peaks and troughs of Dublin market during the 20th Century and early 2000s can be seen in the image below, the most dramatic of which occurred during the recent Celtic Tiger period.

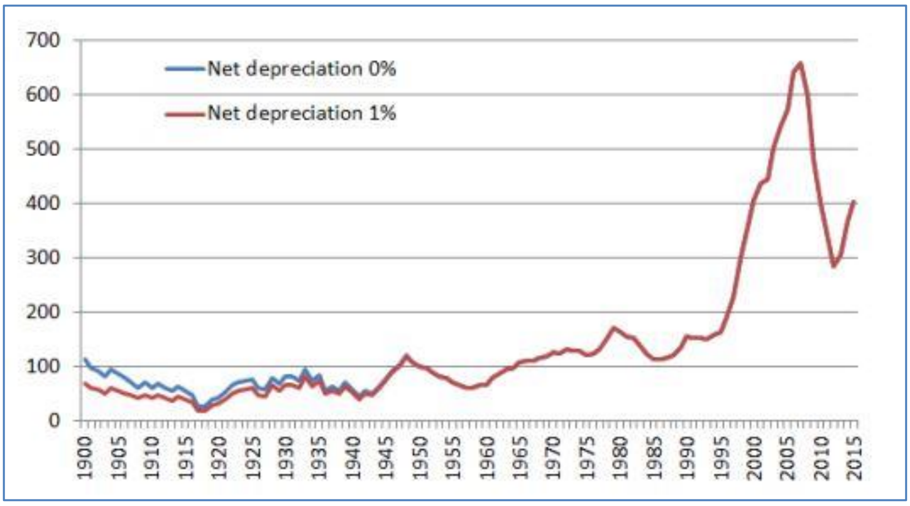


Figure 8.1 - Dublin House Price Index, 1900-2015 (Lyons, 2015)

8.3.2. Key Influence Factors

An extensive exploration of the literature on property market trends, examining an international range of markets, has identified three clusters of factors that influence property markets: the economic environment, location and seasonality. The results of this exploration have been synthesised to produce a diagram that summarises the influencing factors.

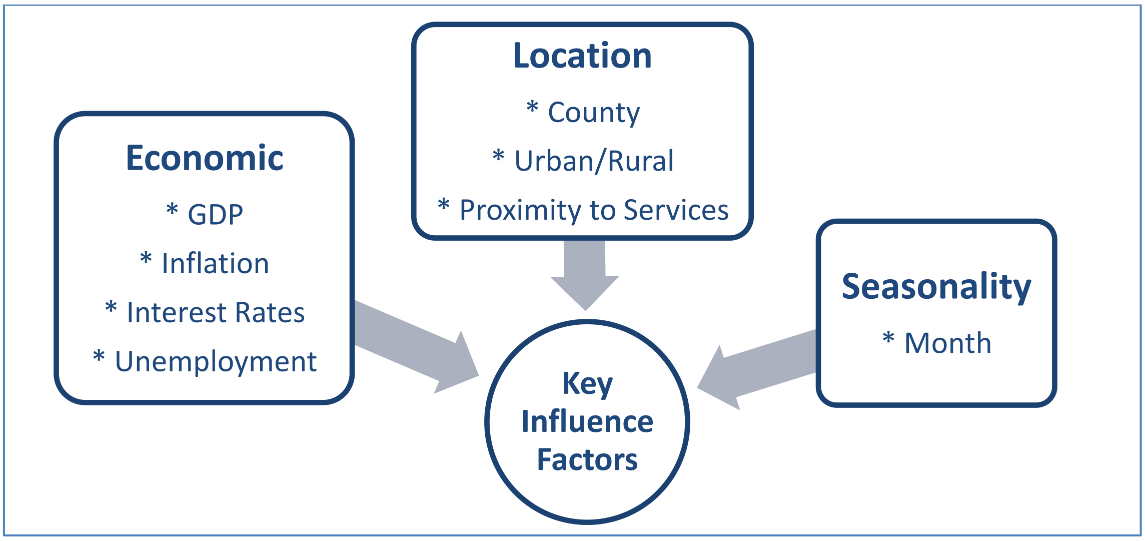


Figure 8.2 - Key Influence Factors

8.3.3. Typical Models

An exploration of the literature in relation to the modelling techniques used to analyse and process property market data identified a range of approaches. In the studies examined, linear regression emerges as the most commonly-deployed methodology, with less frequent usage of decision trees and neural network models. Examples of all of these approaches and their advantages/disadvantages in relation to the present study have been discussed in detail in Chapter 2.

8.3.4. The Need for a Prediction Instrument

The exploration of the property registration process in Ireland identifies serious challenges faced by the PRA: increasing levels of applications for first registration, which when combined with strained staff resources are resulting in slower turnaround times. The key finding is that an instrument enabling PRA management to predict future levels of intake would be a valuable resource. This section also identifies the data stored by the PRA that could be exploited in the design of prediction models.

8.3.5. *Conditioning the Dataset*

The first stage of conditioning the project dataset – statistical analysis and data-profiling – resulted in a range of visualisations that provide insight into the patterns and trends implicit in the dataset. The analysis examined data in relation to population, house sales and price, rates of unemployment, the consumer price index and finally application intake data. A key finding is that the percentage of land already registered in a given county is a significant factor that should be included in the final dataset for modelling.

8.3.6. *Best-Fit Model*

The final stage in this study was the testing of a variety of data modelling processes in order to identify possibilities for using available data to predict the level of applications for property registration. Four modelling approaches were tested – Regression Trees, Random Forests, Multivariate Linear Regression and Neural Networks. The key finding from this chapter is that the model that performed the best in fitting the data and in predicating accurate results was the Random Forest Model.

8.3.7. *Key Prediction Factors*

Looking at findings in relation to the original research question, all of the model types identified the total number of houses sold as the single most important factor influencing rate of application intake, followed by the percentage of the property in a county already registered. A further finding is that the investigation has shown that it is possible to harness the available data to build prediction models in this business area.

8.4. *Future Work & Research*

This project has provided insight into the property market and registration process in Ireland, and has highlighted the significant potential of predictive analytics as a tool for forecasting future intake. However, it is also clear that further work is required in order to develop a predictive process that could be made operational in the PRA business setting. There are several avenues that can be explored to further this research.

8.4.1. *International Comparison with Similar Jurisdictions*

It would be useful to contact jurisdictions with similar systems of registration - for example land registries in other countries - to examine their intake patterns when compared to economic and property market trends, and to investigate if these organisations have conducted any similar research into intake and data analysis.

A logical first venture along these lines would be to look at the activities of Registers of Scotland (RoS); this statutory body is actively working towards the objective of completing a new register of property in Scotland, and thus has parallels with the Irish situation in that both registers are in active development. However, any such initiative would depend on PRA policy lines and would require sanction through formal channels.

Other candidates that may be suitable for comparison would include the United Kingdom, Canada, Australia and New Zealand, since these countries use the same 'Common Law' legal environment as Ireland.

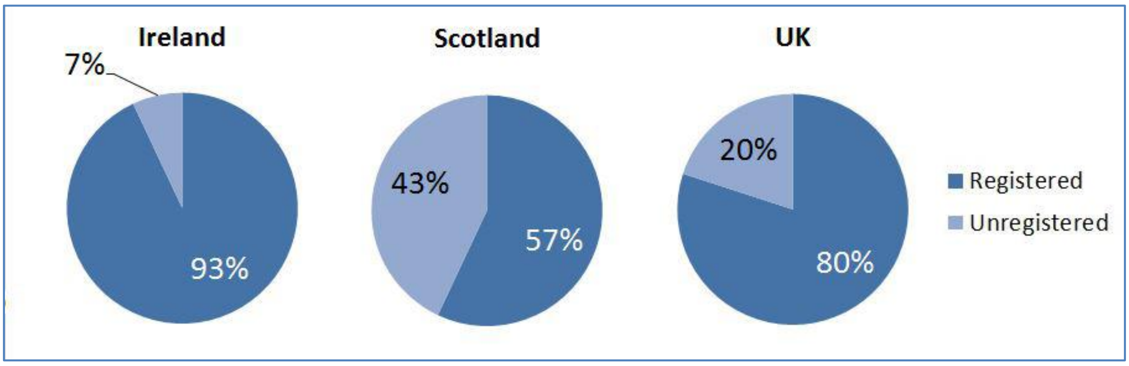


Figure 8.3 - Percentage of land registered, Ireland, Scotland & UK[24]

8.4.2. *Embellishing the Dataset*

It may be advantageous to explore ways to improve and embellish the dataset used – for example:

- The Consumer Price Index data was not found to be significantly influential; it would therefore be unnecessary to proceed with this variable in future approaches;
- A useful potential addition to the dataset could be interest rate values for mortgages and other property loans; this source was not available for this project due to a change in recording processes midway through the time series, but for future experiments interest rates could offer a revealing insight into consumer intentions;
- It would be necessary to address the issue of early and consistent availability of the data streams, as it is necessary that the analytics process provide predictions of level of

[24] Sources: Registers of Scotland. (2015). Land Mass Coverage Report; Land Registry UK. (2016, July). Registered or unregistered land, that is the question. Retrieved from Land Registry UK;

applications in time for management to make the necessary resource allocations. This would require the negotiation of significant inter-agency collaboration.

8.4.3. Outlier Manipulation

During the modelling phase of this experiment, the presence of a large number of outliers was identified, but the decision was taken to leave these observations in the dataset to preserve the integrity of the information contained within. An alternative approach would be to explore these instances more thoroughly, and experiment with removal of outliers identified as having a negative impact on the models created. It would also be useful to conduct further research into the literature surrounding outlier treatment, and to investigate alternatives to removal, such as replacement with the mean or performing a transformation on the whole variable.

8.4.4. Modelling Techniques

In this project, the choice of modelling approaches was guided by the research conducted. However, certain models such as the linear regression model performed poorly, and should not be developed further. Instead, alternative approaches could be explored; for example, time series forecasting has the potential to be advantageous, as a tool for revealing seasonal variation within a dataset. An example of time series forecasting model is the 'ARIMA' model, which combines time series analysis and modelling techniques in an autoregressive approach (Hyndman & Athanasopoulos, 2013).

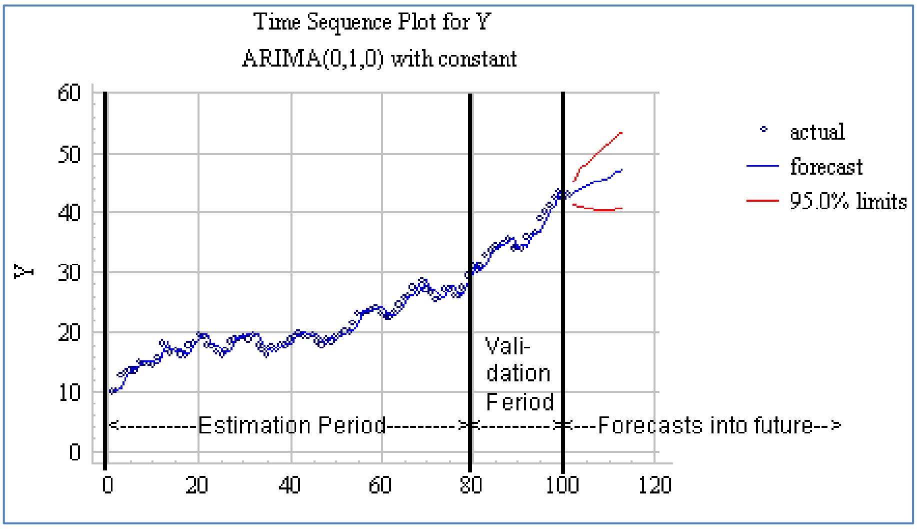


Figure 8.4 - Example of ARIMA Time Series Forecasting Model[25]

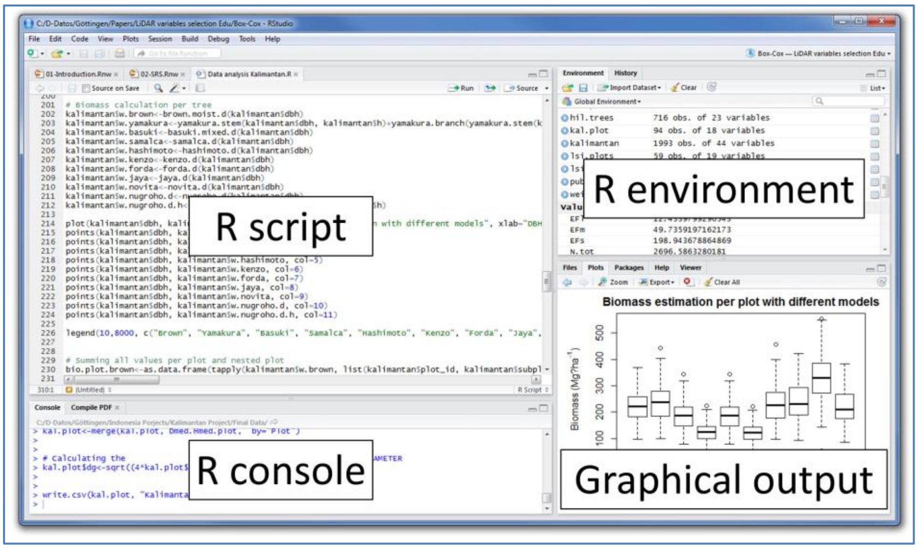
[25] Arsham, H. (2016, July). Time-Critical Decision Making. Retrieved from University of Baltimore: http://home.ubalt.edu/ntsbarsh/stat-data/forecast.htm

8.4.5. *Further Prediction using the Models Created*

In this instance, the models were used to predict the target variable Count. As an extension of the project, they could instead be trained to predict the various application types – Form 1, Form 2 and so on. It may be the case that some models are more successful than others at identifying the intake of a particular application type, and if so it would be interesting to investigate the reasons why.

8.4.6. *Alternative Software*

R was used throughout this process for modelling and evaluation. However, R has certain limitations; for example for the novice user it can be impenetrable to comprehend, and many hours are spent in attempts to overcome coding issues and get to grips with the range of packages required for different models. In comparison, SAS is relatively intuitive to operate and comes with comprehensive documentation and tutorials. Another option to explore would be Python, which has the advantage of simpler syntax and is easily integrated with web applications, if this is required. An interesting experiment would be to build three models in each of the software applications, and then to review and compare the experience of modelling and the quality of the results extracted.

[26]

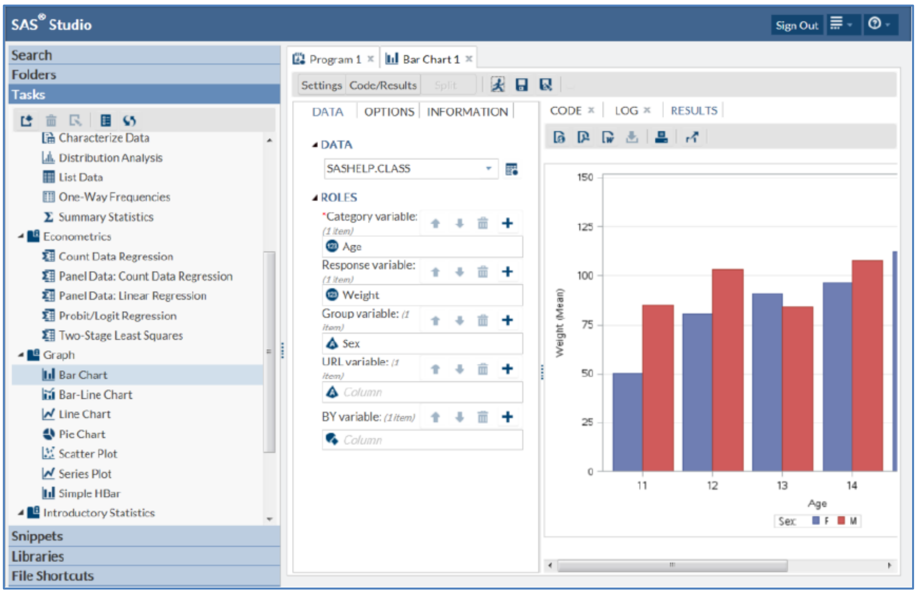
```
# Biomass calculation per tree
kalimantan$w.brown<-brown.moist.d(kalimantan$dbh)
kalimantan$w.yamakura<-yamakura.stem(kalimantan$dbh, kalimantan$h)+yamakura.branch(yamakura.stem(k
kalimantan$w.basuki<-basuki.mixed.d(kalimantan$dbh)
kalimantan$w.samalca<-samalca.d(kalimantan$dbh)
kalimantan$w.hashimoto<-hashimoto.d(kalimantan$dbh)
kalimantan$w.kenzo<-kenzo.d(kalimantan$dbh)
kalimantan$w.forda<-forda.d(kalimantan$dbh)
kalimantan$w.jaya<-jaya.d(kalimantan$dbh)
kalimantan$w.novita<-novita.d(kalimantan$dbh)
kalimantan$w.nugroho.d<-n
kalimantan$w.nugroho.d.h<                                    $h)

plot(kalimantan$dbh, kali                                    n with different models", xlab="DBH
points(kalimantan$dbh, ka
points(kalimantan$dbh, ka
points(kalimantan$dbh, ka
points(kalimantan$dbh, kalimantan$w.hashimoto, col=5)
points(kalimantan$dbh, kalimantan$w.kenzo, col=6)
points(kalimantan$dbh, kalimantan$w.forda, col=7)
points(kalimantan$dbh, kalimantan$w.jaya, col=8)
points(kalimantan$dbh, kalimantan$w.novita, col=9)
points(kalimantan$dbh, kalimantan$w.nugroho.d, col=10)
points(kalimantan$dbh, kalimantan$w.nugroho.d.h, col=11)

legend(10,8000, c("Brown", "Yamakura", "Basuki", "Samalca", "Hashimoto", "Kenzo", "Forda", "Jaya",


# Summing all values per plot and nested plot
bio.plot.brown<-as.data.frame(tapply(kalimantan$w.brown, list(kalimantan$plot_id, kalimantan$subpl
```

```
hil.trees       716 obs. of 23 variables
kal.plot        94 obs. of 18 variables
kalimantan      1993 obs. of 44 variables
lsi.plots       59 obs. of 19 variables
lsi
pub
wei
Valu
  EF
  EFm           49.7359197162173
  EFs           198.943678864869
  N.tot         2696.5863280181
```

```
C:/D-Datos/Göttingen/Indonesia Porjects/Kalimantan Project/Final Data/
> kal.plot<-merge(kal.plot, Dmed.Hmed.plot,  by="Plot")
>
>
> # Calculating the                               AMETER
> kal.plot$dg<-sqrt((4*kal.plot$
>
>
> write.csv(kal.plot, "Kalimanta
> |
```

Figure 8.5 - Example of R interface

[26] AWF-Wiki. (2016, July). Installation and Interface of R. Retrieved from Georg-August-Universität Göttingen, Germany



DATA

SASHELP.CLASS

ROLES

*Category variable: (1 item)
Age

Response variable: (1 item)
Weight

Group variable: (1 item)
Sex

URL variable: (1 item)
Column

BY variable: (1 item)
Column

Figure 8.6 - Example of SAS interface[27]

8.5. *Final Reflection*

This study set out to investigate the possibility of using data analysis techniques and strategies to address a problem identified in a real business context: to identify the factors that influence intake of applications for first registration to the Property Registration Authority; and to test whether these features be used to build models to predict future intake.

The overall conclusion of the study is that the null hypothesis has been rejected and that intake of 'first registration' applications in the Property Registration Authority can be predicted through analysis of historical intake data and external factors. However, it is acknowledged that further work will be required to develop a data gathering and analysis process that can be operationalised in the PRA context; apart from the need to refine the data modelling approach, a crucial issue is the need to gain timely and consistent access to the relevant data.

The predictive modelling techniques explored in this dissertation provide an exciting insight into the power and capability of data analytics in exploiting and harnessing the wealth of data already held both in the PRA and the wider public service. This data must be recognised as a valuable asset, and one which has the power to guide and influence planning and policy making.

[27] http://www.sas.com/

9. BIBLIOGRAPHY

1. Acciani, C., Fucilli, V., & Sardaro, R. (2011). Data mining in real estate appraisal : a model tree and multivariate adaptive regression spline approach. *Aestimum*, *58*(Giugno 2011), 27–45.
2. Attanasio, O. P., Blow, L., Hamilton, R., & Leicester, A. (2009). Booms and busts: Consumption, house prices and expectations. *Economica*, *76*(301), 20–50. http://doi.org/10.1111/j.1468-0335.2008.00708.x
3. Bańbura, M., Giannone, D., & Reichlin, L. (2008). *Large Bayesian VARs* (Working Paper Series No. 966). Frankfurt am Main, Germany. Retrieved from https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp966.pdf?bde67191add82abdf8703c77772d9e45
4. Bin, O. (2004). A prediction comparison of housing sales prices by parametric versus semi-parametric regressions. *Journal of Housing Economics*, *13*(1), 68–84. http://doi.org/10.1016/j.jhe.2004.01.001
5. Box, G. E. P. (1987). *Empirical Model-Building and Response Surfaces*. Wiley.
6. Casey, C. M. (2014). Averting Crisis? Commentary from the International Institutions on the Irish Property Sector in the Years Before the Crash. *The Economic and Social Review*, *45*(4), 537–557.
7. Cebula, R. J. (2009). The Hedonic Pricing Model Applied to the Housing Market of the City of Savannah and Its Savannah Historic Landmark District. *Review of Regional Studies*, *39*(1), 9–22. http://doi.org/http://economy.okstate.edu/rrs/
8. Central Statistics Office. (2012). *This is Ireland, Part 1*. Dublin, Ireland.
9. Chow, S.-M., Tang, N., Yuan, Y., Song, X., & Zhu, H. (2011). Bayesian estimation of semiparametric nonlinear dynamic factor analysis models using the Dirichlet process prior. *The British Journal of Mathematical and Statistical Psychology*, *64 (Pt 1)*, 69–106.
10. Clapp, J. M., Kim, H.-J., & Gelfand, A. E. (2002). Predicting Spatial Patterns of House Prices Using LPR and Bayesian Smoothing. *Real Estate Economics*, *30*(4), 505–532. http://doi.org/10.1111/1540-6229.00048
11. Deeter, K., Quinn, F., & Duffy, D. (2016). *Dublin house prices: A history of booms and busts from 1708-1949*. Dublin, Ireland.
12. Demary, M. (2010). The interplay between output, inflation, interest rates and house prices: international evidence. *Journal of Property Research*, *27*(1), 1–17. http://doi.org/10.1080/09599916.2010.499015
13. Donnelly, R. A. (2007). *The Complete Idiot's Guide to Statistics* (2nd ed.). New York: Alpha Books.
14. Dorgan, S. (2006). How Ireland became the Celtic tiger. *The Heritage Foundation*

Backgrounder, *1945*(1945), 5–7. Retrieved from http://www.heritage.org/research/reports/2006/06/how-ireland-became-the-celtic-tiger?ac=1

15. Duffy, D., & Fitzgerald, J. (2012). The Irish Housing Market. *Quarterly Economic Commentary*, (Summer), 65–78.
16. Engsted, T., & Pedersen, T. Q. (2014). Housing market volatility in the OECD area: Evidence from VAR based return decompositions. *Journal of Macroeconomics*, *42*, 91–103. http://doi.org/10.1016/j.jmacro.2014.07.005
17. Foreman, J. W. (2014). *Data Smart: Using Data Science to Transform Information into Insight*. Indianapolis: John Wiley & Sons.
18. Ge, X. J., & Lam, K. (2001). Building a House Prices Forecasting Model in Hong Kong. *The Australian Journal of Construction Economics and Building*, *2*(2), 57–70.
19. Goldberger, A. S. (1962). Best Linear Unbiased Prediction in the Generalized Linear Regression Model. *Journal of the American Statistical Association*, *57*(298).
20. Goodbody Stockbrokers. (2013). *Irish Property: A detailed analysis of the prospects for Irish Property*. Retrieved from http://static.rasset.ie/documents/business/goodbody-property-market.pdf
21. Goodman, A. C., & Thibodeau, T. G. (2003). Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics*, *12*(3), 181–201. http://doi.org/10.1016/S1051-1377(03)00031-7
22. Hamner, B. (2010). Predicting future traffic congestion from automated traffic recorder readings with an ensemble of random forests. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 1360–1362. http://doi.org/10.1109/ICDMW.2010.169
23. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.
24. Hawkins, D. M. (2004). The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.*, *44*(1), 1–12.
25. Hin Li, L., & Lin Ge, C. (2008). Inflation and housing market in Shanghai. *Property Management*, *26*(4), 273–288. http://doi.org/10.1108/02637470810894902
26. Ho, T. K. (1995). Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition, 1995* (pp. 278–282).
27. Hyndman, R. J., & Athanasopoulos, G. (2013). *Forecasting: principles and practice*. OTexts. Retrieved from https://www.otexts.org/fpp/9/2
28. Kestens, Y., Thériault, M., & Des Rosiers, F. (2004). The impact of surrounding land use and vegetation on single-family house prices. *Environment and Planning B: Planning and Design*, *31*(4), 539–567. http://doi.org/10.1068/b3023
29. Khalafallah, A. (2008). Neural Network Based Model for Predicting Housing Market

Performance, *13*(October), 325–328.

30. Leahy, B. (2014). *Predicting Professional Golfer Performance Using Proprietary PGA Tour "Shotlink" Data*. Dublin Institute of Technology.
31. Li, H. (2014). *The Effects of Demographics on the Real Estate Market in the United States and China*. Pace University. Retrieved from http://digitalcommons.pace.edu/honorscollege_theses/137
32. Li, Y., & Leatham, D. J. (2013). Forecasting Housing Prices: Dynamic Factor Model versus LBVAR Model. *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699. http://doi.org/10.1017/CBO9781107415324.004
33. Lyons, R. C. (2015). *Measuring house prices in the long run : Insights from Dublin , 1900-2015*. Retrieved from http://www.eh.net/eha/wp-content/uploads/2015/05/Lyons.pdf
34. Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.
35. McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133.
36. National Economic Social Council of Ireland. (2014). Homeownership and Rental : What Road is Ireland On ?, (140).
37. Ngai, L. R., & Tenreyro, S. (2009). *Hot and Cold Seasons in the Housing Market* (No. 922). London. Retrieved from http://cep.lse.ac.uk/pubs/download/dp0922.pdf
38. Quinlan, J. R. (2007). Induction of Decision Trees. *Expert Systems*, 81–106. http://doi.org/10.1023/A:1022643204877
39. Samet, H. (2006). High-Dimensional Data. In *Foundations of Multidimensional and Metric Data Structures*. MORGAN KAUFMANN PUBLISHERS.
40. Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications*, *36*(2), 2843–2852. http://doi.org/10.1016/j.eswa.2008.01.044
41. Shearer, C. (2000). The CRISP-DM Model:The New Blueprint for Data Mining, *5*(4), 13–22.
42. Sirr, L. (Ed.). (2014). *Renting in Ireland*. Dublin: Institute of Public Administration. Retrieved from http://arrow.dit.ie/cgi/viewcontent.cgi?article=1001&context=beschrebk
43. Snyder, T., & Ekmekjian, E. (2013). What are the impacts of the home buyer's tax credit on housing and the economy? *Research in Business and Economics Journal*, *8*, 1–10.
44. Stock, J. H., & Watson, M. W. (2002). Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business & Economic Statistics*, *20*(2), 147–162. http://doi.org/10.1198/073500102317351921
45. Stock, J. H., & Watson, M. W. (2010). Dynamic Factor Models. *Oxford Handbook of Economic Forecasting*, (July), 1–43. Retrieved from

http://link.springer.com/article/10.1007/s10182-006-0219-z

46. The Property Registration Authority. (2005). *Annual Report 2004*. Dublin, Ireland. Retrieved from http://www.prai.ie/publications/
47. The Property Registration Authority. (2009). *Annual Report 2008*. Dublin, Ireland.
48. The Property Registration Authority. (2015). *Annual Report 2014*. Retrieved from http://www.prai.ie/publications
49. The Property Registration Authority of Ireland. (2014). *Property Registration Authority Annual Report 2014*. Retrieved from http://www.prai.ie/publications
50. The Property Registration Authority of Ireland. (2016). *Strategic Plan 2016 – 2018*. Retrieved from http://www.prai.ie/publications
51. Valadez, R. M. (2011). The housing bubble and the GDP: a correlation perspective. *Journal of Case Research in Business and Economics*, *3*, 1–18.
52. Welling, M. (2011). *A First Encounter with Machine Learning*. Irvine, California: University of California Irvine 2011.
53. Whelan, K. (2013). *Ireland's economic crisis: The good, the bad and the ugly* (Working Paper Series No. 13/06). Dublin. Retrieved from http://hdl.handle.net/10419/96251