Geospatial Modelling of Property Prices in Limerick City and County

Brian O'Donovan
Department of Mathematics and Statistics
University of Limerick

Supervisor
Dr James Sweeney

Final Year Project
BSc in Mathematics and Statistics
March 8, 2025

Contents

Table of contents i

List of figures iii

List of tables v

Acknowledgements vi

Abstract vii

1 Introduction 1

2 Literature Review 4
- 2.1 Group Literature Review . . . 4
- 2.2 Individual Literature Review . . . 11
  - 2.2.1 The Housing Market in Ireland and in Limerick . . . 11
  - 2.2.2 Geospatial Data & Spatial Effects on Housing Markets . . . 12
  - 2.2.3 Modelling Methods . . . 14
  - 2.2.4 Research Questions . . . 20

3 Data Preparation 21
- 3.1 Limerick Property Dataset . . . 21
- 3.2 Locational Attributes . . . 23
- 3.3 Structural Attributes . . . 27
- 3.4 Complete Dataset . . . 31

4 Exploratory Analysis 32
- 4.1 Limerick Property Dataset . . . 32
- 4.2 Locational Attributes . . . 34
- 4.3 Property Attributes . . . 38
- 4.4 Outliers . . . 42

5 Modelling Methods 43
- 5.1 Multiple Linear Regression . . . 43

5.2 Geographically Weighted Regression . . . 45
5.2.1 Model Diagnostics and Fit . . . 50
5.3 Model Selection and Testing . . . 52
5.3.1 Variable Selection Method . . . 52
5.3.2 Model Comparison . . . 52
5.3.3 Model Accuracy . . . 53
5.3.4 Autocorrelation Test . . . 53
5.4 Data Preparation . . . 54
5.4.1 Dummy Variable Encoding . . . 54
5.4.2 Transformation of the Dependent Variable . . . 56
5.4.3 Out-of-Sample Testing and Validation . . . 57

6 Application to the Limerick Property Dataset 59
6.1 Polynomial Regression . . . 59
6.2 Multiple Linear Regression . . . 61
6.2.1 Basic Linear Regression Model . . . 62
6.2.2 Location Dummy Variables . . . 63
6.2.3 Distance Measure . . . 65
6.2.4 Absolute Location . . . 66
6.2.5 Results . . . 67
6.2.6 Model Interpretation . . . 70
6.3 Geographically Weighted Regression . . . 75
6.3.1 Model Selection . . . 75
6.3.2 Model Calibration . . . 76
6.3.3 Model Diagnostics . . . 77
6.3.4 Results and Model Interpretation . . . 79

7 Results 85

8 Discussion, Conclusions and Further Work 89
8.1 Discussion and Conclusions . . . 89
8.2 Further work . . . 92

Appendices 94
A Exploratory Data Analysis . . . 94
B Data Preparation . . . 96
C Results . . . 97

Bibliography 99

List of Figures

3.1 Missing vs present values of variables of interest. 21
3.2 Summary of the creation of new (green) variables from original (blue) variables. 22
3.3 Map of properties in their respective municipalities with the Limerick city centre point. . 24
3.4 Map of LCS urban areas with notable address locations. 25
3.5 Distribution of price: A - before data cleaning, and B - after data cleaning. 31

4.1 Bar plot of properties across regions, years and property categories. 33
4.2 Map of median prices by municipality with observation locations in red. 34
4.3 Box plot of price across the municipalities and urban areas of Limerick. 35
4.4 Map of property price (incl. outliers) in: A - County Limerick, and B - Limerick city and suburbs. 36
4.5 A - Scatter plot of CBD distance and price by municipality. B - Box plot of price by CBD distance range. 37
4.6 Box plot of price across regions: A - for property categories, and B - for house types. . . . 38
4.7 Map of the most popular property types in Limerick. 40
4.8 Scatter plot of the number of: A - bedrooms, and B - bathrooms, across regions. 41
4.9 Map of the proportion of properties: A - with garages, and B - on cul-de-sacs, across LCS areas. 42

6.1 Polynomial degree selection for 'Beds' and 'Baths'. 60
6.2 Map of the basic OLS model residual values in: A - County Limerick, and B - Limerick city and suburbs. 63
6.3 Map of residual values from the OLS-u model in: A - County Limerick, and B - Limerick city and suburbs. 64
6.4 OLS-cbdr-u model residuals: A - and fitted values, B - histogram plot. 70
6.5 Map of residual values from the OLS-cbdr-u in: A - County Limerick, and B - Limerick city and suburbs. 71
6.6 Relative scalings of 'Property Type', 'Property Condition' and 'CBD Distance Range' coefficients from the OLS-cbdr-u model. 72
6.7 Model selection AICc results for the GWR-3 model. 75
6.8 GWR-3 model residuals: A - and fitted values, B - histogram plot. 79
6.9 Selection of GWR modelling maps. 82
6.10 Selection of GWR modelling maps continued. 83

7.1 Predicted vs actual price values for the OLS-cbdr-u and GWR-3 models. 86
7.2 Predicted price surface in County Limerick using the GWR-3 model. 88

8.1 Map of the Ennis Road in LCS-1, with Ennis Road properties coloured by the correctness of address specification. 93
B.1 Histogram of: A - property price, and B - log($Price$). 96
B.2 Scatter plot of fitted values and residuals for the OLS models of price and log($Price$). . . 96
C.1 Scatter plot of fitted values and residuals for the OLS-cbdr-u and GWR-3 models. 97

List of Tables

3.1 Initial variables and descriptions. 22
3.2 Summary statistics of LCS properties by urban areas. 25
3.3 Property types of houses before data cleaning. 28
3.4 Property categories and types after data cleaning. 29
3.5 Cleaned data by year of sale. 31

4.1 Complete data with variables and descriptions ordered by category. 32
4.2 Summary statistics of all properties in municipalities and urban areas of LCS. 34
4.3 Summary statistics of properties across property types. 39

5.1 Selection of (adaptive) kernel functions used for the GWR model. 48

6.1 OLS model names and descriptions. 61
6.2 10-fold cross-validation results of the OLS models. 67
6.3 Model calibration results for the kernel functions of the GWR-1, GWR-2 and GWR-3 models. 76
6.4 The number of condition numbers over 30 and parameter VIFs over 10 for various bandwidth values for the GWR-2 model with a bi-square kernel function. 77
6.5 Coefficients of the GWR-3 model and the corresponding global model with p values for the global model and a Monte Carlo test of spatial heterogeneity. 79
6.6 10-fold cross-validation results for the GWR-3 model and the corresponding global model. 80

7.1 10-fold cross-validation results of the best OLS and GWR models. 85

A.1 Summary statistics of Limerick houses by property type. 94
A.2 Summary statistics for property attributes and features across municipalities and urban areas. 95
C.1 OLS-cbdr-u model coefficients and confidence intervals in standard and exponential format. 97

Acknowledgements

I would like to express my gratitude and appreciation for my supervisor, Dr James Sweeney, whose guidance and support have been invaluable throughout this project. The many stimulating questions raised during our meetings were vital in inspiring me to think outside the box and form a comprehensive analysis. I would also like to thank the team of lecturers who have always been approachable during the weekly lab sessions, their guidance on various topics has been greatly appreciated.

Abstract

County Limerick contains the third largest city in Ireland, yet there is little empirical research on the modelling of property prices in this region. Within a city marked by evident social segregation and geographically focused poverty, the geospatial modelling of residential prices can be utilised to explain the impacts of location on price. In this final year project, a comprehensive dataset of 1,497 Limerick property transactions from January 2017 to December 2018 is processed and used to develop a suite of models for property price valuation. The challenges associated with a lack of postcode regions in Limerick are mitigated through the development of urban areas using geographic information systems. Natural language processing of text descriptions is utilised to overcome the obstacles relating to poor data quality from property listing websites by learning additional features from the descriptive text.

The traditional non-spatial hedonic regression model is used to quantify the effects of property types, conditions and features on property price. Through the inclusion of spatial variates of price, such as submarket areas and a distance measure, the monetary impacts of living in certain regions in the county are obtained. A parsimonious spatially smooth geographically weighted regression model is used to explore spatially varying relationships between property features and price. The effects of spatial autocorrelation and spatial heterogeneity are considered throughout the analysis and a potential solution is outlined using a combination of the hedonic and geographically weighted regression models as an exploratory tool.

Chapter 1

Introduction

Statistical approaches to property appraisal are used to predict the price of a property based on its unique structural, neighbourhood, and locational characteristics. Property valuations are widely used to inform government policy, calculate mortgage valuations, and conduct financial reporting. Property price indices are a requirement in all European states and inform the population of changes in the average sales price of properties (Eurostat 2013). Corcoran & Liu (2014) outline a requirement for automatically generated housing estimates as an efficient approach to reducing price instability in the housing market. The empirical research on property price modelling in Ireland is generally performed at a national level or is focused explicitly on the cities. There is little research on the modelling of property prices in County Limerick, which has a population of 209,536. County Limerick contains the third most populated urban centre in Ireland, and the demand for housing in Limerick city is likely driven by the number of young families and students. There are over 30,000 students in Limerick city, and over 50% of the population under 35 years of age (Limerick County Council (LCC) 2022*a*). Limerick city and surrounding areas demonstrate urban sprawl and social segregation, where there is an increased concentration of poverty in the city centre and affluence within the suburbs (McCafferty 2011).

The economic effects of rising property prices and instability are apparent, and changes in household wealth and household consumption are driven by these price movements (Campbell & Cocco 2007). As a consequence of increasing house prices and deposit requirements, renters and non-homeowners face increased barriers to entry to the housing market. House prices have a strong influence on the social and economic characteristics of the population and the changes in house prices create a ripple effect, where price diffusion occurs from cities to regional areas (Meen 1999). There is a growing public interest in efficient approaches to automated property valuation for use by government entities to reduce price instability and its effects.

The traditional modelling approach for property prices is the hedonic multiple linear regression model proposed by Rosen (1974). This additive approach models the value of a property as the intrinsic value of its attributes and remains the recommended approach by Eurostat (2013) for government departments. A unitary, global, equation is fitted to the entire market with many pitfalls, including a lack of spatial complexity, a risk of omitted variable bias, and the inadequacy of capturing complex, spatially-varying relationships between the independent variables and price. The underlying issues common to spatial data are spatial autocorrelation, such that neighbouring data points share similar values of the response variable, and spatial heterogeneity, whereby the relationships between attributes vary over space leading to structural instability under the assumption of stationary relationships.

Housing markets are segmented based on shared characteristics and price structures between houses. Submarkets are delineated in the modelling of house prices to capture the effects of segmentation through localised regression models. Local regression models are created for each homogeneous submarket, such that the relationships between the independent variables and dependent variable can be explained within submarkets (Basu & Thibodeau 1998). The geographically weighted regression model is a local estimation approach where a model is calibrated at each observation. This approach is used to explain spatial non-stationary relationships between independent variables and the dependent variable, where the values of model parameters can vary across space forming "valleys and hills" (Fotheringham et al. 1998).

In this final year project, the hedonic regression and geographically weighted regression models are outlined and applied to the Limerick property dataset. The Limerick property dataset is prepared for modelling by refining the initial dataset of 12,435 property transactions to a more informative and complete dataset of 1,497 entries in the period between January 2017 and December 2018. The data exploration is conducted to identify the key predictor variables of property price.

The initial dataset consists of residential, land, and commercial sales transactions from the period of January 2011 to December 2018. The data is relatively sparse, with key structural and location attributes missing for many properties. These issues are persistent in Irish property price studies (O'Hanlon 2011, Maguire et al. 2016). Robust cleaning methods are applied to the data to extract value before removing nonessential entries. Property characteristics are explored across the four municipalities of Limerick county and the nine urban areas within Limerick city and suburbs. The results are compared and variables of interest are identified for the modelling of property price. This analysis aims to develop an interpretable model which can be used to understand the specific contributions of property characteristics and identify those with the greatest impact on property price. Locational attributes are included in the multiple linear regression model to quantify the impact of location on property price, and spatially varying

relationships are explored using a geographically weighted regression model.

The remainder of this paper is organised as follows. First, a high-level group literature review on house prices is presented in Section 2.1. The individual literature review in Section 2.2 focuses on the housing market in Ireland and in Limerick, Irish house price modelling, and the methods used for geospatial approaches to property valuation. In Chapter 3, the Limerick property dataset is introduced and the cleaning methods are outlined. Chapter 4 contains a comprehensive exploration of the cleaned dataset. The modelling methods are outlined and applied to the Limerick property dataset in Chapters 5 and 6, respectively. Overall results are presented in Chapter 7, followed by the conclusions and a framework for further work in Chapter 8. The Appendices A, B and C contain supplementary tables, graphs and model results.

Chapter 2

Literature Review

2.1 Group Literature Review

In this group literature review, we explore the recent fluctuations in house prices, the socioeconomic factors that influence house prices, and the socioeconomic effects of house price changes. We identify the importance of house price modelling and explore the most common property valuation models.

The cyclic behaviour of housing markets and how it impacts the financial sector has come to light in recent years following the recent global financial crisis. Factors such as changing inflation rates and changes in the credit supply regarding mortgages drive major fluctuations in housing markets (Cronin & McQuinn 2021). As the purchase of a property is usually financed by mortgage loans, researchers have suggested a study on house prices should include the financial sector which supplies loans to finance housing (Hillebrand & Kikuchi 2015). Some economists have suggested that the housing market cycles are related to modern economies and key to broader economic fluctuations (Lyons 2015). Researching the factors that drive housing prices is essential to understanding how the housing markets are related to these economic fluctuations. There has been research into how housing prices have a boom-bust behaviour (Case & Shiller 1994, Cesa-Bianchi 2013). The model created by Hillebrand & Kikuchi (2015) suggests that the boom of house prices occurs when the interest rate is positive and there is more flexibility in acquiring mortgage loans. They also suggest that the boom begins when the interest rate is higher than the growth rate, resulting in a positive correlation between loan volume and house prices during a boom-bust cycle. An example of such behaviour arose in Ireland during The Great Financial Crisis in the years of 2007-2012 and resulted in a significant impact on the Irish housing market following the end of the *"Celtic Tiger"* in 2007. This behaviour caused house prices to fall by over 50% (Cronin & McQuinn 2021). This study suggests that the in-

troduction of new mortgage plans in the mortgage credit market and the shift in the general economy resulted in the sharp boom of house prices with an increase of 334% between 1995 and 2007. Following this boom, there was a substantial bust of house prices, with prices falling by more than half, due to the collapse in the provisions in the credit supply market. During this crisis, governments decided to cut national budgets which caused major impacts on the housing market. This saw reduced housing transactions, decreased mortgage activity and a reduction in housing investments (Priemus & Whitehead 2014). (AM)

In a different study published in 2008, Brunnermeier & Julliard (2008) suggest that the sharp fluctuations in house prices are due to money illusion. This is where consumers cannot see whether the changes in nominal value are due to the changes in real fundamentals such as construction and land costs, property taxes, *etc.*, or merely from changes in inflation. Price sizing effects are evident in the housing market according to Case & Shiller (1994) and Costello (2000), and can explain some of the volatility in the housing markets during the boom-bust cycles. In the study by Case & Shiller (1994) in the United States, affordable house prices rose more rapidly during the boom period as they were initially priced lower. However, these houses experienced the sharpest decline due to their rapid rise during the boom and poor economic conditions in lower-income areas. Essentially, the price of more affordable houses tends to be more volatile during boom-bust cycles. Costello (2000) completes a study in Australia and finds that the more affordable houses are, the more competition there is surrounding them, increasing the houses' liquidity, which in turn makes these houses more volatile during these boom-bust cycles. Another study on Australian housing prices finds that inflation appears to be the determinant of the volatility of housing prices at a national level and that volatility has more of a response to the decrease of house prices rather than an incline (Lee 2009). (AM)

There is a common characteristic across housing markets worldwide which is the increase in price inflation due to the global pandemic of COVID-19 (Egan & McQuinn 2023). The Irish housing market is expected to be one of the housing markets in Europe that experienced a major fluctuation in recent years due to this pandemic. Governments worldwide took measures in order to deter the rapid rise of COVID-19 cases by closing all retail and commercial enterprises, excluding essential services. These closures resulted in a demand shock to the economy as there was a significant fall in aggregate consumption (Allen-Coghlan & McQuinn 2021). Allen-Coghlan & McQuinn (2021) present multiple scenario analysis that indicates Irish house prices are due to fall over the next 18 months because of the decline in household disposable income, a decline in mortgage activity and constraints on supply due to the closures of enterprises. This is a similar outlook to Priemus & Whitehead (2014) and Cronin & McQuinn (2021), who had the same conclusions when it came to the reasons for the decline in house prices following the financial crisis in Ireland ending in 2012. (AM)

House prices impact many socioeconomic factors such as household consumption, unemployment, and economic growth. Economic volatility is strongly influenced by the instability of house prices, and there is an asymmetric relationship between property investment returns and economic growth (Asadov et al. 2023). Using property transaction data from 17 OECD countries over 49 years, Asadov et al. (2023) conclude that the economic effect of a negative return on house prices is twice as strong as the effect of a positive return. Consumption at a regional level is influenced by regional house prices, and predictable changes in house prices trigger responses in household consumption (Campbell & Cocco 2007). In an empirical study of 379 municipalities in the U.S.A, Miller et al. (2011) conclude that house price changes influence gross municipal product, where predictable changes are three times stronger than unpredictable changes. A unidirectional causality between house prices and unemployment levels has been identified in 4 EU countries, whereby stabilising house prices contributes to economic stability and stable unemployment levels (Irandoust 2019). (BOD)

Young people have increased struggles in the housing market. Studies in Europe (Lennartz et al. 2016), the UK (Dunne 2012), and the US (Clapham et al. 2010), identify an increase in the number of co-residing adult children. This has given rise to concepts such as the "rental generation" and "failure to launch" (McKee 2012, Mykyta 2012). There is little price elasticity, or responsiveness to price changes, among young renters below 40 years of age, who are most vulnerable to house price increases (Campbell & Cocco 2007). The effect of house prices on ownership has the potential to create a gap in wealth between generations, as wealth accumulation and wealth inequality are strongly influenced by housing (W. Fuller et al. 2020). The increased house prices result in increased deposit requirements, making it more difficult to enter the "property ladder" (Simon & Stone 2017). Individuals from low socioeconomic backgrounds risk homelessness due to this barrier to the housing market (Hoolachan et al. 2017). Increasing house prices impact family life, with current house prices having a stronger influence on birth rates than current unemployment rates (Dettling & Kearney 2014). House prices have the potential to influence inter-urban migration. In a study of 291 micro-residential areas in the USA, housing cost was an influential factor in the choice of location for college-educated males (Plantinga et al. 2013). The importance of house price analysis is emphasised by Asadov et al. (2023), who suggest government policy, influenced by house price analysis and supervision, has the potential to control house price instability and alleviate the economic effects. (BOD)

Housing markets display the characteristics of traditional markets with producers, buyers and sellers (Gallet 2004). The housing market may be divided or segmented into submarkets with different price structures, based on the supply and demand constraints within each submarket (Straszheim 1974). Properties are good substitutes for each other within submarkets and

differ across submarkets (Bourassa et al. 1999). Schnare & Struyk (1976) proposed that housing markets display market segmentation due to a shared inelastic demand for structural and neighbourhood attributes amongst houses. Goodman & Thibodeau (1998) criticise the global approach of traditional house models and propose the examination of submarkets in house price modelling. There is a lack of consensus among academics on the definition of submarkets, leading to difficulty in delineating such areas (Islam & Asami 2009, Watkins 2001). The definition of submarkets based on an a priori understanding of market segmentation has improved accuracy in empirical studies (Bourassa et al. 1999, Adair et al. 1996). Straszheim (1974) proposes that submarkets may be defined by structural type, structural characteristics, or neighbourhood characteristics, and traditionally, authors utilise geographic borders (Goodman & Thibodeau 2003, Basu & Thibodeau 1998, Dubin 1988). Improved methods of submarket definition include cluster analysis (Bourassa et al. 2003, Tu et al. 2007). (BOD)

The availability of geocoded property data has presented many opportunities to model the relationships between and within geographically defined submarkets (Fik et al. 2003). Due to the immobility of houses, location has a significant effect on house price (Kiel & Zabel 2008). The geospatial modelling of house prices has been used extensively in empirical studies, through the modelling of residuals (Dubin 1988), and absolute location (Fik et al. 2003). The impact of neighbouring properties on property price is captured through spatial modelling (Can 1990, Pace & Gilley 1997). The prevalence of market segmentation gives rise to the use of localised regression models, where property coordinates can be used to generate models at every possible location in the data (Farber & Yeates 2006). Oust et al. (2020) and Bourassa et al. (2010) compare a range of models on out-of-sample data, and identify the advantages of using a geospatial model with defined submarkets. (BOD)

Land is a finite commodity, so location plays a big effect on the price of housing. Highly developed areas will have a higher demand than rural areas and therefore incur a higher cost. The price of houses in big cities is typically significantly higher than the price of houses in rural areas. This is down to the role that proximity to employment centres plays, as a lot of jobs are in the city. Other locational factors that play big roles in the pricing of houses are the proximity to green areas, water, and open spaces. A study analysing these effects in the Netherlands found that "attractive landscapes were shown to attract a premium over less attractive landscape types" (Luttik 2000). This article considers a case study of a medium-sized town called Apeldoorn in the Netherlands. The study focused on the effect that a park located in the middle of the district had on house prices. The effect of proximity to the park and a view of the park resulted in price increases while the view of a multi-storey building resulted in a price decrease. The effects were additive and the price difference accumulated to 21% if the house was within 400 metres of the park, had a view of the park, and did not have a view of

a multi-storey building. The effect of urban green space on house prices was also observed by Engstrom & Gren (2017), with their results showing that urban parks have a positive effect on property value and the effect tends to increase when reducing the distance to the park. Other studies that observed an effect of relative location on the price of housing include Heyman & Sommervoll (2019) and Gelfand et al. (2004) in which they both find that spatial effects had high explanatory power in the price of housing. (AC)

To what extent would transport impact the cost of housing? The majority of new home buyers would consider their access to transportation as a way to save costs for things like insurance, petrol, and finance, among other things. In a study published in 2001, Bowes & Ihlanfeldt (2001) examined the relationship between train stations and neighbourhood housing values in Atlanta, Georgia. The positive effects included lowering transportation costs and drawing in retail business. They found that residential growth typically goes hand in hand with the creation of a new transport hub. The relevance of accessibility on land value has been emphasised by many theorists and that distance and accessibility have been widely accepted as a significant factor in the modelling of housing. Hess & Almeida (2007) conclude that there is a positive linear relationship between the prices and distance from railway stations in Buffalo New York, with the help from Weinstein (2002) also reached the same conclusion in another state, Dallas, Texas. Debrezion & Rietveld (2007) found that Dutch houses that are in proximity to railway stations are about 23 percent more expensive than those located 15km away, while on the other hand proximity to the railway stations is a negative factor due to noise pollution. In a study conducted in Shanghai, China, a city which has seen a huge rapid transit expansion in recent years, Haixiao & Ming (2008) found that a house's price rises by 152 yuan (24 USD) per m^2 for every 100 metres that it is closer to a metro station. Hee et al. (2002) discovered that, just before the new transport line was to launch, prices in Seoul were significantly impacted by a property's proximity to any of the metro stations. Overall the studies found that certain transport stations such as railway stations and metros that get developed in residential areas or vice versa have an impact on property value given a proximity to each other, more studies had concluded, such as of Cervero & Landis (1993) and John & G.Stacy (2020). (NS)

When talking about transport, we don't limit ourselves to public transport; the effects of new motorways and roads can also determine a buyer's choice of whether to buy their new property or not. Research by Khanani et al. (2020) looked at how road infrastructure projects affected urban communities in Accra, Ghana, and Kisumu, Kenya. Their research shows that both cities' residential development improved as a result of these projects. The increased value of land and higher housing rentals were largely due to improved road infrastructure, which attracted real estate developers to urban settlements along those routes. Another study done by Boarnet & Chalermpong (2010) looked at the development of new highways in a city in Orange County, United States. With this, they decided to use a hedonic model which has other variables in-

volved which they also controlled for. They looked at a certain area that had built a new highway tolled system. They looked at data from before and after the development of the highway, and conclude that the accessibility for the road had increased in magnitude up until the first portion of the highway was built. Looking at another study done by Levokich et al. (2010) they looked at estimating the effects of the development of two highways in the Netherlands, an important variable that Levokich et al. (2010) had used was they accounted for noise levels in their model. A study similar to this Theebe (2004) had found that noise pollution affects housing prices in the Amsterdam area only at levels above 55 dB. With this, they found that there was a 10 percent–18 percent increase in houses that were located just along the new highway. They also found that the benefit of the highway had resulted in approximately 1,277 million. (NS)

Timely and accurate valuation of properties is of huge importance in the real estate market. Valuations are regularly requested for purposes such as sales and purchase of interest, letting or leasing and more. There are many different methods used for obtaining these valuations. While manual valuation methods have dominated in previous years, it has been said that these methods have flaws, and can provide artificially low valuations in hot markets and artificially high valuations in cold markets (Cannon & Cole 2011). A study from Bogin & Shui (2020) shows that more than a quarter of rural properties are appraised at over 5% above contract price. Additionally, these manual methods come at quite a high cost, as well as taking valuable time to complete (Mooya 2011). Many believe that Automated Valuation Models (AVMs) can overcome many of the disadvantages of on-site manual property valuation, and it is clear that AVMs will be playing a larger role moving forward (Scheurwater 2017). (BQ)

Many market participants have already started using the low-cost appraisals from AVM services, like bank risk managers monitoring the collateral values for mortgage loans, and prospective buyers getting a feel for the expected price of a house (Rainer Schulz & Werwatz 2014). There are two primary approaches to AVMs. One approach is to obtain value estimates directly from regression models, which can be specified in different forms and are estimated using both parametric and non-parametric techniques, and the second approach is to use a comparative sales method (CSM) (Michalis Doumpos & Zopounidis 2020). To make use of CSM, other comparable properties must be identified and referenced when estimating a value for the property in question (Yunjong Kim & Y 2020). While the use of AVMs has become widespread for residential properties, there remain very few applications in the commercial real estate space. This is due in large part to the relative infrequency of trading in this market, meaning less comparable data making the use of a CSM much less effective (Nguyen 2019). (BQ)

Accurate valuation of properties is also of utmost importance from a governmental perspective. Valuations are used to determine the level of property tax payable to the government.

Similar to many other countries, Ireland utilises a system where property tax is based on the value of the property itself as opposed to a Site Valuation Tax (SVT), which takes into account only the value of the site. According to Dye & England (2010), a taxation system of this type "would enhance both the fairness and the efficiency of the property tax". With an SVT, the burden of the tax falls exclusively on landowners, so it could not be passed from landlords to renters. Contrast this to a standard property tax, which taxes both the land and the improvements to the land. This disincentive for landlords to improve their properties also has a negative effect from an environmental standpoint as houses which do not undergo improvements become energy inefficient compared to newer houses and can contribute substantial amounts of greenhouse gas emissions (Hurley & Sweeney 2022). The use of an SVT would mean less of a "penalty" on improvements, and encourage more intensive land use (Oates & Schwab 1997). Additionally, a study from Brueckner (1986) finds that if a land tax encompasses an entire housing market it will reduce the price of housing, and raises the level of improvements. Micheál L. Collins (2011) claims that an SVT could have provided some stability during the 2008 crash, noting the lack of a property tax and Ireland's pro-cyclical fiscal property in the years leading up to the crash. Collins states that an SVT could have provided the government with some stable revenue with its broad base and frequent revaluations. (BQ)

House price modelling is extremely important for both buyers and sellers, as it can help the buyer determine when the best time to buy is, and help the seller determine what the right price should be (Alfiyatin et al. 2017). There are many different ways to model prices. Time-series analysis is used in many studies, such as by Gupta & Miller (2012) which, coupled with co-integrating techniques, helped to anticipate the peaks and downturns in the house price indexes for some of the metropolitan statistical areas they looked at with up to a three-quarter lead time. Another method of analysis is Particle Swarm Optimization (PSO), which is an optimisation based technique used for the selection of affect variables before using a regression model to determine the optimal coefficient of variables (Alfiyatin et al. 2017). This method is used by Wang et al. (2014) and Zhou (2021), where Wang *et al.* show that optimisation has better forecasting accuracy than a BP neural network. Hedonic price analysis is the most commonly used type of model for house prices. Hedonic regression models involve breaking up the house into its different characteristics such as number of bedrooms, number of bathrooms, size, and location, and analysing the effect of these characteristics. This method can be useful because it measures the value of each feature of the house individually, controlling for the other factors. However, it also has some limitations, such as endogeneity issues. Endogeneity issues are ones that arise when the explanatory variable is correlated with the error in the model. Some variables may be simultaneously correlated with house prices, which may give rise to omitted variable bias (Engstrom & Gren 2017). (AC)

2.2 Individual Literature Review

2.2.1 The Housing Market in Ireland and in Limerick

Market Overview

There are 1.86 million habitable dwellings in Ireland (Kennedy et al. 2019), and Ireland has the highest share of people living in houses among all European Union member states, with a proportion of 90% compared to the EU average of 53% (Eurostat 2021). The average Irish house is transacted every 60 years (Maguire et al. 2016), and the infrequency of property transactions has been a key concern for traditional valuation methods (Hurley & Sweeney 2022, O'Hanlon 2011). In January 2023, 3,675 dwelling sales were filed in Ireland with a median price of €305,000, this is an increase of 4.3% from the previous year (Central Statistics Office (CSO) 2023*b*).

County Limerick is home to 205,444 inhabitants, 105,326 of whom live in Limerick city and surrounding areas, Ireland's third largest urban centre (LCC 2023, CSO 2023). In the period between 2016 to 2022 the housing stock in Limerick grew by 4% to 85,363 units, 57.7% of these properties are owner occupied which is considerably lower than the national average of 66.6% (CSO 2023*a*). McCafferty & O'Keeffe (2009) report that apartments make up 70% of inner Limerick city properties and identify a large amount of rented accommodation in Limerick. 13% of city dwellings are rented from the council and 40% of properties in the "studentified" regions surrounding the University of Limerick and Mary Immaculate College are privately rented.

Urban sprawl is apparent in Limerick city, with large-scale housing developments in the Castletroy and Dooradoyle areas (McCafferty 2011). County Limerick consists of four municipalities, including the Metropolitan District of Limerick which contains the city and surrounding areas, and spreads approximately 10km to the villages of Castleconnell and Patrickswell (LCC 2023, Local Area Electoral Boundary Committee (LEABC) 2018). The remaining municipalities constitute the surrounding county region with a population of approximately 100,000 people. Limerick city displays social segregation stemming from socially selective suburbanisation, where more affluent families choose to live in the suburbs, leaving a concentration of low-income households in the city (McCafferty 2009). There is a spatial concentration of poverty in public housing in the city, which displays an imbalanced demography and level of household consumption compared to the rest of the city. McCafferty (2011) identifies the highest national rate of lone-parent households in the poverty-stricken areas of Limerick city along with an increased number of non-Irish nationals compared to the national average. McCafferty encapsulates this social segregation by proposing 6 social area types in Limerick city, defined by family and socioeconomic status, labour market status, and age demographic.

Property Price Modelling

In the modelling of Irish property prices, research is primarily focused on nationwide price modelling. Amongst others, Morgenroth (2014) explored the fundamental influencers of Irish house prices and Lyons (2015) analysed the variability of house prices. O'Hanlon (2011) developed a property price index for the CSO, and Maguire et al. (2016) similarly developed a property price index using sparse national data. Many authors have analysed property prices in the capital city of Dublin, Mayor et al. (2009) study the influence of green spaces on prices, and Roche (2001) models property prices using economic factors. Kitchin (2013) identifies issues with the property sales records in Ireland, which are inadequate in spatial terms and missing features in essential areas. O'Hanlon (2011) reports a similar issue and applies many methods for cleaning the data and imputing missing values from the Residential Property Price Index (RPPI) and financial institutions. The availability of geospatial property data in Ireland has supported the development of geospatial models (Hurley & Sweeney 2022, Dupré 2020).

The implicit modelling of property prices in Limerick has not been deeply explored in the literature. Shinnick (1997) conducted a property price analysis of Ireland, considering Limerick, amongst other cities, as a submarket. The price diffusion of property prices between Limerick and Galway has been explored by Stevenson (2004) and later Gray (2013), who discovered the presence of a weak bi-directional price diffusion. McCafferty & O'Keeffe (2009) spatially analysed properties in Limerick, with a focus on the proportion of apartments and rented properties in the city and surrounding areas.

2.2.2 Geospatial Data & Spatial Effects on Housing Markets

Geographical Information Systems (GIS) support the analysis of data that are spatially referenced to the earth (Burrough & McDonnell 1998, p. 11). This spatial data has coordinates for all observations listed as pairs of longitude and latitude values based on the prime meridian at Greenwich and the equator (Bivand et al. 2008, ch. 1). Spatial data often has attributes containing information for each coordinate point and the World Geodetic System (WGS84) is the most commonly used global reference system for exploring these geospatial instances. The most common formats of spatial data include: spatial points which represent a single point location, such as the geocoded addresses in the Limerick property data; lines representing a set of points joined by line segments; polygons outlining shapes enclosed by lines, such as the Limerick county and municipalities; and grids, which are an evenly distributed collection of points or lines, forming a lattice. Geospatial approaches to property price modelling are predominantly explored using point or polygon data. Point data are used to represent individual properties (Hurley & Sweeney 2022, Oust et al. 2020, Bourassa et al. 2010). Polygons are utilised to represent regions, such as American states, with aggregated census information for each region (Dubin 1988, Adair et al. 1996). Alternative methods of analysis are used to model spatial data. Spatial points pattern

analysis is used to model the observations of a random process within a geographic region, geostatistics is used to model data that could occur anywhere but is measured at a specific location, and analysis of areal data concerns a selection of aggregated polygons (Bivand et al. 2008).

Spatial autocorrelation and spatial heterogeneity are two elements of spatial effects (Anselin & Griffith 1988). Spatial autocorrelation is a spatial dependence between observations. Tobler's first law captures the concept of spatial autocorrelation, it states that "everything is related to everything else, but near things are more related than distant things" (Tobler 1970). Can (1992) describes spatial autocorrelation as a "spatial spill", wherein house prices are affected by the characteristics of neighbouring properties. Spatial heterogeneity represents spatial non-stationary, or variation in space. In the context of a housing market, heterogeneity is demonstrated by a varying relationship between house prices and their characteristics over space (Peng & Inoue 2022). Spatial heterogeneity can be present in the form of spatial heteroscedasticity, where the residual variance is non-constant over space (Anselin 1988). Can (1992) proposes that neighbourhood effects, the exogenous geographical features common amongst neighbouring properties, demonstrate spatial heterogeneity.

Basu & Thibodeau (1998) claim that spatial heterogeneity is demonstrated by segmentation of the housing market. Market segmentation involves dividing the overall market into distinct subsets with shared attributes. The property market is then composed of a set of submarkets, consisting of groups of similar dwellings that are accurate substitutes for one another within, but poor substitutes of properties outside of their groups (Bourassa et al. 1999, Pryce & Gwilym 2013). Homogeneity is apparent within these submarkets due to shared structural and local characteristics of the dwellings (Adair et al. 1996). Basu & Thibodeau (1998) argue that shared neighbourhood characteristics, such as education quality, along with similar economic characteristics, such as household income, also cause this market segmentation.

Spatial autocorrelation occurs when members of the population are related through their geographic location. Dubin (1988) proposes that the distance between points is related to the time lag in a time series where the association between observations is expected to decrease as the distance between them, or lag, increases. Basu & Thibodeau (1998) propose that spatial dependence in house prices is due to properties having a common construction date and similar access to local amenities. Therefore, properties demonstrate common structural and accessibility attributes. Farber & Yeates (2006) use the Moran's I coefficient to measure the spatial autocorrelation (Moran 1950). Using the property characteristics and locations, this test specifies the level at which values are clustered, or randomly dispersed. Geary's contiguity ratio (Geary's C) is an alternative measure (Geary 1954). The presence of spatial autocorrelation has the potential to break the assumptions of independence for ordinary least squares (OLS) models, resulting in inefficient estimates and biased test statistics (Basu & Thibodeau 1998).

2.2.3 Modelling Methods

Empirical Studies

Many authors have analysed the methods for modelling and predicting property prices demonstrating spatial effects.

- The multiple linear regression model (MLR) is a hedonic model proposed by Rosen (1974). This is typically used as a baseline model in empirical studies, and parameters are commonly estimated using **ordinary least squares (OLS)**.
- Dubin (1988) estimates the spatial autocorrelation of Baltimore house prices and uses a maximum likelihood (ML) approach to estimate the regression coefficients, comparing the results to the OLS model.
- Local market disaggregation is modelled using a switching **local regression model** proposed by (Straszheim 1974). In this approach, authors include a dummy variable for market segments in the OLS model (Bourassa et al. 1999, Basu & Thibodeau 1998, Goodman & Thibodeau 1998).
- Dubin (1992) applies a **regression Kriging** approach to Baltimore houses, combining ordinary Kriging and the estimation of coefficients by OLS. Basu & Thibodeau (1998) enhance this approach, calculating the localised regression models for 8 submarkets of 5,000 property transactions in Dallas, Texas.
- **Spatial autoregressive models** are used to model the spatial lag between neighbouring properties (Can 1990, Can & Megbolugbe 1997*a*).
- Properties are modelled geostatistically using localised regression models in a **geographically weighted regression (GWR)** model (Fotheringham et al. 1998). Farber & Yeates (2006) compare this approach to other local regressions for 19,007 Canadian property transactions. Sisman & Aydinoglu (2022) apply a GWR model to 1,475 properties in Istanbul.
- A **generalised additive model (GAM)** is used with smoothing splines to estimate the non-linear relationships between variables. Hurley & Sweeney (2022) use this approach to examine the impact of the reported property addresses of 5,208 properties in Dublin. Dupré (2020) and Shimizu et al. (2014) have similarly applied this approach to property transactions in Dublin and Tokyo, respectively.

Oust et al. (2020) compare the regression Kriging approach and the GWR model, among others, to the OLS modelling approach for 16,417 property transactions in Oslo from August 2016 to December 2017. Similarly, Bourassa et al. (2010) compare a selection of the above models using 13,000 sales transactions from Louisville, Kentucky, and test the results using out-of-sample data.

Multiple Linear Regression Model (MLR)

Hedonic multiple linear regression models have been traditionally used to model house prices. In the European Commission Handbook on Residential Property Price Indices, the hedonic method is the recommended approach for property price indices (Eurostat 2013). Proposed by Rosen in 1974, the model assumes the value of the residential property is the sum of the market value of its characteristics (Rosen 1974, Taylor 2008). The independent variables in the regression fall into three broad categories, location descriptors such as distance to the city centre, structural variables describing the house itself, and neighbourhood variables that categorise income, education levels, and race in the local area (Dubin 1988).

The variables in the multiple linear regression model are adequate for assessing the impacts of externalities on house prices (Oust et al. 2020). In previous studies, structural variables include: house age, living area, and plot size (Farber & Yeates 2006); and the number of bedrooms, and floor area (Shinnick 1997). Hurley & Sweeney (2022) and O'Hanlon (2011) specify property types as detached, semi-detached and bungalow, for example. Pace et al. (1998) incorporate dummy variables for the presence of the property features, such as "carport", "aluminium doors", and "aircon", and Hurley & Sweeney (2022) use similar variables for "attic conversion", "cul-de-sac" and "fireplace". Distance variables predominantly consist of distance to the Central Business District (CBD) (Dubin 1988, Hurley & Sweeney 2022). Farber & Yeates (2006) additionally include distance to the nearest shopping mall.

The parameters of the MLR model are frequently estimated using ordinary least squares whereby the variables are assumed invariant across space and time (Farber & Yeates 2006). Hence, this is a global regression model, which fits a single equation to the relationship between the independent variables and the dependent variable. This leads to difficulty in capturing spatial autocorrelation and spatial heteroscedasticity in the data, and the spatial dependence is captured either in the regressors or in the error term (Jahanshiri et al. 2011). In a study of 16,417 property transactions in Oslo from August 2016 to December 2017, Oust et al. (2020) implement a hedonic regression model with parameters estimated using least absolute distance to reduce the effect of outliers, as a baseline model. Oust *et al.* conclude that the accuracy of the traditional hedonic model is vulnerable to the identification of housing attributes in addition to the effects of time and location. In a similar study, de Haan & Diewert (2013) identify the risk of omitted variable bias in the model where one or more relevant variables are excluded from the model. Basu & Thibodeau (1998) outline that similar values for omitted variables amongst neighbours lead to spatially correlated error terms.

Localised Linear Regression Models

While the global models described thus far model a unitary market, local linear regression models such as a "switching regression" involve a different set of parameters between, but not

within market segments (Goodman & Thibodeau 2003). This allows for the delineation, or clear definition of, submarkets (Straszheim 1974).

There is extensive literature on the methods of assigning houses to submarkets. The a priori knowledge of geographical boundaries, through the use of postcode regions or census blocks, is traditionally used. In a study of 13,000 transactions from Louisville, Kentucky, Bourassa et al. (2010) identify housing submarkets using census demographic data. In a Dallas house price study, Goodman & Thibodeau (2003) utilise ZIP codes to create submarkets. Hurley & Sweeney (2022) define submarkets in Dublin using postcode regions. Basu & Thibodeau (1998) use school district boundaries and a larger neighbourhood region composed of census blocks. The use of statistical classification methods, such as K-means clustering, has been incorporated to group similar properties in distinct submarkets (Bourassa et al. 1999, Helbich et al. 2013). Fuzzy clustering, where observations can belong to multiple clusters, is used to cater for intersecting submarkets (Gabrielli et al. 2017, Hwang & Thill 2009).

Dummy variables are used to represent submarkets in the MLR model, resulting in separate intercepts, and thus equations for each submarket. In a comparison of multiple linear regression methods, Eckert et al. (1990) observe the best predictive abilities in the regression model with well-defined submarkets. Goodman & Thibodeau (2003) report a 50% reduction in the mean and variance of the proportional error when submarkets are used in regression modelling. Improvements in prediction accuracy have been identified when smaller submarkets are used, in an empirical study by Basu & Thibodeau (1998), ZIP code boundaries produced more accurate results than the larger census groups. Goodman & Thibodeau (2003) argue that modelling smaller submarkets is constrained by the definition of their boundaries. Additionally, Bourassa et al. (2010) argue that increasing the number of submarkets is limited by the amount of data for such groups. In Ireland, large rural areas are defined by county councils (Collins & Larragy 2011). Postcode regions are defined in Dublin, however, Hurley & Sweeney (2022) outline a lack of well-defined postcode areas outside of Dublin. O'Hanlon (2011) concludes that house price analysis at a neighbourhood level is limited in regions outside of Dublin due to the lack of well-defined submarkets.

The addition of dummy variables representing submarkets does not appear to mitigate the issue of spatial autocorrelation. Basu & Thibodeau (1998) study 5,000 dwellings in Dallas, Texas, reporting spatial autocorrelation of house prices within 1,200km of 4 out of the 8 submarkets, and spatially correlated residuals throughout 2 of the 8 submarkets. In a study of spatial autocorrelation among 221 Baltimore dwellings, Dubin (1988) assumes the residual correlation of properties follows a negative exponential of the distance between them. Dubin uses a maximum likelihood estimation method (Mardia & Marshall 1984) to estimate the regression parameters based on the autocorrelation function.

Interaction Effects

Additive multiple linear regression models examine the price as a sum of the predictor variables and only model the additive effects of each independent variable on price. Hence, including location terms results in modelling location as an independent premium added to house prices. Fik et al. (2003) propose the use of interaction terms in the modelling of house prices, suggesting an interaction be included between absolute location (Cartesian coordinates) and structural attributes, which have spatial regularities, or a pattern across space. Such interaction terms have been used in various studies; Troy & Grove (2008) identify a significant interaction between distance to the nearest park and the combined rape and robbery rate of a neighbourhood. In a study of 2,907 property transactions from Tucson, Arizona, Fik et al. (2003) use a second-order polynomial regression of the housing attributes, Cartesian coordinates, and location dummy variables. Similarly, Bitter et al. (2007) include mean-centred Cartesian coordinates in an interactive model. By including interactions between Cartesian coordinates and other predictor variables, the parameters can vary over space. Fik et al. (2003) and Bitter et al. (2007) report an improvement in the predictive performance of the model and a decrease in the autocorrelation of model residuals after including the interaction effects, thus, removing the spatial dependence in the residual term.

The expansion model proposed by Casetti (1972) is often used to include interaction terms in the model. The model is used to understand the spatial drift observed in geographical processes such as the labour market (Pandit & Casetti 1989). It isolates terms demonstrating spatial heterogeneity or "spatial drift" (Kestens et al. 2006). The model creates a prediction for each point in space, where each parameter is a function of the other variables; polynomial regression can be used for such estimations (Fotheringham et al. 1998). Hence, the parameters can vary over space, whereby spatial complexity increases with the complexity of parameter estimation.

Geographically Weighted Regression (GWR)

Multiple linear regression models consist of a single set of parameters and a unitary, or global, equation for all observations. Since the model involves a single equation and intercept for all observations, the MLR model represents the global average of all observations (Wheeler & Calder 2007). There is a need for parameter variation across space to capture spatially varying relationships between variables (Jing & Fotheringham 2016). The geographically weighted regression (GWR) model is a local regression model, where each point has a distinct set of parameters and intercept term (Farber & Yeates 2006). The set of parameters of each regression model has a corresponding distance decay matrix based on neighbouring entries. In the case of non-stationary relationships, the GWR model has been found to outperform the global hedonic MLR models. Oust et al. (2020) report improved accuracy in a study of property transactions in Oslo, where the median error for the OLS and GWR models were 12.1% and 6.65%, respectively. This is

attributed to the improved spatial complexity of the GWR models, such that the spatial patterns demonstrated by the dependent variable are captured through modelling spatially varying relationships (Fotheringham et al. 1998). However, Farber & Yeates (2006) note the presence of extreme coefficients in the model, where a coefficient value is an outlier compared to the values of all other observations. This may be due to the sensitivity of the parameters to local irregularities in the data used to calibrate the model (Oust et al. 2020). In a comparison of the GWR model with the expansion model, Bitter et al. (2007) conclude that the GWR model accounts for heterogeneity at both local and global levels. Increased predictive ability and explanatory power are reported for the GWR model as it captures the spatially varying relationships between variables. The GWR model is often used as a secondary exploratory tool, for example, Kestens et al. (2006) use the expansion method along with the GWR model to explain both the complex interactions and non-stationary spatial effects influencing house prices.

Spatial Autoregressive Models

Can (1990) identifies spatial autocorrelation and heterogeneity of linear regression model residuals when the observations are on a two dimensional plane, thus, causing a violation of the model assumption of independence. Can proposes two autoregressive models to incorporate spatial dependence in the market. The autocorrelation error model assumes the error is a weighted average of neighbouring residuals. The spatial autoregressive model includes a spatially lagged price variable, representing the average sales price of neighbours with a distance weight (Farber & Yeates 2006, Can 1992). A mixed autoregressive model includes the spatial lag variable in combination with other MLR model predictors, this assumes price is a function of the structural, neighbourhood, and accessibility attributes in addition to the values of neighbouring properties (Kobylińska 2021, Can & Megbolugbe 1997*b*).

The weight matrix, proposed by Can & Megbolugbe (1997*b*), uses the Euclidean distances between properties within a certain radius. Oust et al. (2020) use this approach to collect 15 neighbouring properties per observation. They raise awareness of rural dwellings with relatively little or no neighbours, and in such cases, the local average value is used. Spatial autoregressive models follow the simultaneous autoregressive model (SAR) or the conditional autoregressive model (CAR). The former, described above, is most applicable to geographical studies (Pace & Barry 1997). Pace & Gilley (1997) employ a SAR model by including the weighted residuals of neighbouring properties in the MLR model and find the relative goodness of fit to be 1.86 times that of the OLS model. Bourassa et al. (2003) compare the same approach to localised regression models and observe little improvements in predictive abilities when submarkets are geographically defined in the localised regression model.

Regression Kriging

Kriging is a method of spatial interpolation with origins in mining geology, the spatial dependence between neighbouring observations is modelled as a function of the distance between them (Thompson et al. 2012). Predictions at new locations are made using a weighted linear combination of neighbouring values, with weights chosen to minimise error variance (R. Webster 2001). Basu & Thibodeau (1998) estimate the weights using a semivariogram, such that the weights are based on the spatial correlation between neighbouring entries. Following on from her previous work, Dubin (1992) uses data on Baltimore house sales to examine the spatial autocorrelation of residuals in regression house price models. Omitting all neighbourhood and accessibility measures, she examines residual spatial autocorrelation using a negative exponential correlogram and concludes that the geostatistical approach outperforms the localised linear regression model.

Residual Kriging is an explicit two-stage process, which combines the regression modelling of the variables of interest with ordinary Kriging of residuals from the regression model (Hengl et al. 2007). In the modelling of single-family houses in Dallas, Texas, Basu & Thibodeau (1998) apply a localised regression model with spherical autocorrelation to the 8 submarkets using estimated generalised least squares. Basu & Thibodeau conclude that the geostatistical technique outperforms the OLS model in terms of predictive ability in 6 out of 8 of the submarkets. Bourassa et al. (2010) apply the regression Kriging approach to localised regression models of disaggregated submarkets and conclude that the model outperforms both the OLS and GWR models, with 47.4% of predictions within 10% of the true value. In a simulation study comparing the GWR model with the Kriging approach, Harris, Fotheringham, Crespo & Charlton (2010) report that the regression Kriging approach with submarkets outperforms the GWR model.

Generalised Additive Models

Assuming linear relationships between variables in hedonic regression is not sufficiently complex to account for "non-linear structured patterns" within variables (Hurley & Sweeney 2022). Shekhar et al. (2015) outline that the parameters of the covariance errors in the spatial autoregressive models are uniform over space, thus, the model is inadequate at capturing the underlying landscape variation. Generalised additive models (GAM), an extension to linear regression models, allow for the estimation of price as a series of general functions of the independent variables, where each function is estimated nonparametrically through smoothing splines, locally weighted regression or the kernel method (James et al. 2023, p.289-311). Pace (1998) concludes improved prediction using a GAM compared to the MLR model to predict the prices of 442 Memphis houses.

In two separate studies of Dublin house prices, GAMs are used to geospatially model house prices (Dupré 2020, Hurley & Sweeney 2022). Hurley & Sweeney model a closed interval subset of the data using smoothing splines. Dupré notes that such models with continuous Bayesian smoothing splines are sensitive to geographical features, such as rivers or coastal boundaries. Approximating values across such boundaries is known as "finite area smoothing", and soap film smoothing is a computationally efficient proposed solution (Wood et al. 2008). The benefits of GAMs are outlined by James et al. (2023), where transformations do not have to be attempted for each variable, and smoothness can be graphically summarised for each spline. Hurley & Sweeney note the ease of interpretability of the smoothing functions in the model compared to alternative regression and machine learning models.

Including non-linearity in the model increases its explanatory power, however, there is evidence of overfitting. Shimizu et al. (2014) compare the out-of-sample prediction accuracy of linear models with a non-linear GAM using 500 random samples of 9,682 out-of-sample data points, and they conclude that the GAM failed to outperform linear models.

2.2.4 Research Questions

Following the above review of the statistical models for property prices and associated empirical studies, the following research questions are used to outline the aims of this analysis.

1. What are the structural and locational attributes of properties in Limerick that can be used to build an interpretable multiple linear regression property valuation model with a high level of accuracy, and what attributes significantly impact property price?
2. Are the differences between properties in the rural and urban areas of County Limerick indicative of market segmentation and how can this be incorporated in the modelling of property price?
3. By implementing a geographically weighted regression model, is there evidence of spatially varying relationships between certain independent variables and property price?

Chapter 3

Data Preparation

3.1 Limerick Property Dataset

The Limerick property dataset has details of over 12,435 property transactions in County Limerick ranging from January 2011 to December 2018. The data, provided by 4Property Ltd, has been scraped from property listing websites Daft.ie™ and MyHome.ie™, as well as the Residential Property Price Register from the Property Services Regulatory Authority (PSRA). Due to the combination of sources, the data is relatively sparse, as demonstrated in Figure 3.1. The original variables and respective descriptions are outlined in Table 3.1.

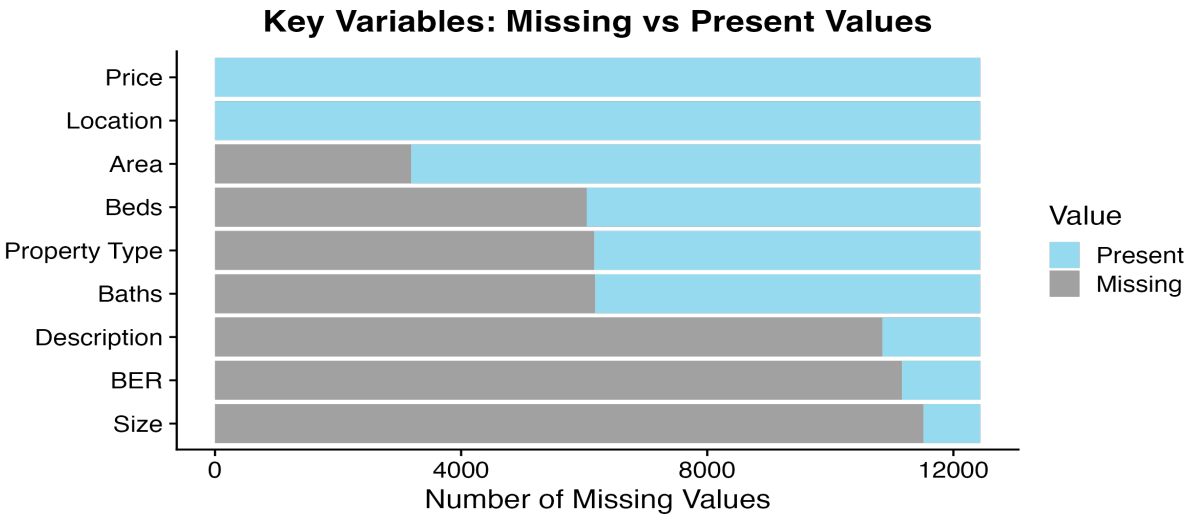

Figure 3.1: Missing vs present values of variables of interest.

Table 3.1: Initial variables and descriptions.

Variable	Description
Price	Property price in euro (€)
Property Category	Classifies properties as 'House', 'Apartment' or 'Land'
Original Property Type	Text mined property type describing structural and design styles
Location	Property locations in spherical coordinates
Area	Regional area of property based on raw address
Beds	The number of bedrooms
Baths	The number of bathrooms
Description	Text description of the property
BER	Building Energy Rating value
Size	Property size in units squared

Approximately half of the entries are from the property listing websites and contain information on the property type, the number of bedrooms and the number of bathrooms. Only 100 of the overall listings contain information on both the building's energy rating (BER) and property size. A proportion (1,587) of the entries from listing websites also contain a property description which is mined for additional structural and design characteristics of the house. All properties have a raw text address and spherical coordinates (longitude and latitude). Over half (5,888) of the entries have no information on the property type or structural characteristics, such as the number of bedrooms and bathrooms. These entries are most likely from the Residential Property Price Register, which only stores the property address and sale price. Due to a lack of information on the properties, these entries are excluded from the analysis.

The creation of new location and property attributes from original variables are described in sections 3.2 and 4.3, respectively. A summary diagram of the creation of these attributes is shown in Figure 3.2.

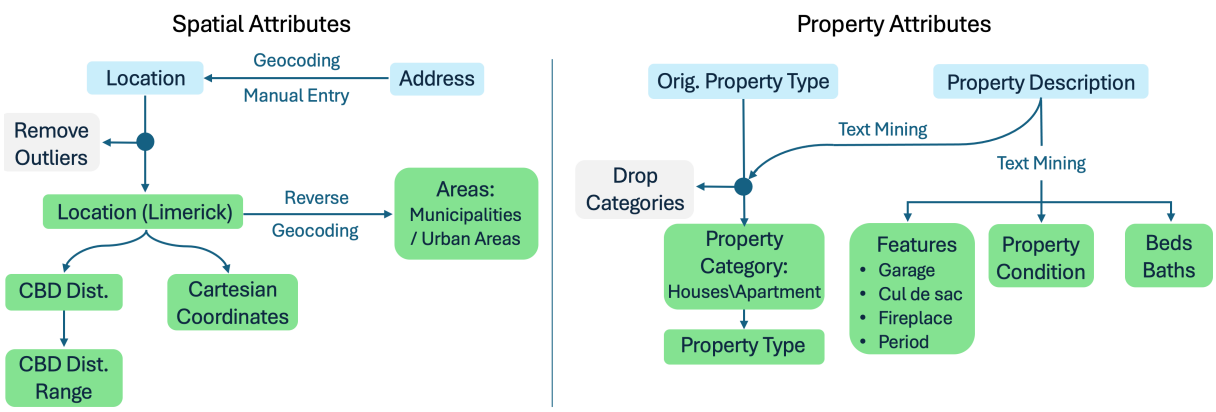

Figure 3.2: Summary of the creation of new (green) variables from original (blue) variables.

3.2 Locational Attributes

The *sp* package developed by Pebesma & Bivand (2005) is used to analyse the spatial data, this contains various classes for analysing and transforming spatial points and polygons stored as shapefiles. Upon plotting the entries using the original longitude and latitude coordinates, properties with extreme coordinate values were identified. These properties were incorrectly geocoded to the United States, France, and other counties in Ireland. To rectify this issue, a radius of 50km was created around the centre point of County Limerick, and any points outside this radius were inspected. For each of the outliers, new spherical coordinates have been obtained using a *Google Maps* API and the *ggmap* package in *R*. Due to ambiguities in the text address, the approach was unsuccessful for some entries, and such entries were manually assigned coordinates based on the provided address and Eircodes. On the final iteration, 9 properties were removed as they are located outside of County Limerick. A shapefile of County Limerick was obtained from Tailte Éireann (2024), and an additional 12 non-Limerick properties were removed when the dataset was merged with this geographically defined county border.

Properties were incorrectly assigned to the regions "Limerick city" or "County Limerick". Through a visual analysis, a large proportion of properties in "County Limerick" were found to be located in the inner city and suburbs. These properties were investigated to rectify the cause of the error. The city properties with erroneous region assignments contain the word "Limerick" in the address, whereas correctly assigned city properties contain the words "Limerick city".

Similarly, pre-assigned property areas have been investigated. These values were originally taken from the raw addresses, and thus a proportion of properties had missing values for this variable due to an insufficient number of terms in the addresses. Many of these entries were incorrectly assigned to areas. This became apparent when the plotted values for certain areas revealed a dispersion across the county. For the purposes of this analysis "Limerick" refers to all observations in the county and "Limerick county" refers to all properties outside of the Limerick city and suburbs area.

Property Areas

Due to the inaccuracy of region and area assignments, there is potential for uncertainty and bias when data is modelled at a local level. The spherical coordinates have been utilised to assign the correct region and area values to each property. For each property, the longitude and latitude values were reverse geocoded using the *tidygeocoder* package in *R*. This procedure resulted in a classification of properties into suburbs, villages, towns and municipalities. The process was unsuccessful for 26 properties, information for these properties was obtained from *Google Maps* and manually entered. The municipalities represent 4 broad county regions such as "Cappamore

and Kilmallock Area" (CK) and "Limerick city and suburbs" (LCS).[1]

The shapefile of County Limerick contains polygons representing the municipalities geographically. Additional errors in the assignment of municipalities were discovered when observations were plotted within the shapefile regions. These errors were rectified similarly to the County Limerick border, where the shapefile was merged with the data to assign the correct municipalities to each point. The geographical area of these municipalities with their corresponding observations can be observed in Figure 3.3. The urban sprawl of Limerick city, described by McCafferty (2011), is evident in this map, whereby the entries in Limerick city and suburbs spread far beyond the city centre point.[2]

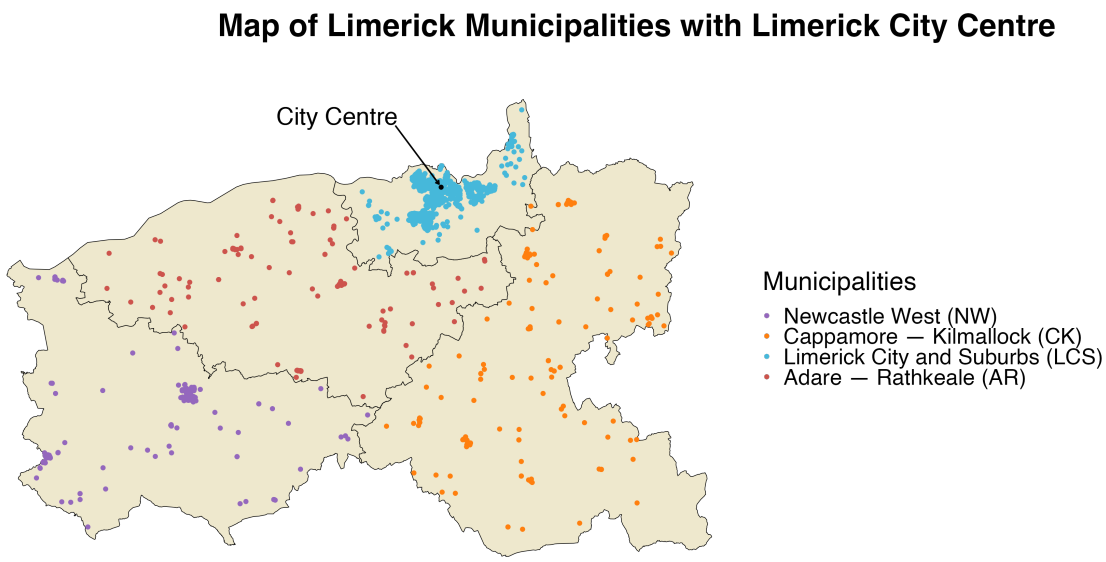

Figure 3.3: Map of properties in their respective municipalities with the Limerick city centre point.

In the results of the reverse geocoding, suburbs are only present for the Limerick city and suburbs municipality, while villages are only present for properties outside of this municipality. These entries have been merged into a new area column with 112 unique areas in the dataset. Following the approach of Adair et al. (1996), who amalgamate the planning authority regions to form submarkets in Belfast, electoral divisions in Limerick city and suburbs are merged to define broader urban areas in a shapefile using the geographic information systems software *QGIS*. These urban areas follow the general shape of the regions defined by Limerick County Council (LCC 2022*b*, p. 4). Further subdivisions are created to define the Outer City South and

[1]The Limerick municipalities are formed by the electoral areas of Limerick City & County Council (LCC 2023). The city and suburbs, formally the Metropolitan District of Limerick, will henceforth be referred to as "Limerick city and suburbs" to improve clarity.

[2]The Limerick city centre point is taken to be the Limerick Milk Market due to its history as the centre of trade in Limerick. The term "central business district" (CBD) is used in line with previous empirical studies (Hurley & Sweeney 2022, Oust et al. 2020).

City North, and to divide the city into City North and City South. The urban areas have been assigned codes to aid visualisation and the communication of results. These areas are outlined in Table 3.2. Notable addresses, including the Ennis Road, North Circular Road (N.C. Road) and Castletroy, are mentioned throughout the analysis. A map of LCS divided into urban areas with the locations of such addresses is shown in Figure 3.4.

Table 3.2: Summary statistics of LCS properties by urban areas.

Identifier	Urban Area	Frequency
LCS-1	North	141
LCS-2	City North	47
LCS-3	Corbally	68
LCS-4	City South	98
LCS-5	West	33
LCS-6	Ballycummin	209
LCS-7	Outer City South	152
LCS-8	Ballysimon	196
LCS-9	East	68

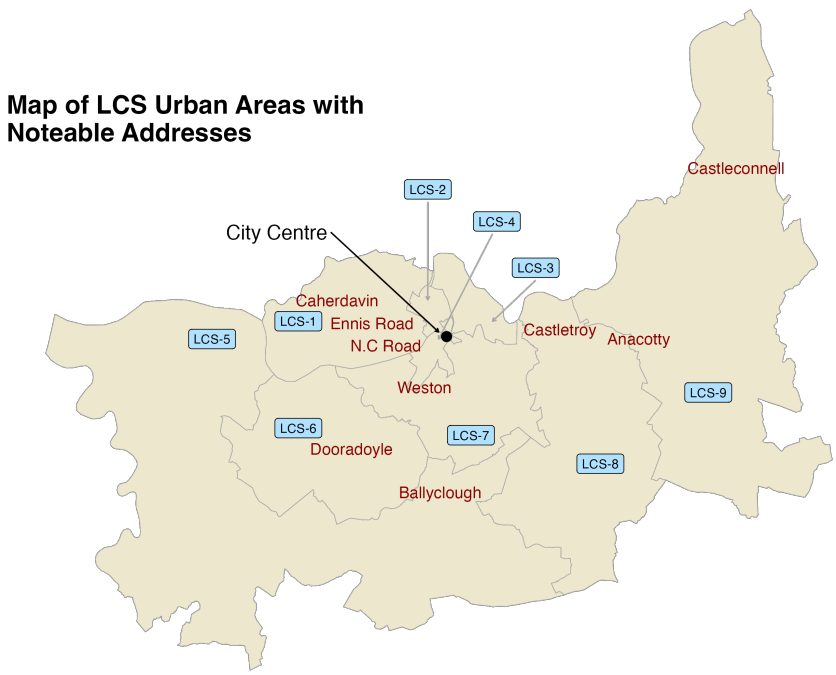

Figure 3.4: Map of LCS urban areas with notable address locations.

CBD Distance & Cartesian Coordinates

Inverse distance calculations require non-zero distance values between values for finite results. Thus, properties such as multi-storey apartment blocks with duplicate coordinate values were rectified by adding a minor increment to the longitude and latitude of the duplicate coordinates. The distance between each property and Limerick city centre is calculated using the haversine method, which calculates the shortest distance in metres between points on a spherical earth (Robusto 1957). A categorical variable is created to categorise properties based on their CBD distance values. Properties are labelled as either less than 5km, between 5km and 10km, between 10km and 20km, or over 20km from the city centre. Cartesian coordinates are assigned to the properties based on longitude and latitude values using a Pseudo Mercator projection, a cylindrical map projection for 85.06°S and 85.06°N (Stefanakis 2017).

3.3 Structural Attributes

Proceeding with houses and apartment entries in the dataset, all attribute levels and missing values are explored before removing unwanted entries, ensuring the integrity of assigned labels and retaining as much of the remaining data. Data with a populated text description is utilised for the text-mining of information, these descriptions consist of web text from the property listing websites. The text-mining phase is a lengthy and iterative process, where each step begins with the investigation of random samples of the property descriptions of interest. These investigations allow for the identification of common words, phrases, and special characters, from which regular expressions have been developed. On each iteration of the text-mining, results are clarified using samples of the relevant data, and expressions are adapted if necessary. For each attribute of interest, the common search terms and key findings are outlined.

Property Category

Properties were pre-assigned property categories from either house, apartment or land. The original property types are a result of the text-mining by the data provider, these have been utilised to categorise entries. Common entries are grouped into property categories, for example, listings such as "Investment Property", "Investment Opportunity" and "Investment Property for sale" are assigned to an "Investment" category.

Entries from each newly defined category are examined, whereby relevant entries are extracted and assigned to either the "House" or "Apartment" categories. The key actions undertaken are outlined as follows:

- Sixty-two "New Development" properties with a missing property category have been reclassified as houses. These consist of properties in housing estates in the Castletroy and Annacotty areas.
- Entries in the "Investment" category are removed, such entries contained no information on property type or structural attributes. The mean price for this category is €327, 571, compared to the overall mean of €157,355. Due to the lack of information on the properties, it is difficult to identify the reasons for such price differences.
- Properties from the "Hospitality", "Commercial" or "Industrial" categories were removed, these are not of interest in this analysis of residential properties.
- Before removing the "Land" category, the descriptions have been mined for keywords such as "house" and "dwelling". There are multiple occurrences of houses with large sites. Such entries have been reclassified as houses.

Similar text-mining expressions were run within the property categories. This resulted in the clarification of 7 apartments that were misspecified as houses.

Property Type

The property types consist of many interlinking designs and structural classifications of properties. Commonly, design types consist of "Bungalow", "Dormer" and "Cottage", while property (structural) types consist of "Detached", "Terraced House", and "Townhouse". The property types of houses are shown in Table 3.3 and there is clear ambiguity surrounding the terms. Design types are subsets of property types, for example, a bungalow is a subset of other structural types such as detached or semi-detached. A similar case is true for cottage and dormer dwellings. There are also missing values and an ambiguous "House" type that requires investigation.

Table 3.3: Property types of houses before data cleaning.

Property Type	Count	Property Type	Count
Bungalow	46	House	2
Cottage	5	Semi-Detached House	578
Detached House	392	End-of-Terrace House	61
Terraced House	203	Townhouse	42
Dormer	1	NA	24

Taking a similar approach to Adair et al. (1996), distinct levels outlining the structural characteristics of houses were created. These categorise houses as detached, semi-detached, terrace, end-of-terrace or townhouse. Each house is assigned to one of these property types based on the text-mined keywords. The approach is successful for a large proportion of entries, the challenges encountered are outlined as follows:

- On the final iterations of the process, the remaining properties were bungalows. Based on the text descriptions, these properties were found to be located on individual sites in the countryside; thus, these properties are classified as detached houses.
- Four properties were found to be commercial listings, consisting of shop units with overhead accommodation. These units are not included in the analysis as they cannot be compared with residential properties.
- Many house listings contain keywords from the land category such as "plot", "site" and "development". These entries consist of sites for sale with descriptions outlining the planning permission, such entries have been removed from the analysis.

A similar approach is repeated for the apartment category, where keywords include "studio", "apartment", "penthouse" and "flat". Duplex properties remain a unique property type, aligning with the original property type. The property types and their respective categories from the complete dataset are shown in Table 3.4.

Table 3.4: Property categories and types after data cleaning.

Property Category	Property Type	Count
Apartment	Apartment	160
	Duplex	26
House	Detached	425
	Semi-Detached	594
	Terraced	179
	End-of-Terrace	61
	Townhouse	52

Property Features

Property features are obtained by mining for keywords in the text description. The resulting property features are described below.

- Garage: The presence of a garage was identified by mining for words such as "outbuilding", "shed", and "barn".
- Cul-de-sac: Properties situated on a cul-de-sac have been identified. Variations and mistakes in the realtor's spelling of cul-de-sac were accounted for.
- Fireplace: The presence of a fireplace was identified by mining for words such as "mantlepiece", "open fire", and "fireplace".
- Parking: As the data contains dwellings in the countryside, many detached properties have parking as standard and the word "parking" may not be mentioned in text descriptions.
- BER: As mentioned in Section 3.1, there are missing BER values. Text mining is not possible for these values since possible rating values are not mentioned in property descriptions.
- Glazing: It is not required to list the glazing specifications of each property in the property advertisement. Mentions of window specifications range from specific terms such as "triple glazing" to general terms such as "PVC windows" or "big windows". Due to the lack of information surrounding window glazing specifications, this attribute is not used for the analysis.

The variables 'Garage', 'Cul-de-sac' and 'Fireplace' were created to mark the presence of such features.

Property Design

The potential to assign property designs based on text mining results is explored. The following insights were attained from examining the property design types for certain samples of the data.

- Property descriptions often list specific design styles for houses, such as "Edwardian", "Tudor" or "Georgian". These properties are grouped as period properties.
- Some property styles are related to the locations of properties, such as "Country House" and "Farmhouse" for properties located in rural areas.
- Property design types are not described for a large proportion of common property types, especially 3-bedroom semi-detached houses located in suburban estates.

Due to the lack of well-defined design types, this category is not used for this analysis. Period properties were identified by assigning a Boolean variable, 'Period'.

Property Condition

Property condition was predefined as either new or second-hand, this characteristic has been extended to include properties in need of renovation and derelict properties. Insights that follow include:

- Text descriptions mentioning "derelict" houses are categorised as derelict.
- Properties in need of renovation have been identified by mining for phrases such as "upgrade", "refurbishment", and "renovation". Expressions were modified to exclude renovated properties or properties with features in need of upgrade, such as decking or gardens.

The Number of Bedrooms and Bathrooms

The text descriptions are used to identify the number of bedrooms and bathrooms for properties missing such entries. For listings missing the number of bathrooms, descriptions have been mined for the count of key terms such as "en-suite", "main bathroom" and "WC". Many properties listed the number of bedrooms in phrases such as "four-bed house" or "3 bedroom property". To account for the variation in such descriptions, the characters before the words "bed" or "bedroom" were extracted. Where this attempt was unsuccessful, the occurrences of the letters "bed" were counted. It became apparent that such keywords are often repeated in property descriptions. Thus, results were checked systematically against the true values.

3.4 Complete Dataset

the cleaned dataset consists of 1,532 properties ranging from 2014 to 2018, the dispersion across the years is not equal as seen in Table 3.5. Due to a low density of entries in the years 2014 to 2016, data from 2017 and 2018 is selected for analysis, ensuring the temporal effect can be captured through the use of dummy variables representing time periods in a similar manner to Oust et al. (2020). There are now only two variables of interest with missing values, 'BER' has 264 (18%) missing values and 'Size' has 775 (52%) missing values.

Table 3.5: Cleaned data by year of sale.

Year	2014	2015	2016	2017	2018
Count	1	3	18	722	775

The data cleaning and filtering described in the previous sections have resulted in reduced variability of the dependent variable, apparent in Figure 3.5. The interquartile range (IQR) of price has reduced from €133,000 to €105,000, following the data cleaning. There is a reduction in the magnitude of outliers in Figure 3.5B, this is likely due to the removal of non-residential properties. The median value of price has increased from €135,000 to €183,000 following the process. The increase in median value is likely attributed to the removal of the "Land" category, which contains a large number of site listings and has a median value of €89,000.

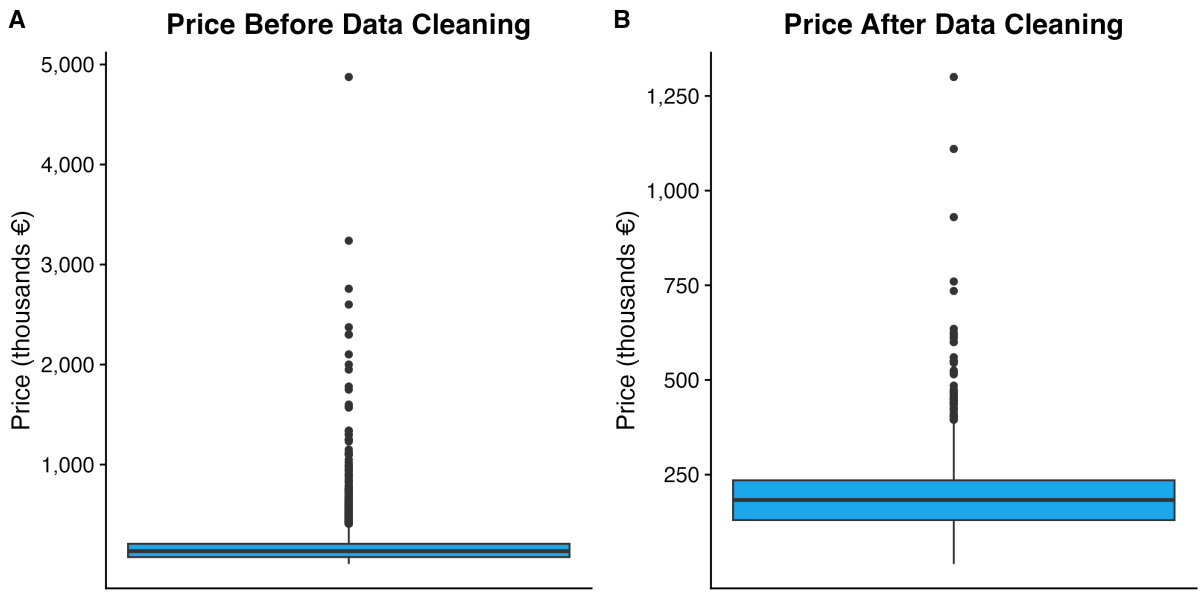

Figure 3.5: Distribution of price: A - before data cleaning, and B - after data cleaning.

Chapter 4

Exploratory Analysis

4.1 Limerick Property Dataset

The dataset described in Section 3.4 consists of 1,497 property transactions in Limerick from January 2017 to December 2018. Combining the results of previous demonstrations, a summary of the variables of interest is presented in Table 4.1. In a similar manner to Dubin (1988), property attributes are divided among structural and neighbourhood categories. Structural characteristics are specific to the property while neighbourhood characteristics are shared among neighbours.

Table 4.1: Complete data with variables and descriptions ordered by category.

Category	**Name**	**Description**
Dependent Variable	Price	Property price in euro (€)
Structural	Beds	The number of bedrooms
	Baths	The number of bathrooms
	Property Category	Classification as a house or apartment
	Property Type	Structural classification of the property
	Property Condition	Classification of property condition
	Period	The property is a period property
	Garage	The property has a garage
	Fireplace	The property has a fireplace
Neighbourhood	Cul-de-sac	The property is located on a cul-de-sac
Locational	Region	Limerick county or Limerick city and suburbs
	Municipality	One of AR, CK, LCS, or NW
	Urban Area	Dummy variables for areas in LCS
	CBD Distance	Distance to Limerick city centre (km)
	CBD Distance Range	Ranges of distance to Limerick city centre (km)
	(x, y)	Cartesian x and y coordinates
Temporal	Year	Year of property sale
	Months	Month and year combination of property sale

The remaining categories consist of locational variables, the dependent price variable, and the year and month-year combination as temporal variables.

The data is relatively equally divided over the two years, with 52% (775) of observations in 2018, and 48% (722) of observations in 2017. The distribution of properties in respective categories and regions is outlined in Figure 4.1. There are 1,012 properties in the Limerick city and suburbs area, and 485 properties are in the remaining. There are 93 apartment listings each year, and there are 629 and 682 house listings in 2017 and 2018, respectively.

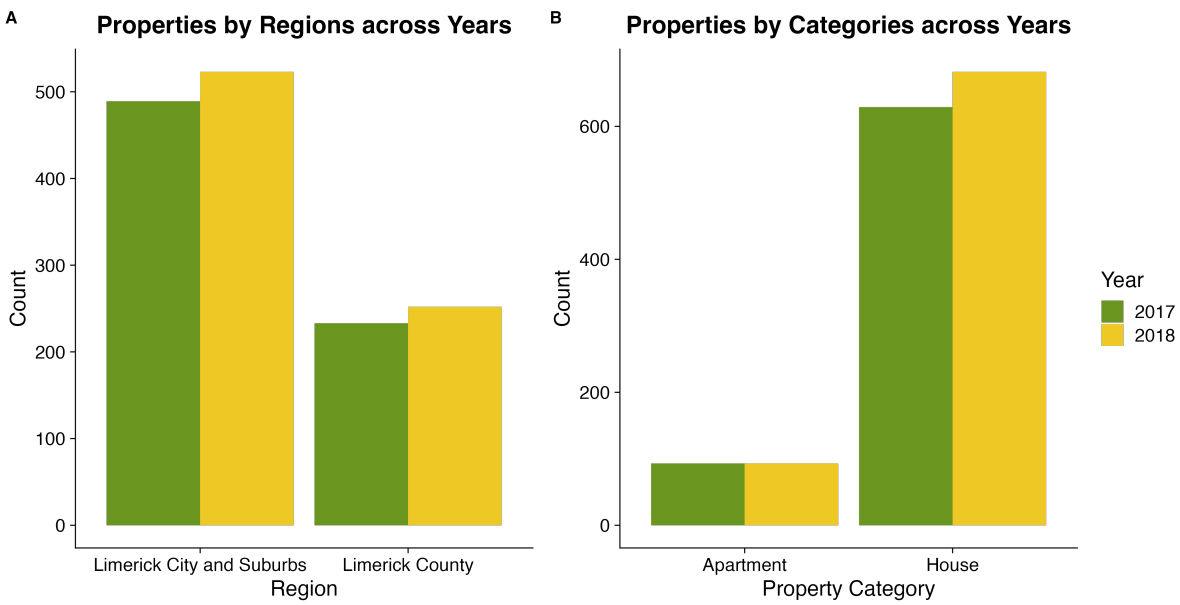

Figure 4.1: Bar plot of properties across regions, years and property categories.

The mean price of observations in 2017 and 2018 are €187,105 and €199,347, respectively. Daft.ie™ report mean property prices of €169,812 and €189,218 for Limerick in 2017 and 2018, respectively (Daft.ie 2017, 2018). The average values from the Limerick property dataset are greater than those reported by the property listing site but demonstrate a similar increase and difference between the years. This difference may be due to the sampling methods used, such as the period of selection and property types included in the calculation.

The following exploratory analysis is divided among locational variables and property attributes, namely structural and neighbourhood categories. To aid visualisation, 12 outliers with prices over €600,000 are removed from the graphs following the initial price exploration. These observations, representing 0.8% of the overall data, are described at the end of this section and are included in the modelling of property price.

4.2 Locational Attributes

Using the reverse geocoded addresses from Section 3.2, the properties are analysed in their respective municipalities or urban areas. The number of properties along with the median price and interquartile range (IQR) of observations within each area is outlined in Table 4.2. At a municipal level, the Limerick city and suburbs region has the greatest median house price of €196,000, while Newcastle West has the lowest median house price of €135,000. The greatest variability of property prices is in the Adare-Rathkeale region, with an interquartile range of €145,000.

Table 4.2: Summary statistics of all properties in municipalities and urban areas of LCS.

Identifier	Municipality	Urban Area	Frequency	Median Price (€)	IQR (€)
AR	Adare - Rathkeale	-	125	170,000	145,000
NW	Newcastle West	-	198	135,000	69,300
CK	Cappamore - Kilmallock	-	163	140,000	101,500
LCS	Limerick city & suburbs		1,012	196,000	95,600
LCS-1		North	141	225,000	105,000
LCS-2		City North	47	180,000	145,000
LCS-3		Corbally	68	166,900	97,000
LCS-4		City South	98	132,000	56,100
LCS-5		West	33	183,000	173,500
LCS-6		Ballycummin	209	212,000	66,000
LCS-7		Outer City South	152	155,800	83,100
LCS-8		Ballysimon	196	235,700	83,800
LCS-9		East	68	221,000	93,900

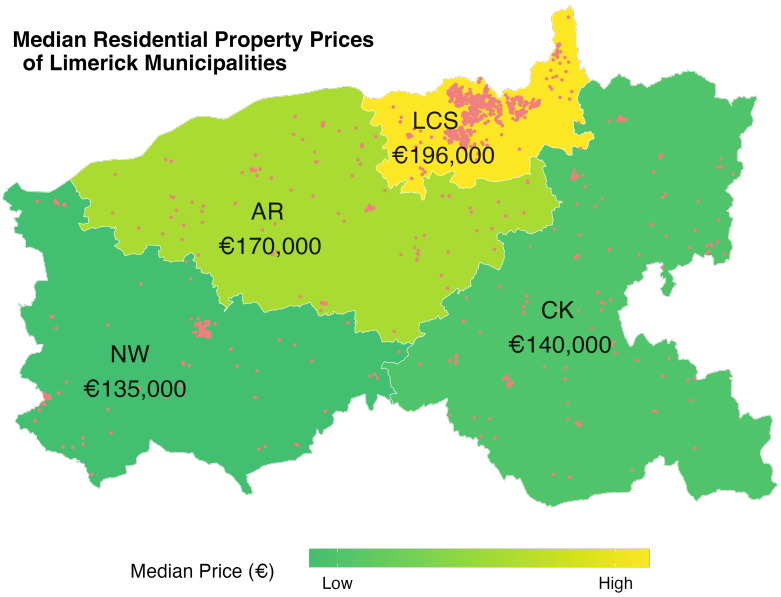

Figure 4.2: Map of median prices by municipality with observation locations in red.

The municipal median prices are plotted by location in Figure 4.2. The differences in the median price of properties in LCS compared to NW and CK are apparent. The median property price in the AR municipality for this period is closer to that of LSC compared to the other two municipalities. The locations of observations are plotted in red to highlight the sparsity of property transactions among the municipalities of AR, NC and CK compared to LCS.

The distribution of property prices by municipalities and urban areas is plotted in Figure 4.3, where the observations in LCS are analysed by the corresponding urban areas. The variability of AR is evident in comparison to the other areas. In LCS, there is an observable decrease in median price from €225,000 in LCS-1 to €132,000 in LCS4, while the median values are higher for LCS-5 to LCS-9, save for LCS-7. Analysing the map of urban area locations in Figure 4.4B the numbering of urban areas starts at north Limerick and the location moves in a clockwise direction towards the inner city centre to LCS-4. This is suggestive of a decrease in property prices as the distance to Limerick city centre reduces in an eastward direction. Similarly, the indexes from LCS-5 to LCS-9 correspond to the southern urban areas moving eastward. The proximity to Limerick city centre is greater for such areas except for LCS-7, which is the outer city with a reduced median price of €155,800.

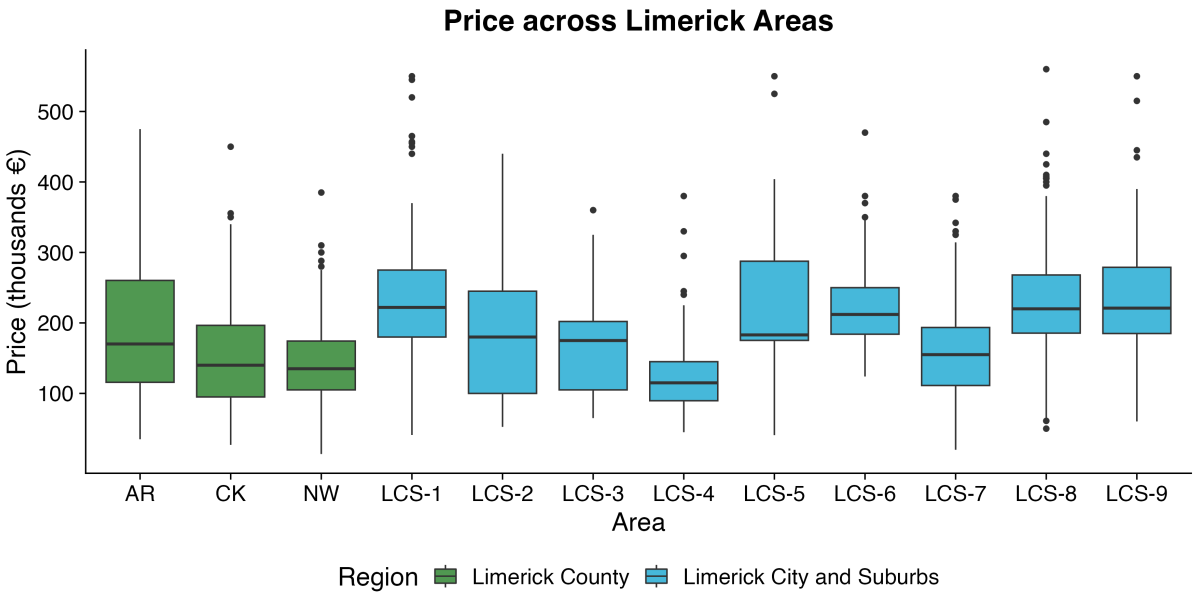

Figure 4.3: Box plot of price across the municipalities and urban areas of Limerick.

In Figure 4.4A, the sparsity of data in NC, AR and CK challenges the identification of spatial patterns. In NC, there are many lower-priced properties near the west border and a "hot spot" of similarly priced properties in the centre of the municipality at the town of Newcastle West. There is an apparent belt of higher-priced properties in AR, wrapping around the border with LCS, this spatial pattern may be explained by accounting for distance to the city. The town of Adare consists of a cluster of high-priced properties compared to the surrounding properties in the AR municipality. In CK, the southern properties appear to be lower priced and gathered in clusters, while there is a similar band of higher-priced properties on the northern side of the municipality. The most apparent clusters of similarly priced properties are in LCS.

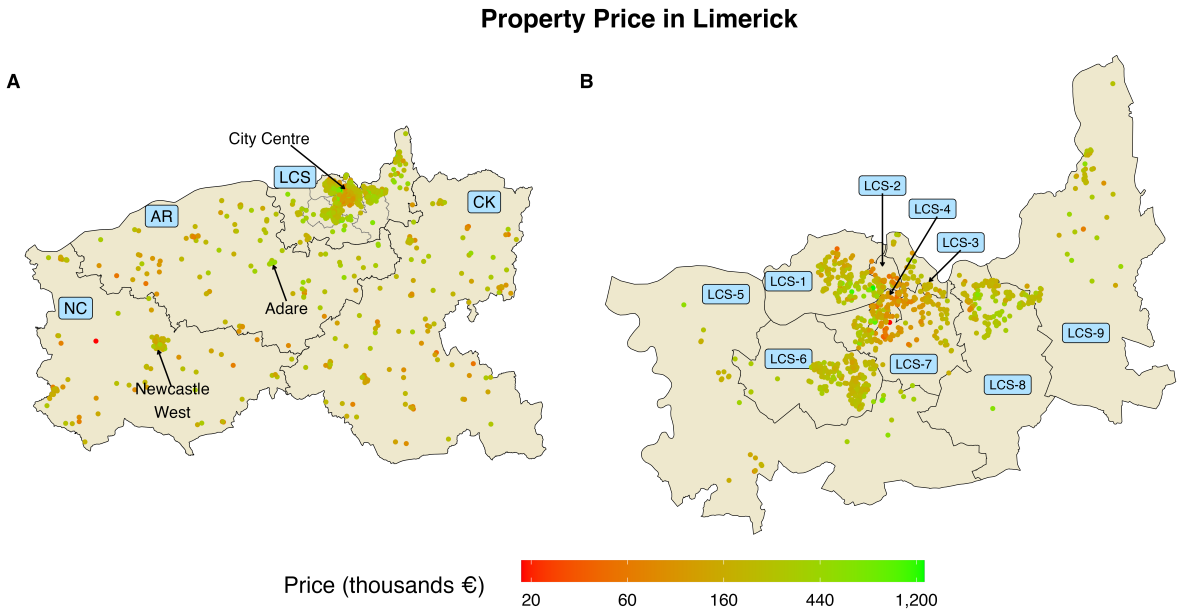

Figure 4.4: Map of property price (incl. outliers) in: A - County Limerick, and B - Limerick city and suburbs.
Note: $\log(Price)$ is plotted using an exponentiated legend to aid visualisation.

The property prices within the urban areas of LCS are plotted in Figure 4.4B and the addresses highlighted in Figure 3.4 should be referenced to aid interpretability. The densely populated inner city areas of LCS-2, LCS-4 and LCS-7 contain most of the low-priced properties. The cluster of high-priced properties on the south side of LCS-1 are those with the notable addresses of the Ennis Road and the North Circular Road. Properties in LCS-6 appear to be similarly priced with a median price of €212,000, most likely due to homogeneous properties within the many housing estates in Dooradoyle. Castletroy, in LCS-8, contains many expensive properties with a relatively large price range, evident from the spread of the box plot in Figure 4.3. This is likely related to the prevalence of larger properties in this area, with an average of 3.69 bedrooms, which is greater than the LCS average of 3.3 (Table A.2 of Appendix A). The surrounding areas of LCS-9 and LCS-5 contain fewer observations compared to the inner city areas. LCS-5 contains the greatest variability, with an IQR of €175,000, this is likely due to the large spatial scale of

the area. LCS-9 has a similar price structure to LCS-1 with a median price of €221,000.

Distance to Limerick City Centre

The CBD distance for each municipality is plotted against property price in Figure 4.5A. There is a distinction between LCS and NW at the lower and upper scale of CBD distance, respectively. The distance values for CK and AR have a similar range, aligning with the map of property locations in Figure 4.4. There are notable breaks in the relationship between price and distance, highlighted by vertical lines at the 5, 10 and 20-kilometre points. Properties are grouped by these ranges and plotted by distance ranges in Figure 4.5B. There appears to be a positive relationship between distance and price for observations within 5km, where property price increases as distance from the CBD increases. The observations between 5 and 10km of the city centre are relatively uniform in price and have the greatest median price among all groups. Properties within 10km of the CBD represent LCS, the above findings align with the visual analysis in Figure 4.4B, where a cluster of lower-priced properties is apparent in the city centre and similarly priced properties are located in the surrounding areas.

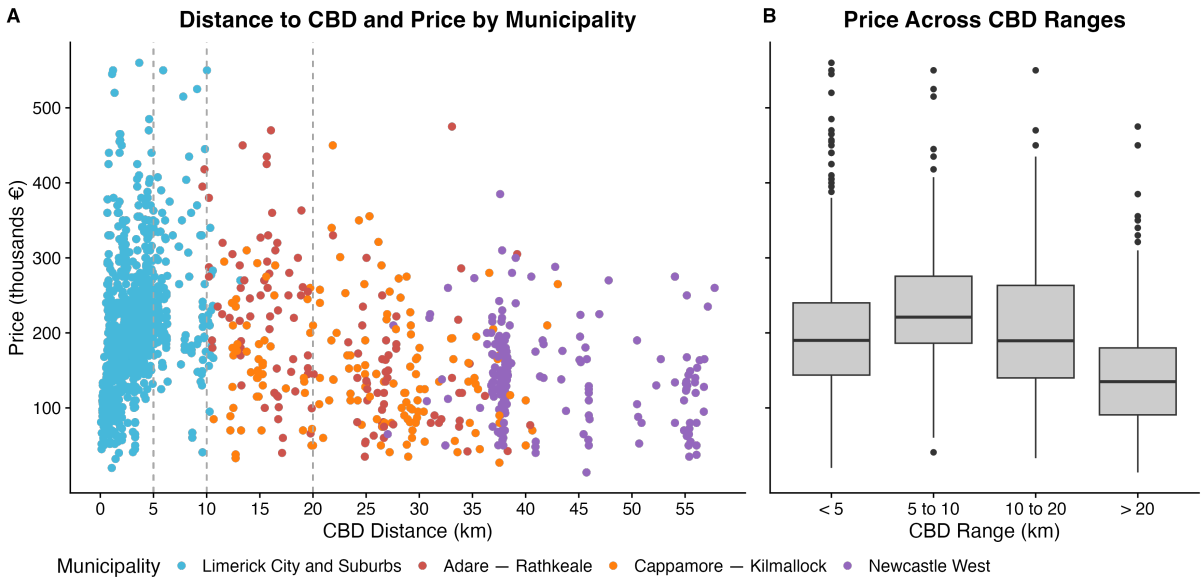

Figure 4.5: A - Scatter plot of CBD distance and price by municipality. B - Box plot of price by CBD distance range.

The relationship between price and CBD distance appears to be negative for distances between 10 and 20 kilometres, suggesting a decrease in property price for increases in distance from the city. The median price for properties over 20km from Limerick city centre is the lowest among all bands at €130,000. The relatively constant relationship between price and distance within this band suggests a diminishing negative effect of distance on price, where the price decreases associated with increased distance in this range are less than those in the 10-20 kilometre band.

At a macro-scale, the general trend aligns with the median price pattern observed in Figure 4.2, where the median property prices are lower for municipalities further from LCS.

4.3 Property Attributes

Property Type

Using the property types defined in Chapter 3, the submarkets are compared in terms of property type in a similar manner to Adair et al. (1996). The differences in price between groups across regions can be observed in Figure 4.6. The left plot compares the prices of houses and apartments across regions and the right plot compares the prices of house types across regions.

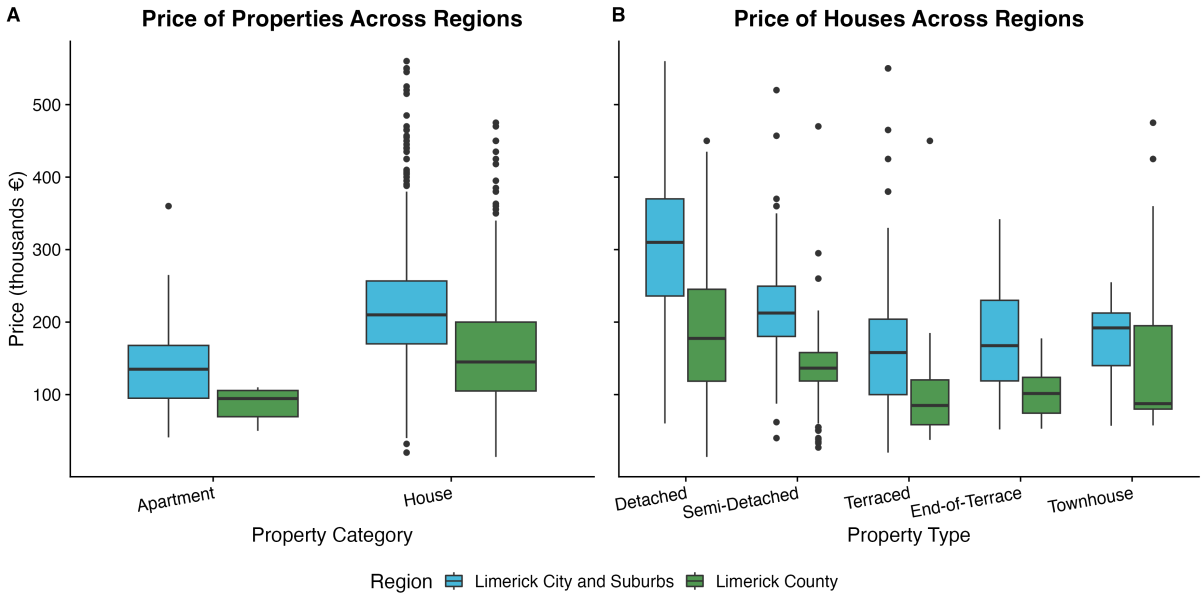

Figure 4.6: Box plot of price across regions: A - for property categories, and B - for house types.

The summary statistics outlined in Table 4.3 aid the interpretation of Figure 4.6. Detached houses have the highest sales price and greatest variability among property types, with a median value of €220,000 and an IQR of €160,300. Semi-detached is the most popular property type demonstrating reduced variability compared to that of detached houses, there are 594 observations with an IQR of €79,800. Terraced properties have a median price of €130,000, while end-of-terrace properties have an increased median price of €166,000. Duplex properties and apartments are combined in the apartment category of this analysis. The median price of a duplex property in Limerick is €176,000, which is greater than that of an apartment (€126,500). The highest priced apartment is a 2 bedroom property with a price of €360,000. This is located in LCS-4 and mentions a "prestigious" building complex.

Table 4.3: Summary statistics of properties across property types.

Property Type	Frequency	Median Price (€)	IQR (€)
Apartment	160	126,500	63,700
Duplex	26	176,000	52,800
Detached	426	220,000	160,300
Semi-Detached	594	192,000	79,800
Terraced	179	130,000	109,000
End-of-Terrace	61	166,000	115,000
Townhouse	52	185,000	102,400

In Figure 4.6A, the mean price of both houses and apartments is higher in LCS compared to Limerick county. The median prices of houses are €196,000 and €142,500 in LCS and Limerick county, respectively. There is a marginal difference in the variability of prices in the two regions, the IQR in Limerick county is €97,500, and the IQR is €95,600 in LCS. It is apparent in Figure 4.6B that the median price is greater in LCS than in Limerick county across all property types. Summary statistics for property types of houses across regions can be found in Table A.1 of Appendix A and key findings are outlined as follows:

- The median price of a detached house is €316,000 in LCS, and €176,000 in Limerick county. The IQR of detached houses in LCS is €137,500 compared to an IQR of €128,000 in Limerick county.
- The median price of a semi-detached house is €71,000 greater in LCS than in Limerick county. The number of transactions of semi-detached houses is 450 in the city and suburbs, compared to 114 in the Limerick county region.
- Price differences are apparent across terrace, end-of-terrace and townhouses, the median price of all three property types is higher in LCS compared to Limerick county. There is also an apparent difference in the sales count of such property types, most notably, there are 55 and 6 end-of-terrace house observations in LCS and Limerick county, respectively.

The distribution of property types in the county is shown in Figure 4.7, the most popular property types are selected to aid visualisation. The predominance of detached houses among the sparsely distributed Limerick county region is apparent. Semi-detached houses mainly occur in clusters in these areas, representing local towns and villages. In LCS, 18% of property observations are apartments (Table A.2), mainly located in the city centre. The centre is surrounded by semi-detached properties in all directions. This aligns with the prevalence of housing estates in the city-surrounding suburbs and may be related to the homogeneity of property price within the 5-10 kilometre band from the CBD in Figure 4.5.

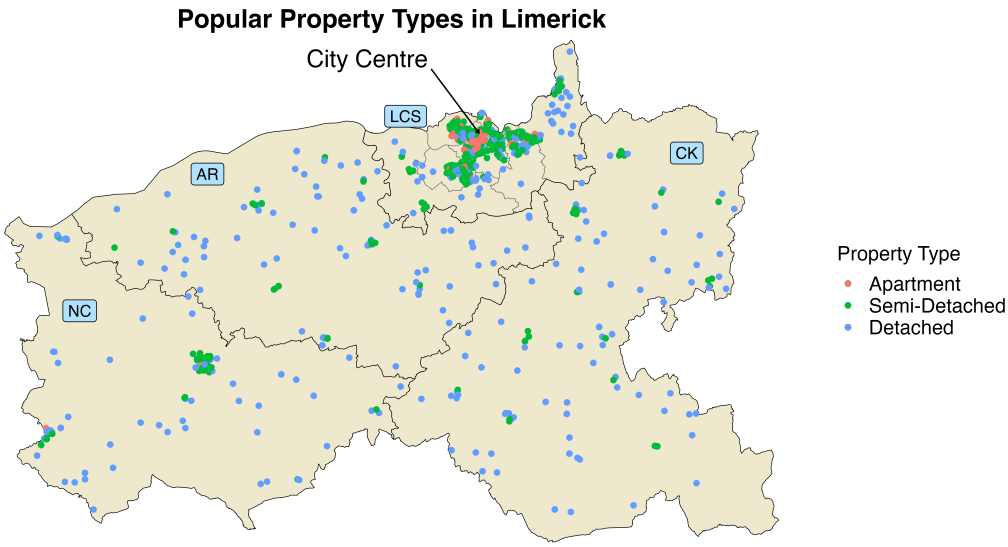

Figure 4.7: Map of the most popular property types in Limerick.

The Number of Bedrooms and Bathrooms

In Figure 4.8, there is an observable positive relationship between price and both the number of bedrooms and the number of bathrooms. Pearson's correlation coefficients are 0.48 and 0.54 for price and the number of bedrooms and bathrooms, respectively. These indicate a moderate positive linear relationship in each case. There are apparent outliers in each of the below scatter plots. The Pearson's correlation coefficient between the number of bedrooms and the number of bathrooms is 0.45, suggesting a moderate positive linear relationship between the variables. This does not raise immediate concerns with collinearity, where predictor variables are highly correlated with one another, these issues are investigated in the following sections (James et al. 2023, pp. 99-102).

Observing Figure 4.8A, there are apparent outliers in the number of bedrooms. The 7-bedroom properties are investment properties for student accommodation in the Castletroy area of LCS-8. The properties with 5 and 6 bedrooms consist of detached bungalows, semi-detached family properties, and country houses located across all municipalities. There are 17 properties with 1 bedroom in LCS, 14 of which are apartments. In Limerick county, there are 5 houses with 1 bedroom, these consist of cottages and a terraced house.

In Figure 4.8B, there are apparent outliers in the number of bathrooms. There are 4 properties with 0 bathrooms located in the Limerick county region. The property condition is described as derelict or in need of restoration, these properties have been assigned to the "derelict" property condition as defined in Chapter 3. The properties with a greater number of bathrooms align

Note: The values are jittered for interpretability, where a small increment is added or subtracted from each region to distinguish points at discrete values.

Figure 4.8: Scatter plot of the number of: A - bedrooms, and B - bathrooms, across regions.

somewhat with the houses with over 5 bedrooms. For example, there is a 6-bedroom and 6-bathroom Georgian country house located in CK with a sales price of €360,000. There are outliers in the prices of medium-sized city properties, these are highly priced 3, 4, and 5-bedroom houses in LCS. These properties mainly correspond with the previously seen houses with highly sought-after addresses in LCS-1 and LCS-8. The lower-priced medium-sized properties are predominantly located in the Limerick county region. Many of these properties outline a need for "refurbishment" or "modernisation". Property types in the lower price range mainly consist of terraced houses in many rural towns and villages such as Cappamore, Doon, and Glin.

Property Features

In Figure 4.9, the proportions of garages and cul-de-sacs are plotted for each urban area in LCS. LCS is plotted here rather than the whole of Limerick as the smaller urban areas demonstrate more observable differences. LCS-4 contains the lowest proportion of both garages and cul-de-sacs, this effectively captures the architecture of the city centre which is made up of townhouses, apartments and terraces (McCafferty 2011). In comparison, LCS-1 and LCS-9 contain higher proportions of garages, likely due to the increased space in the areas. LCS-6 contains the highest proportion of cul-de-sacs, this is most likely due to the prevalence of housing estates in Dooradoyle, where each estate may have multiple cul-de-sac lanes.

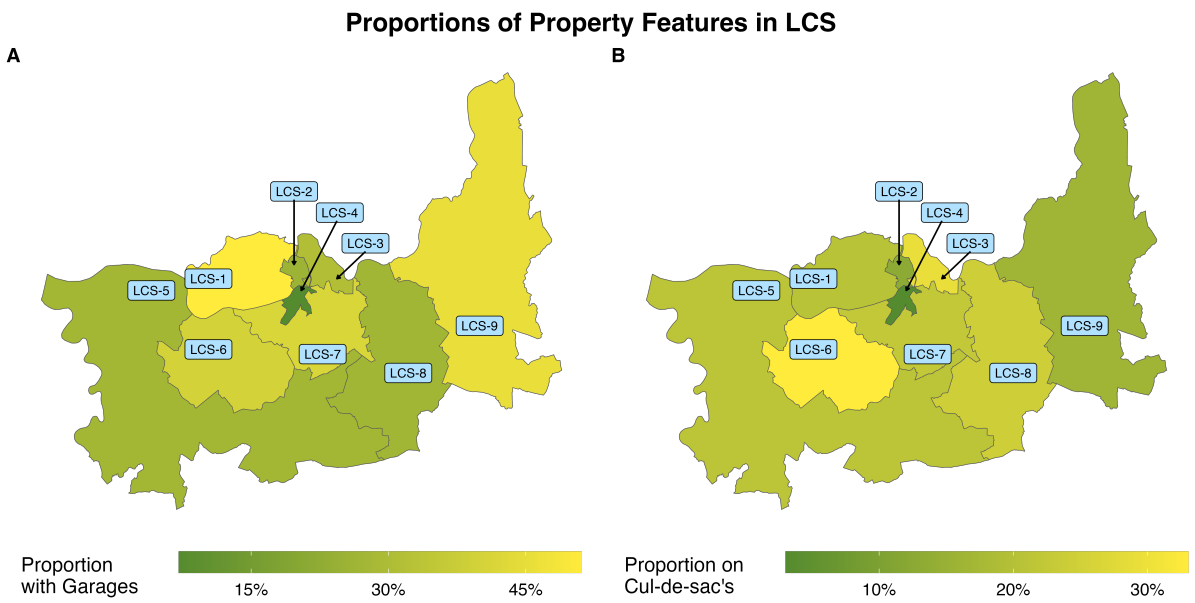

Figure 4.9: Map of the proportion of properties: A - with garages, and B - on cul-de-sacs, across LCS areas.

4.4 Outliers

The observations removed from the previous graphs are 12 properties with values over €600,000, these are mainly located in the LCS region. These outliers consist of period houses in highly sought-after areas, such as the Ennis Road and the North Circular Road (LCS-1), Ballyclough (LCS-5), and Castletroy (LCS-8). The most expensive property in the dataset is an Edwardian family home on the Ennis Road with a sales price of €1,300,000. Two of these properties are 7-bedroom student-letting properties in Castletroy while the remaining outliers contain 5-6 bedrooms. All of the outliers are houses with a range of property types. There is a 6-bedroom Georgian end-of-terrace house located on O'Connell Avenue in LCS-4 with a sale price of over €900,000. The highest-priced terraced house is a 3-bedroom period home on the Ennis Road with a sales price of €735,000.

The outliers on the lower end have been retained during the exploratory analysis and consist of small apartments, houses in need of refurbishment, and derelict properties. The lower-priced houses are more dispersed amongst the regions of NW, CK, and LCS. The lowest observed entry is a derelict cottage with a price of €14,000 in NW. The lowest-priced terraced house is a 3-bedroom property located in Weston (LCS-7) with a sales price of €20,000, the property description mentions "complete refurbishment required".

Chapter 5

Modelling Methods

In this chapter, the model theory and estimation method for the linear regression model and the geographically weighted regression model are described in Sections 5.1 and 5.2, respectively.

5.1 Multiple Linear Regression

Multiple linear regression (MLR) models are used to model the linear relationship between the independent variables and the dependent variable. A unitary equation fitted to all observations is termed the global model and is represented in the general case in Equation 5.1 (James et al. 2023, ch. 3.2).

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \qquad i = 1, .., n \tag{5.1}$$

where y_i is the dependent variable, the price of property i for n observations. Here, β_0 is the constant intercept representing the mean of $\log(Price)$. The jth predictor variable for the ith observation is represented as x_{ij}, with a corresponding scaling coefficient β_j. The error terms are assumed to be independently and identically normally distributed with constant variance, $\varepsilon_i \sim N(0, \sigma^2)$. The equation can be represented in matrix form as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{5.2}$$

where $\boldsymbol{Y} = [y_1, ..., y_n]^\mathsf{T}$ is the observed dependent variable in vector form and $\boldsymbol{X}$ is the design

matrix containing the explanatory variables $[\boldsymbol{X}_1, ..., \boldsymbol{X}_p]$ from Equation 5.1 along with a leading column of ones for the intercept. Here, $\boldsymbol{X}_j = [x_{1j}, ..., x_{nj}]^\mathsf{T}$ represents the jth explanatory variable in vector form. The intercept β_0 and coefficient values β_j are represented in vector form as $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_p]^\mathsf{T}$ and the residual values ε_i correspond to a $n \times 1$ vector $\boldsymbol{\varepsilon}$.

Ordinary least squares (OLS) is used to estimate the coefficients $\boldsymbol{\beta}$ by minimising the sum of squares residual which is the difference between the true value and the OLS model prediction value of the dependent variable. In this case, the parameters are estimated as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{Y}. \tag{5.3}$$

For linear regression models, there is assumed to be a linear relationship between the independent variables and the dependent variable (James et al. 2023, pp. 93-103). Osborne & Waters (2019) recommend the need for normally distributed variables that are measured without error. Residual assumptions are assessed by plotting the standardised residuals against fitted values and the normality of residuals can be assessed using a Q-Q plot of the standardised residuals or by using a Kolmogorov-Smirnov test.

The assumption of independent residuals means residuals are uncorrelated with each other. Residual correlation can lead to underestimating the standard error of model coefficients, in turn reducing the width of confidence intervals (James et al. 2023). Additionally, the p values may be underestimated, potentially resulting in erroneously concluding significant effects. James et al. (2023) outline additional concerns with outliers and leverage points in addition to collinearity, which inflates the standard error of coefficients.

The MLR model can include dummy variables to represent regional areas, resulting in a model for each region. Location is alternatively incorporated by including the distance measures, such as distance to the city centre, or Cartesian coordinates. Dummy variables could also represent the property type and the presence of certain property features, such as a "fireplace", or neighbourhood characteristics, such as a "cul-de-sac".

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{k=1}^{q} \gamma_k z_{ik} + \varepsilon_i \qquad i = 1, .., n \tag{5.4}$$

Equation 5.4 follows from Equation 5.1 with the addition of dummy variables. Now, x_{ij} is the jth independent variable for observation i, capturing the continuous effects of the number of bedrooms or the number of bathrooms, for example. Dummy variables for observation i are represented as $\boldsymbol{Z}_i = [z_{i1}, ..., z_{iq}]$, where z_{ik} indicates the presence of a certain property

characteristic or area with a regression scaling coefficient γ_k. In Equation 5.4, the regression terms β_0, β_j, and γ_k have estimates $\hat{\beta}_0$, $\hat{\beta}_j$, and $\hat{\gamma}_k$, respectively.

Due to the positive skewness in property price, the natural log transform of price is often applied and modelled as the dependent variable (Bourassa et al. 2010, Oust et al. 2020, Hurley & Sweeney 2022). This is further explored in Chapter 6. Equation 5.5 follows directly from Equation 5.4 with a natural log transformation applied.

$$\log(y_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{k=1}^{q} \gamma_k z_{ik} + \varepsilon_i \qquad i = 1, .., n \tag{5.5}$$

5.2 Geographically Weighted Regression

The geographically weighted regression (GWR) model is an extension of the traditional model in Section 5.1. This non-parametric technique consists of an ensemble of local regression models where parameters are estimated using subsets of the data at different locations, rather than a global model which uses all observations (Wheeler & Páez 2009). Estimating the parameters at different locations allows one to model spatially varying relationships between the independent variable and the dependent variable. The model is fitted at regression points i, which can correspond to locations in the observed data with Cartesian coordinates (u_i, v_i). The observations used to estimate the model coefficients are called the calibration points. Using a similar notation to Wheeler & Páez (2009), the model is represented as

$$y_i = \beta_{i0} + \sum_{k=1}^{p} \beta_{ik} x_{ik} + \varepsilon_i \qquad i = 1, .., n \tag{5.6}$$

where y_i is the dependent variable or price of property i for n observations. The intercept, β_{i0} represents the mean $\log(Price)$ for the ith property. There are p explanatory variables in the model and x_{ik} is the value of the kth explanatory variable for property i with corresponding local regression coefficient β_{ik}. The error terms, $\varepsilon_i \sim N(0, \sigma^2)$, are assumed to be independently and identically normally distributed with constant variance (Farber & Yeates 2006). In a similar manner to Equation 5.5, the natural log transformation could be applied to the dependent variable, and dummy variable covariates could be included in Equation 5.6.

The regression model is alternatively represented as

$$y_i = \mathbf{X}_i \boldsymbol{\beta}_i + \varepsilon_i \tag{5.7}$$

where $\mathbf{X}_i = [1, x_{i1}, ..., x_{ip}]$ is a $1 \times (p+1)$ row vector of explanatory variables for location i with a leading 1 for the intercept term, β_{i0}. The term $\boldsymbol{\beta}_i$ is a $(p+1) \times 1$ vector of the corresponding regression coefficients and the intercept value. The observed price at location i is y_i.

The assumptions of the GWR model naturally follow from the MLR model (Comber et al. 2023). However, there is an awareness of Tobler's first law and the inherent spatial dependence between observations. This is demonstrated by the spatial structure of residuals, where the value of the dependent variable at location i depends on the independent variables of surrounding locations. In a list of model assumptions for the GWR model, Harris, Fotheringham & Juggins (2010) state that residual correlation is assumed to have a negligible effect on model results. Non-constant residual variance has been observed through the geographical space, contributing to biased and inefficient coefficient estimates (Charlton et al. 2009).

The GWR model enables local parameter estimation, the vector of parameters $\boldsymbol{\beta}_i$ in Equation 5.7 can be thought of as a realisation of the continuous functions $\boldsymbol{\beta}$ at the ith location (u_i, v_i) (Brunsdon et al. 1996). The model allows for the estimation of parameters based on the dependent variable and a set of independent variables measured at known locations (Fotheringham et al. 1998). To estimate the parameters of the model at a specific point, it is necessary to weight the surrounding observations such that the parameters of the ith regression point are based on the calibration points proximal to this location.

This is achieved using weighted least squares, where the weights constituting the diagonal of a weight matrix, $\mathbf{W}$, are used for the coefficient estimation (Charlton et al. 2009). The least squares estimation in Equation 5.3 includes $\mathbf{W}$ to become a weighted least squares estimation in Equation 5.8. Here, a constant weight w_{ii} is applied to each square difference before minimising them, such that observations with increased inaccuracy are penalised more.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\boldsymbol{Y}. \tag{5.8}$$

The weighting of the GWR model coefficients is similar to Equation 5.8, however, the weight assigned to a calibration point j is based on its proximity to regression point i. Thus, the regression equation for location i has an individual distance decay weight matrix (Fotheringham & Oshan 2016).

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}^\mathsf{T}\mathbf{W}_i\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}_i\boldsymbol{Y} \tag{5.9}$$

In Equation 5.9, $\boldsymbol{\beta}_i$ is estimated from 5.7 using 5.8 with an individual weight matrix. The diagonal distance-decay weight matrix for regression point i is $\mathbf{W}_i = \text{diag}[w_{i1}, ..., w_{in}]$. The entry w_{ij} is based on the proximity of the calibration point j to the location i. The design matrix, $\mathbf{X}$, follows from Equation 5.3, and $\boldsymbol{Y}$ is a $n \times 1$ vector of the dependent variable for each calibration point.

Model Calibration

The weight matrix must be calculated at each location i before estimating the parameters of the local regression model for such locations (Wheeler & Páez 2009). The weight matrix is defined using a weighting function called a kernel, this function assigns more weight to entries that are proximal to location i. Weight functions have a bandwidth, this dictates the spatial range and is used to specify the number of neighbouring observations used to estimate the parameters of the local regression for the ith location (Comber et al. 2018).

Using the notation of Gollini et al. (2013), a selection of common kernel functions is outlined in Table 5.1 where w_{ij} corresponds to the weight for observation j from the calibration point i with coordinates (u_i, v_i). The distance between entries i and j is represented as d_{ij}, and b is the kernel bandwidth parameter.

The global model assigns equal weight to all observations, the discontinuous boxcar function is a step function that assigns a weight of 1 to all points within a distance d from regression point i and assigns a weight of 0 to all other observations. The bi-square function is continuous and differentiable with a similar cut-off distance to the boxcar function. The exponential and Gaussian kernel functions are both continuous incorporating a distance decay relationship, such continuous functions are most typically used (Charlton et al. 2009, Hu et al. 2015).

Kernel bandwidth, b, is an unknown quantity and must be estimated using the observations. The bandwidth choice depends on whether the kernel is adaptive or fixed (Wheeler & Páez 2009). For a fixed kernel, the bandwidth b is a distance measure with the same units as d_{ij}. This is constant for all locations and weights are affected by the volume and distribution of entries in the two-dimensional space. Fotheringham et al. (2003) outline the need for homogeneously geographically distributed data for fixed kernels.

For the adaptive kernel, the bandwidth is based on a specification of the number of nearest neighbours to include in the model kernel, whereby the spatial extent of the kernel adjusts for variations in density to keep this count constant. In Table 5.1, the value b is the distance to the

Table 5.1: Selection of (adaptive) kernel functions used for the GWR model.

Global Model	$w_{ij} = 1$
Boxcar	$w_{ij} = \begin{cases} 1 & \lvert d_{ij} \rvert < b \\ 0 & otherwise \end{cases}$
Bi-square	$w_{ij} = \begin{cases} (1 - (d_{ij}/b)^2)^2 & \lvert d_{ij} \rvert < b \\ 0 & otherwise \end{cases}$
Gaussian	$w_{ij} = exp(-\frac{1}{2}(\frac{d_{ij}}{b})^2)$
Exponential	$w_{ij} = exp(-\frac{d_{ij}}{b})$

furthest calibration point to include in the kernel. To ensure the count is constant within the kernel, the spread is reduced for dense clusters of data. For sparse data, the spread of the kernel increases. In the case of the Gaussian and exponential kernel functions, the adaptive bandwidth is a representation of the local number of neighbours used for model estimation, whereas the actual number of neighbours is all observation points due to the continuity of the functions (Gollini et al. 2013).

It is assumed that the choice of bandwidth and kernel function is suitable for the study (Harris, Fotheringham & Juggins 2010). Brunsdon et al. (1996) propose the use of cross-validation (CV) or Akaike information criterion (AIC) as the most widely used approach for bandwidth estimation. The CV process uses a subset of the data to iteratively search for the kernel bandwidth such that the prediction error of all y values in the subset is minimised. The bandwidth estimate, $\hat{b}$, is chosen such that it minimises the cross-validation score (sum of squared errors)

$$\text{CV} = \sum_{i=1}^{n} [y_i - \hat{y}_{\neq i}(b)]^2. \tag{5.10}$$

In Equation 5.10, $\hat{y}_{\neq i}$ is the prediction for location i by using bandwidth b and the parameters from Equation 5.9. The ith location is omitted from the estimation of y_i to prevent perfect estimation in the event of a reduced bandwidth, where observation i would receive the most weight for predicting itself. CV is the sum of squared prediction errors and $\hat{b}$ is the value that minimises this quantity (Brunsdon et al. 1996). The AIC approach iteratively finds the optimal bandwidth by minimising the AIC value.

There is a bias-variance trade-off when estimating the kernel bandwidth. A large bandwidth

results in a GWR model similar to the global model with greater bias due to the spatial drift, or heterogeneity, of parameters. A small bandwidth can result in an increased variance of local coefficients due to the reduced number of calibration points (Fotheringham et al. 1998).

Comber et al. (2023) outline an implementation approach for GWR modelling, this involves the determination of the most suitable modelling approach from either basic, mixed, or multi-scale GWR models. The basic GWR model has been described thus far, this assumes the same bandwidth for each parameter. Multi-scale GWR modelling drops the assumption that relationships operate over the same scale for each variable and determines a separate bandwidth for each parameter, allowing relationship scales to vary. The mixed GWR model contains parameters with either global effects or a common bandwidth. Comber *et al.* suggest initially implementing a multi-scale approach to determine the optimal bandwidth, or nearest neighbours, for each parameter. One can then decide upon the modelling approach by comparing these bandwidth values to each other and the total number of observations.

Variable Selection and Significance

Gollini et al. (2013) suggest a forward selection approach with an arbitrary bandwidth value and kernel function. This involves adding each variable based on the lowest AICc value until all independent variables are included. AICc is the corrected Akaike information criterion, where smaller sample sizes are considered for the AIC calculation. Using this model, the most suitable bandwidth and kernel are selected based on the AICc score.

For each of the n observations and p coefficients, a parameter estimate and corresponding standard error are obtained; hence, a local t-statistic can be used to explore the spatial relationships. This is not a preferred inferential statistic due to multiple testing with potentially $n \times p$ hypotheses. Additionally, the overlapping sets of calibration data and the use of such data for bandwidth selection is a concern for independent testing. Following such an approach, da Silva & Fotheringham (2016) outline an increase in type I errors or the number of false positives. The need for correction of the decision rule, relating to the critical value of the confidence level in the conventional t-test, is outlined. Common approaches are the Bonferroni Correction and the Byrne, Charleton, and Fotheringham (BCF) correction which is based on the Benjamini & Hochberg (1995) false discovery rate (Byrne et al. 2009). BCF contains the risk of increased type I error without sacrificing the statistical power of tests to detect significance, which is commonly an undesirable outcome of such corrections. Fotheringham et al. (1998) do not employ any correction and utilise the t-statistics to spatially explore the variation in the strength of each relationship. In a similar manner, Harris, Fotheringham, Crespo & Charlton (2010) use the GWR model as an exploratory non-inferential tool, by visualising the t-values within ± 1.96 (the critical values of a 95% confidence interval) to understand the strength of the relationship between each independent variable with the dependent variable at different locations.

5.2.1 Model Diagnostics and Fit

In a suggested route map for GWR modelling, Comber et al. (2023) outline the main areas of concern for the model, namely collinearity, influential outliers, and dependence in the error term. Such occurrences could lead to localised changes in the sign and magnitude of coefficient estimates which potentially compromise significance testing and bandwidth estimation. It is recommended to first explore these areas of concern using a global OLS model, and then use the GWR model results to explore the independence of residuals.

Due to the possibility of small sample sizes in the model calibration and variation in the correlation structures of the data, there are many concerns with collinearity in the GWR model (Gollini et al. 2013). However, unlike simulated data, spatial data has observations made in situ, often leading to unavoidable collinearity. Harris, Fotheringham & Juggins (2010) outline the assumption that collinearity between predictor variables has a negligible effect on model results. Fotheringham & Oshan (2016) address collinearity concerns by concluding that multicollinearity among the parameters of the GWR model should be treated the same as if it were a MLR model.

Fotheringham & Oshan (2016) suggest the combined use of local condition numbers and the variance inflation factor (VIF) of each coefficient to investigate collinearity, which can be localised to detect the areas of concern. The variance inflation factor is a measure of multicollinearity, measuring how much standard error is inflated by the existence of correlation among variables. In a global model, the condition number is the ratio of the largest to the smallest eigenvalue of the matrix $\mathbf{X}^T\mathbf{X}$. For GWR modelling, the local condition number of observation i is assessed using the matrix $\mathbf{X}^T\mathbf{W}_i\mathbf{X}$, where $\mathbf{W}_i$ is the unique weight matrix. The condition number is very large when the smallest eigenvalues are close to 0, this is due to strong collinearity within the design matrix $\mathbf{X}$. Myers & Myers (1990), Belsley et al. (2005) and Fotheringham & Oshan (2016) suggest using a threshold of 30 for the local condition numbers, and investigating any coefficients with a VIF>10 as unstable estimates.

A robust GWR model is explored by Harris, Fotheringham & Juggins (2010) to account for outliers in the model calibration by removing observations with a residual value over a cut-off value. The possibility of matrix instability due to local collinearity for parameter estimation using a discontinuous weighting kernel, such as the boxcar function, is identified.

Local R^2 values are obtained for each observation point, where the accuracy at which the model replicates the observed value at each point of the model calibration can be observed (Fotheringham et al. 1998). When the regression points are the same as the calibration points, the fitted values have a corresponding true value and the residual plot can be explored to further understand model fit (Charlton et al. 2009).

Spatial Heterogeneity

Testing for spatial variation of model coefficients is a suitable method of comparing the global and local modelling methods to describe the data (Fotheringham et al. 1998). The intercept and parameter estimates can be plotted, ceteris paribus (with all else equal), to observe changes in magnitude and sign throughout the space. The variability of parameter estimates is calculated using the estimated values from n observations, this can be used to check for non-stationary relationships, where one can imagine the parameter values form "valleys and hills" across the geographical space (Brunsdon et al. 1996). Under the hypothesis of stationarity, any permutation of the coordinate pairs is equally likely to occur. A Monte Carlo simulation is used to fit the GWR model to data from 99 random permutations of the coordinates. The observed variance for each parameter is compared to that of the random permutations, forming the basis of a significance test. If the true variance of the coefficient does not fall outside of the upper 5% tail of the ranked results of the Monte Carlo simulation, the null hypothesis of stationarity is accepted at the 5% level of significance (Lu et al. 2014).

5.3 Model Selection and Testing

5.3.1 Variable Selection Method

Common methods of variable selection include forward and backward selection (James et al. 2023, p. 79). Hurley & Sweeney (2022) follow the approach of Gelman & Hill (2006, p. 43) by including all relevant variables with the expected sign and magnitude. Similarly, Adair et al. (1996) include variables based on an a priori economic reasoning rather than statistical significance. Principal component analysis (PCA) is commonly used for dimensionality reduction of many correlated predictor variables to an orthogonal set. Panduro & Thorsen (2014) compares a mixed selection approach to the PCA for a hedonic house price model with 86 predictor variables, and concludes the PCA results in a loss of valuable information compared to the alternative.

A mixed selection method, combining both forward and backwards selection, is used for the MLR models presented in this study. This method involves iteratively adding single variables which would result in the lowest residual mean squared error (RMSE) while checking the results of t-tests and removing insignificant variables by order of the highest p value. A forward selection approach is used for the GWR models since the significance levels of each variable are non-inferential. In favour of parsimony, the model is selected by minimising the number of variables using the AICc score.

A function is created to report metrics in terms of price, which is more interpretable than the log-transformed alternative. These include the root mean squared error in euros, adjusted R^2, Alkaline information criterion (AIC), and statistics on model prediction accuracy. The purpose of each statistic and testing criterion is outlined in the following subsections.

5.3.2 Model Comparison

Following the approach of Fox & Weisberg (2011), F tests are used to calculate the significance of categorical variables with more than 3 levels and interactions with other variables. This is calculated using an ANOVA type III sum of squares which tests each term in the model when it is added to the reduced model of all remaining variables. The null hypothesis, which assumes the reduced and full models equally fit the data, is tested against the alternative hypothesis, stating that the full model is superior; thus, the individual contribution of each main effect and interaction term is tested. The Akaike information criterion (AIC) is used to compare model fit by maximising the likelihood function while accounting for parsimony by penalising the number of parameters in the model. For the geographically weighted regression model, a corrected Akaike information criterion (AICc) is used to compare models, this corrects for small sample sizes. An adjusted R^2 value is reported for each model as a measure of goodness of fit. R^2 is

the proportion of variability in the dependent variable explained by the model, the adjusted R^2 accounts for the number of variables in the model, and has the potential to decrease given a variable is added that does not contribute significantly to the model (James et al. 2023, p. 212).

5.3.3 Model Accuracy

The root mean squared error RMSE is commonly used as a measure of model accuracy. In the case of a natural log-transformed price variable, the RMSE in euros is calculated for cross-validation results using

$$\sqrt{\frac{\sum_{i=1}^{n}(e^{\hat{y}_i} - p_i)^2}{n}}$$

where p_i and $\hat{y}_i$ are the actual property price and the fitted log price for observation i, respectively. The median absolute error is the middle value of the percentage errors ordered by magnitude, where the percentage error of observation i is defined as

$$\frac{|p_i - e^{\hat{y}_i}|}{p_i} \times 100.$$

This is less sensitive to outliers compared to the mean absolute percentage error. The percentage of price observations within their corresponding 95% prediction interval is calculated. The prediction interval is the range of values that an individual observation is expected to fall in given what is observed, it is expected that 95% of the actual observations lie within this interval Harris, Fotheringham, Crespo & Charlton (2010). The proportion of predicted price values within 10% and 20% of actual price are also calculated.

5.3.4 Autocorrelation Test

Moran's I is the most common test for spatial autocorrelation, this is the Pearson's product moment correlation coefficient, which measures the correlation of an observation with itself through the weights of a distance-based weight matrix (Getis 2009, pp. 262-265). The matrix $\mathbf{W}$ contains entries $w_{ij} = d_{ij}^{-1}$, where d_{ij} is the Euclidean distance between observations i and j, the matrix has a zero diagonal and is row normalised such that all rows sum to one. The covariance matrix $\mathbf{Y}$, with autocovariance values $(y_i - \overline{y})(y_j - \overline{y})$ between observations i and j, is used to obtain a measure for Moran's I in Equation 5.11.

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \overline{y})(y_j - \overline{y})}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \quad i \neq j. \tag{5.11}$$

Moran's I has an expected value of $-1/(n-1)$. The values of I lie within [-1,1], where a positive value signifies similarity amongst neighbours, and a negative value indicates dispersion (Gaetan et al. 2010). Spatially independent observations will have a Moran's I value close to 0. The test for spatial autocorrelation is based on the assumption of randomness, where observations y are assumed to be realisations of a uniformly distributed Y variable, such that each observation is equally likely. Normality is also assumed, such that each observation is a random realisation of a different normal distribution. The following test is conducted:

- H_0 : Observations are random, there is no spatial autocorrelation.
- H_A : There is spatial autocorrelation between observations.

During the subsequent modelling applications, the autocorrelation of a vector of model residuals $\boldsymbol{\varepsilon}$ is tested using the alternative formula for Moran's I from Getis (2009)

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\boldsymbol{\varepsilon}^T \mathbf{W} \boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}}. \tag{5.12}$$

Geary's C is an alternative test for autocorrelation. While Moran's I considers the covariance structure, Geary's C considers the differences between neighbours with the expectation that neighbouring units do not differ from one another (Getis 2009, p. 265). In this case, there is no consistency in the magnitude of differences between observations.

5.4 Data Preparation

Relevant data preparations are performed to satisfy modelling assumptions and improve the interpretability of coefficients, these include dummy variable encoding and transformation of the dependent variable. Additionally, the number of bedrooms and bathrooms are centred at their median values of 3 and 2, respectively. This allows the regression intercept to be interpreted in terms of a 3-bedroom house with 2 bathrooms, which represents 14% of the overall data. Finally, the data is proportioned into folds for cross-validation testing.

5.4.1 Dummy Variable Encoding

Categorical variables must be represented as factors with respective numerical variables to be included in the regression models. Sum-to-zero contrast encoding is used for categorical variables in the Limerick property dataset. Following the methods of van den Berg (2021, ch. 10), categorical variables with levels representing groups of observations are considered and used to compare the contrasts. The resulting regression models corresponding to dummy encoding and sum-to-zero encoding using the municipality variable are demonstrated as an example.

Treatment Contrasts

Dummy encoding is used to represent K levels of a categorical variable with K-1 binary variables, with the remaining group serving as a reference level (James et al. 2023, p. 86). Treatment contrasts are made from these encodings, where the mean of each group is compared to the mean of the reference or "treatment" group. The variable 'Municipality' has four levels NS, AR, CK and LCS. In Equation 5.13, the coding scheme is represented in $\mathbf{S}$, and $\mathbf{L}$ is the contrast matrix.

$$\mathbf{S} = \begin{array}{c} \\ AR \\ CK \\ LCS \\ NC \end{array} \begin{array}{c} \begin{array}{cccc} & D_1 & D_2 & D_3 \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} \end{array}, \quad \mathbf{L} = \begin{array}{c} \\ L_1 \\ L_2 \\ L3 \\ L4 \end{array} \begin{array}{c} \begin{array}{cccc} \overline{AR} & \overline{CK} & \overline{LCS} & \overline{NC} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix} \end{array} \tag{5.13}$$

The first column of $\mathbf{S}$ represents the intercept, and the remaining three columns are the dummy variables, the rows are the four levels of municipality. The municipality AR is represented by the dummy variable D_1. NC has no corresponding dummy variable, it corresponds to the zero values of D_1, D_2, and D_3, and thus, is represented by the intercept. The contrast matrix $\mathbf{L}$ is such that $\mathbf{L} = \mathbf{S}^{-1}$. The row sums of $\mathbf{L}$ represent contrasts and the columns are the means of each group, the first two contrasts are presented below.

$$L_1 = \overline{NC}$$
$$L_2 = \overline{AR} \text{ - } \overline{NC}$$

The first parameter (or contrast) is the difference between the mean of NC and 0, representing the intercept, the second parameter contrasts the mean of AR and the reference level, NC. Similarly, the remaining rows of $\mathbf{L}$ contrast CK and LCS with NC. These contrasts demonstrate the popular dummy encoding, where each level is compared to the reference group.

Effects Encoding

Using sum-to-zero encoding for dummy variables results in more interpretable and relevant model output, more specifically, sum contrasts are used to compare the mean of each group to the "grand mean", or the mean of all group means. Consider the example in Equation 5.13, the matrices $\mathbf{S}$, and $\mathbf{L}$ become

$$\mathbf{S} = \begin{array}{c} \\ AR \\ CK \\ LCS \\ NC \end{array} \begin{array}{c} \begin{array}{cccc} & D_1 & D_2 & D_3 \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \end{pmatrix} \end{array}, \quad \mathbf{L} = \begin{array}{c} \\ L_1 \\ L_2 \\ L3 \\ L4 \end{array} \begin{array}{c} \begin{array}{cccc} \overline{AR} & \overline{CK} & \overline{LCS} & \overline{NC} \end{array} \\ \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 3/4 & -1/4 & -1/4 & -1/4 \\ -1/4 & 3/4 & -1/4 & -1/4 \\ -1/4 & -1/4 & 3/4 & -1/4 \end{pmatrix} \end{array}, \tag{5.14}$$

where the rows and columns of **S** represent the same levels and variables as the dummy encoding. Consider the first two contrasts from matrix **L**,

$$L_1 = \tfrac{1}{4}(\overline{AR} + \overline{CK} + \overline{LCS} + \overline{NC})$$
$$L_2 = \overline{AR} - \tfrac{1}{4}(\overline{AR} + \overline{CK} + \overline{LCS} + \overline{NC}).$$

The first row of **L** represents the intercept, which is the grand mean. The second parameter is a contrast between the mean of AR and the grand mean. Similarly, the third and fourth rows are the deviations of CK and LCS from the grand mean. The missing level, NC, can be obtained by subtracting L_1, L_2 and $L3$ from 0, since all parameters sum to 0.

Model Comparison

The interpretability of parameters based on the above methods of dummy encoding are compared using a linear regression equation with municipality levels from the Limerick property dataset. The dummy encoding and sum-to-zero encoding are represented in Equations 5.15 and 5.16, respectively.

$$\mathbb{E}[Price] = 141,484 + 50,560AR + 12,438CK + 68,682LCS \tag{5.15}$$

$$\mathbb{E}[Price] = 174,399 + 17,645AR - 20,477CK + 35,747LCS \tag{5.16}$$

Using Equation 5.15 the mean price of properties in AR is €50,560 + €141,484 = €192,044 and the difference in mean price between AR and NW is €50,560. In Equation 5.16, the coefficient of AR is the difference in mean of AR properties to the grand mean, where the mean of AR properties can similarly be obtained using €174,399 + €17,645 = €192,044.

The improved interpretability of sum encoding is apparent from the above example, the categorical variables are essentially centred around the mean; thus, each level can be compared to the grand mean.

5.4.2 Transformation of the Dependent Variable

The assumption of constant variance for the linear regression model is crucial for the calculation of confidence intervals and standard errors and the dependent variable is often transformed to resolve the issues of non-constant variance (James et al. 2023, pp. 95-96). Transforming the dependent variable y results in "shrinkage" of the larger responses, resulting in less skewness of the distribution and reduced heteroscedasticity of residuals. Concave functions such as $\log(y)$ and $\sqrt{y}$ are commonly used.

Following from empirical studies by Oust et al. (2020), Hurley & Sweeney (2022) and Liu (2013), it is expected that a natural log transformation would be suitable, resulting in a log-linear model; that is, modelling the log of price against a linear combination of the independent variables. The distribution of price is explored in Appendix B, it is heavily right-skewed due to outliers (Figure B.1A) and the log-transformed price in Figure B.1B demonstrates normality.

An additive linear regression model of price against all predictors is used to explore the residuals and test the linear regression assumption of constant variance, or homoscedasticity of residuals. The residuals are plotted against fitted values in Figure B.2 of Appendix B, and there is an observable funnel pattern (heteroscedasticity) of residuals. As in previous studies, the log-transformed price will henceforth be modelled as the dependent variable.

5.4.3 Out-of-Sample Testing and Validation

Common validation methods include the validation set approach, leave-one-out cross-validation (LOOCV) and K-fold cross-validation (James et al. 2023, pp. 203-206). Fik et al. (2003) and Bourassa et al. (2010) test model accuracy using out-of-sample test validation sets, while Hurley & Sweeney (2022) and Oust et al. (2020) use 5 and 10-fold cross-validation, respectively. The validation set is an out-of-sample proportion of the data used to test the model fitted on sample data, there is greater bias in the test results since the model is trained on a proportion of the data (often around 50%). LOOCV involves training the model on $n-1$ observations and testing on the nth observation, repeating this for all n observations, and averaging the error rate. There is reduced bias associated with the LOOCV, which uses almost all of the data for model training; however, there is an increased variance associated with the LOOCV compared to that of the validation set approach since many $n-1$ correlated error values are averaged (James et al. 2023, pp. 205-206).

K-fold cross-validation involves dividing the data into K folds, training on K-1 folds, and testing on the Kth fold, repeating this K times and averaging the error rate. The bias-variance trade-off outlined above is related to the choice of K, which can lead to LOOCV or the test validation set approach at the upper and lower extremes, respectively. K=10 is commonly suggested to be a suitable medium and thus, is applied to the Limerick property dataset, each fold has approximately 150 observations. Following the suggested approach of James et al. (2023, pp. 203-206), MLR models are initially explored using all observations and relevant predictors are identified. The mean metrics on model accuracy and goodness of fit described in Section 5.3 are reported from the 10 training results of the MLR models.

To align the GWR model analysis with that of the MLR model, a 10-fold cross-validation is conducted using the best model. For each iteration, 9 folds make up the calibration points while observations in the 10th fold are used as regression points for prediction. Using the

suggested approach of Brunsdon et al. (1996), bandwidth is optimised for the selected kernel at each regression point in the validation set. The kernel function is used to assign weights to the surrounding calibration points for parameter estimation, and thus, a fitted value and corresponding residual value are calculated.

Chapter 6

Application to the Limerick Property Dataset

The multiple linear regression and geographically weighted regression models are applied to the Limerick property dataset in Sections 6.2 and 6.3, respectively. Univariate models are first used to explore non-linear relationships in Section 6.1.

The variables used for the following models are selected from Table 4.1. The number of bedrooms and bathrooms are median-centred and the Cartesian coordinates are mean-centred to improve the interpretability of the intercept.

6.1 Polynomial Regression

A 10-fold cross-validation is performed to explore the polynomial degree of univariate regression models of $\log(Price)$ with the number of bedrooms and the number of bathrooms. The polynomial degree for each model is compared by analysing the mean squared error and estimated variance $\hat{\sigma}^2$. Following the approach of James et al. (2023), only polynomials with significant higher-order terms are explored.

Bathrooms and Price

The polynomial degree for the number of bathrooms, 'Baths', is explored and the greatest reduction in mean squared error and estimated variance $\hat{\sigma}^2$ is observed for the second-degree polynomial in Figure 6.1B. The improvements from degrees 2 to 7 are marginal compared to this. The second, third, and seventh-order polynomials are the only ones with significant higher-order terms. Comparing each of these polynomials fitted to the data in Figure 6.1A, the 7th order polynomial is too complex for the data. The 3rd order polynomial fits the data well, however,

rises at the 7th bedroom. Comparing this approach to the second-order polynomial and linear model, there appears to be an improved fit; hence, the third-order polynomial is chosen.

Bedrooms and Price

In a similar manner, the polynomial fit for 'Beds' is explored, the second and sixth-order polynomials are the only ones with significant terms of the highest order. The greatest improvement in the MSE arises for second-order polynomial, this is apparent from the scatter plot in Figure 6.1C, where the line captures the diminishing returns of additional bedrooms. For example, the added value of a sixth bedroom is much smaller than that of a second bedroom. The 6th degree polynomial has the lowest MSE as seen in Figure 6.1D, however, this model has an increased number of terms and there is an apparent risk of overfitting. The second-order polynomial is chosen in favour of a more parsimonious model.

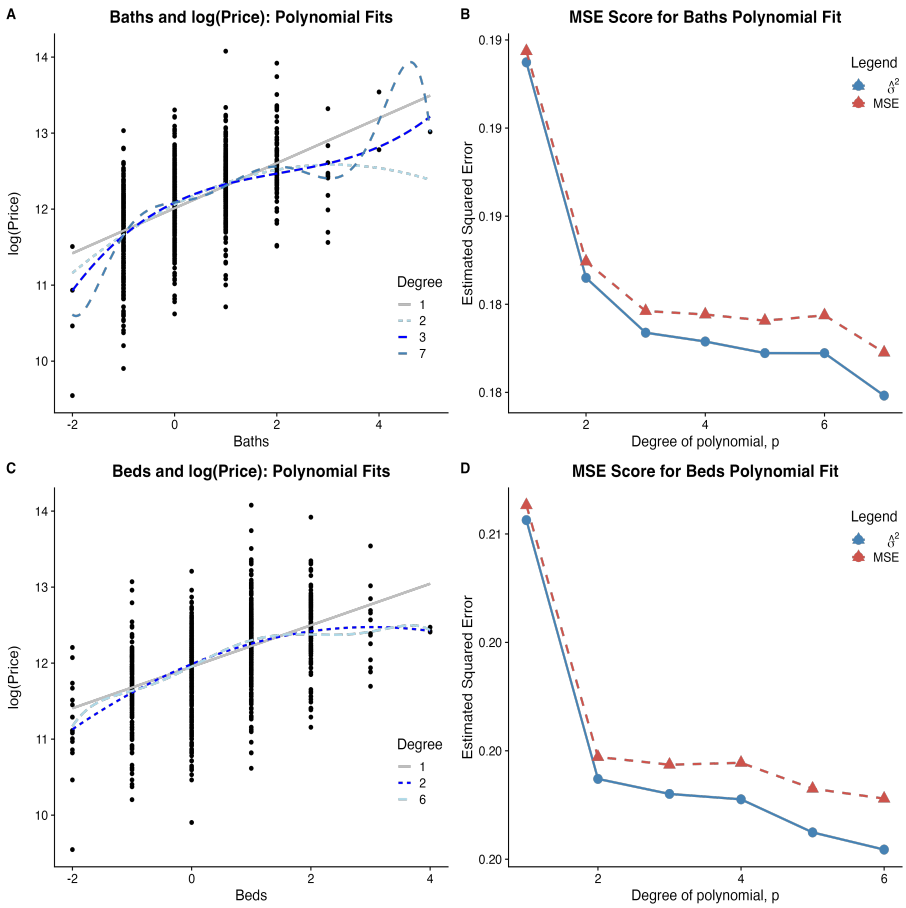

Figure 6.1: Polynomial degree selection for 'Beds' and 'Baths'.

6.2 Multiple Linear Regression

A selection of MLR models are fitted in the following section, the coefficients for each model are estimated using ordinary least squares as described in Section 5.1. The terms MLR and OLS are used interchangeably to refer to these models. The models and their descriptions are outlined in Table 6.1. The residuals of each model are found to have significant autocorrelation using the Moran's I hypothesis test from Section 5.3.4 at the 95% level of significance. This is indicative of model misspecification, where the model fails to represent the data, and is most likely due to a lack of spatial complexity in the models (Gaetan et al. 2010). Due to space constraints, the residual maps will not be included for every model, however, key insights will instead be described.

Table 6.1: OLS model names and descriptions.

Section	Model Name	Model Description
Basic Linear Regression Model	Basic OLS	Linear regression model with property and neighbourhood attributes and no spatial component
Location Dummy Variables	OLS-m	Basic OLS with municipality dummy variables
	OLS-u	Basic OLS with urban area dummy variables
Distance Measure	OLS-cbd	Basic OLS with distance to Limerick city centre
	OLS-cbd-u	OLS-cbd with urban area dummy variables
	OLS-cbdr	Basic OLS with distance ranges to Limerick city centre
	OLS-cbdr-u	OLS-cbdr with urban area dummy variables
Absolute Location	OLS-xy	Basic OLS with Cartesian coordinates
	OLS-xy-u	OLS-xy with urban area dummy variables

Multicollinearity & Influential Points

The generalised variance inflation factor (GVIF) is used to test collinearity between variables with a varying number of factor levels by adjusting for the degrees of freedom (Fox & Monette 1992). For each term, this is interpreted as the increase in the size of the ellipsoid of the coefficients of the predictor variables, or confidence ellipsoid, in comparison to that of uncorrelated regressors. The $2p^{th}$ root of the GVIF is taken to adjust for the dimensions of the confidence ellipsoid, where p is the number of levels of the term, and this is used to measure the inflation of the standard error. The values are initially obtained for an additive model of all independent variables and also throughout the following sections. In each case, the GVIF values do not exceed 5, the suggested threshold for further investigation (Fox & Monette 1992). Cook's distance (Cook's D) is used to identify influential points in the following regression models. Cook's D measures the effect of removing such influential points on the fitted values. Any values greater than 1 are investigated (Weisberg & Cook 1982).

6.2.1 Basic Linear Regression Model

Variable Selection

The third-degree polynomial terms for the number of bathrooms are selected as starting variables as they collectively have the lowest RMSE. The monthly dummy variables are subsequently included to account for the temporal effects, followed by variables representing property and neighbourhood attributes. The following insights have resulted from the mixed selection approach:

- The third-degree bath polynomial is insignificant following the addition of the property condition, suggesting that the non-linear relationship in the number of bathrooms is explained somewhat by the property condition. The 0-bathroom houses are expected to be sufficiently explained by the 'derelict' condition.
- The dummy variable representing a fireplace is found to be insignificant. The effect of a fireplace is possibly accounted for by the property type since many apartments or modular houses do not have fireplaces.

Following from the general MLR model definition in Equation 5.5, the model equation takes the form

$$\log(y_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{k=1}^{q} \gamma_k z_{ik} + \sum_{l=1}^{r} \delta_l t_{il} + \varepsilon_i \qquad i = 1, .., n \tag{6.1}$$

where y_i represents the price of property i, and ε_i is the iid residual value $\varepsilon_i \sim N(0, \sigma^2)$. Again, β_0 is the intercept term representing the mean $\log(Price)$. Continuous effects and dummy variables for observation i are grouped as $\mathbf{X}_i$ and $\mathbf{Z}_i$, respectively.

$$\begin{aligned} &\mathbf{X}_i = [\text{Beds}_i, \quad \text{Beds}_i^2, \quad \text{Baths}_i, \quad \text{Baths}_i^2], \\ &\mathbf{Z}_i = [\text{Property} - \text{Condition}_i, \quad \text{Property} - \text{Type}_i, \quad \text{Garage}_i, \quad \text{Period}_i, \quad \text{Cul} - \text{de} - \text{sac}_i]. \end{aligned} \tag{6.2}$$

The temporal effects for observation i are modelled as monthly dummy variables grouped as $\boldsymbol{T}_i$. For every observation i, the independent variables in $\boldsymbol{X}_i$, $\boldsymbol{Z}_i$ and $\boldsymbol{T}_i$ have corresponding regression coefficients $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ in vector form, respectively.

Model Residuals

The residuals of the basic OLS model are plotted for County Limerick and LCS in Figure 6.2, the residuals follow a pattern quite similar to that of price in Figure 4.4. This is expected, as the model contains no spatial component, and thus cannot account for the spatial effects on price. From the residuals in County Limerick in Figure 6.2A, most notably in NC, there is an

apparent hotspot of similarly priced properties in the middle of the municipality. Additionally, negative values indicate overestimated properties in the west. The residual pattern in AR is less uniform, there are many positive residuals in a belt surrounding LCS, indicative of under-estimated observations. Including the distance of the property from the city in the model may provide a solution to the residual patterns in these municipalities. The residuals in LCS are shown in Figure 6.2B. Many properties are underestimated in LCS-1, this area contains the affluent addresses of the Ennis Road and the North Circular Road. There are clusters of similar residual values in LCS-6, this aligns with the prevalence of housing estates in Dooradoyle. The overestimation of property prices in LCS-4 and the northern areas of LCS-2 complies with the presence of lower-priced properties in these areas (McCafferty 2009).

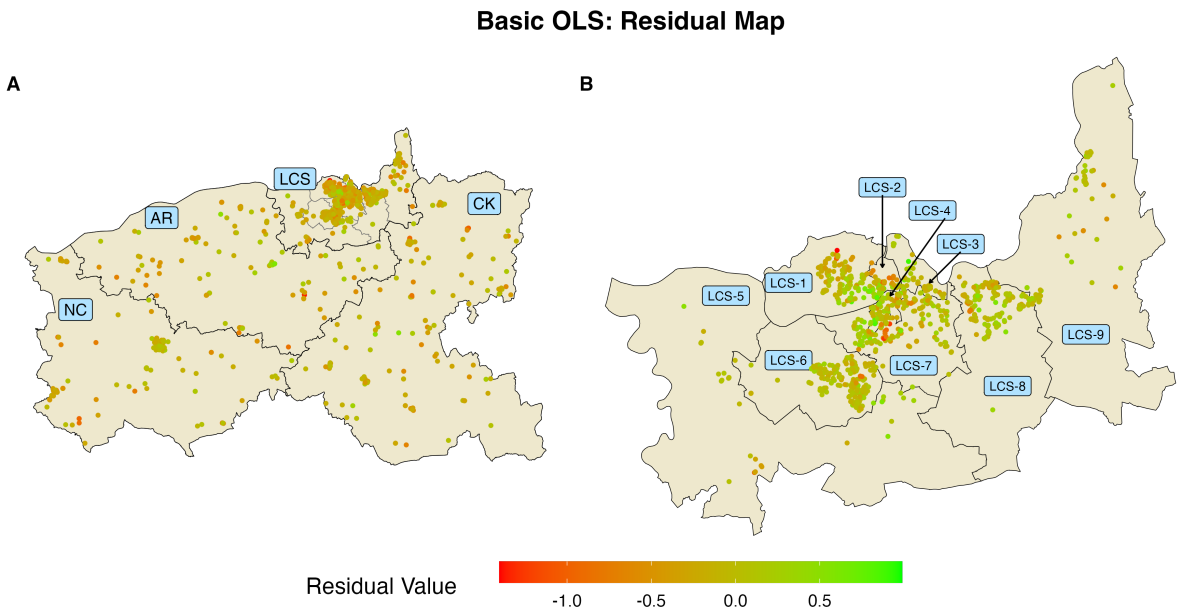

Figure 6.2: Map of the basic OLS model residual values in: A - County Limerick, and B - Limerick city and suburbs.

6.2.2 Location Dummy Variables

To account for differences between Limerick municipalities and urban areas, dummy variables representing these locations are added to the basic OLS model. The municipal dummy variables are added initially followed by the urban area dummy variables, which represent smaller boundaries within LCS. Each regional dummy variable corresponds to the average monetary effect of living in that specific region, this accounts for differences between regions, but not within regions, since the same average is applied to each property within that region (Bourassa et al. 2003).

The municipal dummy variables are added to the basic OLS model. Using a mixed selection method, the third-degree 'Baths' polynomial term is found to be significant. Adding the

municipal dummy variable likely introduces more flexibility for the model to capture spatial variations, thus allowing for varying non-linear relationships between the number of bathrooms and log($Price$) within each municipality. This is suggestive of local variation in the relationship between the number of bathrooms and property price. The 'Garage' variable is found to be insignificant, this is possibly a proxy for urbanisation or municipal differences, since the proportion of properties with garages is lowest in LCS, at 34% (Table A.2). The residual pattern is similar to that of the basic OLS model in Figure 6.2 with a reduced magnitude of residuals around the county border.

The urban area dummy variables consist of the three municipalities of Limerick county, along with the 9 urban areas of LCS (Figure A.2). There is a significant interaction between 'Urban Area' and 'Fireplace' using an ANOVA type III test, suggesting local variation of the effect of a fireplace on property price. This may account for the lower-priced apartments in the city centre with no fireplaces. For example, LCS-4 has the lowest median price of €132,000 and the lowest proportion of properties with fireplaces (31%, Table A.2). Similarly, there is a significant interaction between 'Urban Area' and 'Beds', which suggests the effect of an additional bedroom varies across urban areas. The above interactions are insignificant in the previous model with municipality dummy variables, this suggests that smaller submarket borders can be used to account for spatially varying relationships (Bourassa et al. 2010). The 'Cul-de-sac' dummy variable does not significantly add to the model and is removed. The proportion of cul-de-sacs in each urban area can be observed in Figure 4.9, it is likely that the dummy variables representing suburbs in south LCS, where cul-de-sacs are more popular, account for these effects.

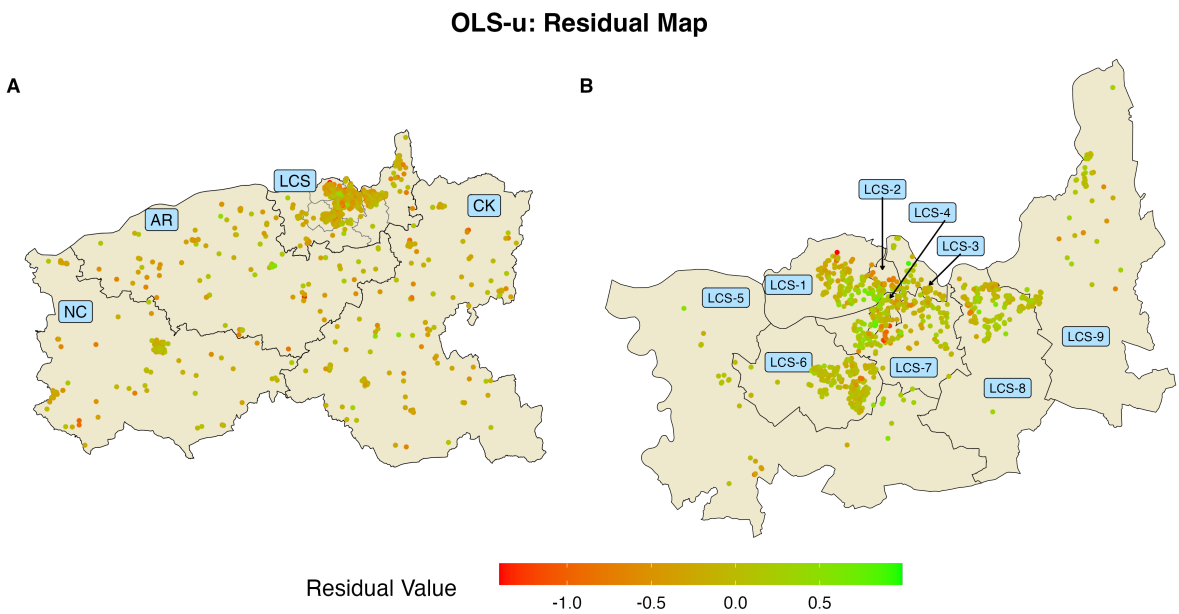

Figure 6.3: Map of residual values from the OLS-u model in: A - County Limerick, and B - Limerick city and suburbs.

The residuals of the urban area OLS model can be observed in Figure 6.3. In County Limerick (Figure 6.3A), there remain apparent clusters of similarly priced properties in NC and CK. Compared to the basic OLS model residuals in Figure 6.2, there is a reduced magnitude of residual values for observations along the south and west county border, as previously described for the municipal OLS model. In Figure 6.3B, the residual pattern in LCS is similar to the residual pattern from the basic OLS, with slightly reduced magnitude for overestimated values in the city centre areas of LCS-2 and LCS-4.

6.2.3 Distance Measure

To further account for the effect of location on property price, the distance from the city centre is included in the basic OLS model from Section 6.2.1. This is an attempt to decrease the appearance of a belt of similar residuals surrounding LCS in Figure 6.2A. Through mixed selection and the inclusion of significant interactions, the distance to the city centre ('CBD Distance' abbreviated to 'CBD') is included in the continuous covariates from Equation 6.2 to form

$$\begin{aligned}\mathbf{X}_i = [&\text{Beds}_i, \quad \text{Beds}_i^2, \quad \text{Baths}_i, \quad \text{Baths}_i^2, \quad \text{Baths}_i^3, \quad \text{CBD}_i, \quad \text{CBD}_i \cdot \text{Fireplace}_i, \\ &\text{CBD}_i \cdot \text{Period}_i], \\ \mathbf{Z}_i = [&\text{Property} - \text{Condition}_i, \quad \text{Property} - \text{Type}_i, \quad \text{Garage}_i, \quad \text{Period}_i, \quad \text{Cul} - \text{de} - \text{sac}_i, \\ &\text{Fireplace}_i]. \end{aligned} \tag{6.3}$$

where the interaction terms between 'CBD' and 'Fireplace' and 'Period' are included. The temporal effects remain as in the basic OLS model from Equation 6.1.

The interaction between 'Property Type' and 'CBD Distance' is significant using an ANOVA type III test at the 5% level of significance. This results in the same mean squared error as including the two interactions present in Equation 6.3; however, this interaction results in a highly influential point with a Cook's D of 1.126, which is greater than the suggested value of 1 for further investigation (Weisberg & Cook 1982). This observation, with a price of €110,000, is the only duplex property in Limerick county, more specifically in NC. It is likely that the relatively low property value in combination with the property type and distance measure introduces bias in the regression model. Adair et al. (1996) describe a similar situation when modelling property types across the urban areas of Belfast. There is little improvement in the residual pattern in comparison to the basic OLS model, dummy variables representing 'Urban Area' are subsequently included to account for differences in submarkets.

Using a mixed selection approach for the OLS-cbd model with 'Urban Area', 'Garage' remains insignificant. Similar to the OLS-u model, significant interactions between 'Urban Area' with 'Fireplace' and 'Beds' are included. It is likely that the distance component along with regional

differences, has created more flexibility for spatial variation of model coefficients, rendering the global effects of garage insignificant.

As an alternative to distance in kilometres, a categorical variable of distance ranges, 'CBD Distance Range', is included in the model. This variable categorises observations as within 5km, 5-10km and 10-20km of Limerick city centre, as well as properties outside of the 20km radius. In contrast to the OLS-cbd model, the OLS-cbdr model does not contain an interaction between 'CBD Distance Range' and 'Period', this could not be calculated due to lack of such properties within certain radii from the city centre. The OLS-cbdr-u model resembles the OLS-cbd-u model, where 'CBD Distance Range' is used as a replacement for 'CBD Distance'.

6.2.4 Absolute Location

Absolute location is incorporated in Equation 6.1 by including Cartesian coordinates in the continuous effects in Equation 6.4 below. Following a similar approach to Fik et al. (2003), a second-degree interactive polynomial of mean-centred Cartesian coordinates is fitted. Fik et al. (2003) outline an improvement in the interactive approach over the non-interactive model. These polynomial terms are included in the model along with the variables of interest from Table 4.1. Including absolute location in the model may account for omitted variables such as property age or distance to amenities, which are otherwise averaged out using dummy variables. The spatial polynomial is included as a base model to account for spatial effects, and a mixed selection model is used to include the subsequent property attributes, following which 'Garage' is found to be insignificant using an ANOVA type III test.

$$\begin{aligned}
\mathbf{X}_i &= [\text{Beds}_i, \quad \text{Beds}_i^2, \quad \text{Baths}_i, \quad \text{Baths}_i^2, \quad \text{Baths}_i^3, \quad x_i, \quad y_i, \quad x_i \cdot y_i, \quad x_i^2, \quad y_i^2] \\
\mathbf{Z}_i &= [\text{Property} - \text{Condition}_i, \quad \text{Property} - \text{Type}_i, \quad \text{Period}_i, \quad \text{Cul} - \text{de} - \text{sac}_i, \quad \text{Fireplace}_i]
\end{aligned} \tag{6.4}$$

The OLS-xy-u model is obtained by including the dummy variables representing urban areas among the categorical covariates, $\mathbf{Z}_i$, in Equation 6.4, this model aims to account for differences between the areas in Limerick. The mixed model selection results in similar significant variables to the OLS-cbd-u and OLS-cbdr-u models from Section 6.2.3, the interaction between 'Urban Area' and 'Fireplace' is found to be significant using an ANOVA type III test.

6.2.5 Results

The results of the 10-fold cross-validation of each OLS model are presented in Table 6.2. The basic OLS model from Section 6.2.1 has the worst model fit with a median absolute percentage error of over 20% and an adjusted R^2 of 0.53, meaning approximately 50% of the variation in the price variable is explained by the basic OLS model. Firstly, this non-spatial model is compared to the OLS-m, the OLS-cbd, OLS-cbdr, and the OLS-xy models from Sections 6.2.2, 6.2.3 and 6.2.4. The model fit improves to a similar level for each of these models with location attributes, and the increase in model fit from the OLS model to the OLS-m model aligns with the results of Bourassa et al. (2010), who note an increase in R^2 value from 0.70 to 0.75 upon the addition of submarket dummy variables. Among the five models, the OLS-cbdr has a slightly higher adjusted R^2 along with a lower AIC and reduced median APE. There are differences in the proportions of predictions within 10% and 20% of price, the OLS-cbdr has the greatest proportion out of these 5 models, suggesting improved predictive accuracy. The OLS-cbd has the greatest proportion of predictive intervals containing the correct value of price, suggesting wider intervals compared to the basic OLS model and the OLS-m which both have less than 94% of actual values lying within the predictive intervals.

Table 6.2: 10-fold cross-validation results of the OLS models.

Model	Adj. R^2	AIC	RMSE	Median APE %	Within 10% of Price	Within 20% of Price	Within 95% Prediction Interval	Moran's I
Basic OLS	0.53	1093	€73,665	21.57%	22.44%	47.22%	93.98%	0.203
OLS-m	0.63	767	€65,053	17.63%	28.57%	56.16%	93.98%	0.103
OLS-u	0.66	686	€64,574	16.72%	32.46%	57.30%	94.18%	0.095
OLS-cbd	0.63	786	€70,497	18.14%	28.98%	53.28%	94.78%	0.103
OLS-cbd-u	0.66	671	€63,927	16.75%	31.53%	57.50%	94.11%	0.090
OLS-cbdr	0.64	739	€65,313	17.02%	31.64%	57.76%	94.05%	0.096
OLS-cbdr-u	0.66	667	€63,663	16.31%	32.72%	57.37%	93.98%	0.081
OLS-xy	0.63	770	€65,577	18.35%	29.84%	54.43%	94.18%	0.116
OLS-xy-u	0.66	687	€63,310	16.99%	31.13%	57.51%	94.11%	0.091

10-fold cross-validation is performed on models to calculate the adjusted (Adj.) R squared, Akaike information criterion (AIC), root mean squared error (RMSE), median absolute percentage error (APE), the proportion of observations within 10% and 20% error intervals around actual price and within their 95% prediction intervals. Moran's I is specified for each model fitted to all observations.

Considering the Moran's I values for the basic model, a value of 0.203 is a highly significant observed value with a p value approximately equal to 0, there is sufficient evidence to reject H_0 at the 5% significance level and conclude that there is positive spatial autocorrelation among the model residuals. This is a similar result to Farber & Yeates (2006) and Basu & Thibodeau (1998), who report spatially biased basic linear regression models identified by spatial association

among residuals. Including municipal dummy variables results in a reduction of the spatial autocorrelation of residuals in this model, with a Moran's I of 0.103. Including the distance from Limerick city centre results in the same Moran's I as the OLS-m model, suggesting the same level of spatial autocorrelation among residuals. Compared to all previous spatial enhancements to the basic OLS, including the Cartesian coordinates results in the greatest degree of spatial autocorrelation among residuals, with a Moran's I of 0.116 suggesting a lack of spatial complexity compared to the other models. This is likely due to the relative simplicity of the equation compared to the expansion method of Brunsdon et al. (1998), aiming for interpretation and parsimony, complexity is minimised in the OLS-xy model.

Consider the OLS-u, OLS-cbd-u, OLS-cbdr-u and OLS-xy-u models, where dummy variables representing urban areas in LCS are added to the basic OLS model and 3 locational models. In the OLS-u model, there is a further reduction in the Moran's I value from 0.103 to 0.095 as the boundary size of regional areas decreases, this is a similar result to Bourassa et al. (2010) and Oust et al. (2020), who note improvements for smaller submarkets. Increasing the number of regions results in a further increase in adjusted R^2 and a reduction of prediction error to 16.72%, this contradicts the results of Dubin (1988), who identifies the lack of value from arbitrarily defined neighbourhood variables. By increasing the spatial complexity of the OLS-cbd model to include urban areas, there is a reduction in the degree of spatially autocorrelated errors, with a Moran's I of 0.09. This is slightly lower than the OLS-u model and is a similar result to that of Hurley & Sweeney (2022), who note an improvement in model fit when including CBD distance and dummy variables compared to the basic OLS. With an increased proportion of estimates within 10% and 20% of price, the OLS-cbd-u model has an increased predictive accuracy compared to the OLS-cbd model. There is a reduction in the number of observations within their 95% predictive intervals for the OLS-cbd-u model, suggesting reduced uncertainty.

The OLS-cbdr-u model has further increased accuracy compared to the OLS-cbd-u model, with an AIC of 667 and RMSE of €63,663. The AIC values of these models differ by over 2 units suggesting the OLS-cbdr-u model is more suitable (Gelman & Hill 2006, p. 525). The OLS-cbdr-u model has the largest proportion of predictions within 10% of actual price, and the lowest Moran's I value of all models.

Increasing the spatial complexity by including urban area dummy variables in the OLS-xy model results in a similar improvement to the OLS-cbd-u and OLS-cbdr-u models in reducing the spatial bias of the OLS-xy model. The Moran's I value of 0.091 is similar to that of the OLS-cbd-u model and is slightly lower than the OLS-u model. In a similar manner to Fik et al. (2003), there is an improved model fit compared to the basic OLS model when absolute location is included. Including urban area dummy variables with the absolute location results in a slightly higher median absolute percentage error compared to the OLS-u and OLS-cbd-u models, suggesting

the model does not fit the data as well, and prediction errors compensate for this. There is an improved model fit when urban area dummy variables are included in locational models, contradicting the results of Heyman & Sommervoll (2019), who note diminished improvements to locational models following the addition of submarket dummy variables.

Accounting for location in the basic OLS model results in a reduction of the spatial dependence among residuals, and an increase in model fit and predictive accuracy. Including the 'Urban Area' variable in the basic OLS, OLS-cbd, OLS-cbdr and OLS-xy models results in an increase in model fit compared to such initial models. This highlights the importance of locational dummy variables representing submarkets and aligns with the results from empirical studies by Basu & Thibodeau (1998) and Fik et al. (2003). The OLS-xy-u and OLS-cbd-u models have similar values for predictive accuracy, spatial autocorrelation of residuals, and model fit. The OLS-cbdr-u model has slightly reduced spatial association among residuals and increased predictive accuracy. The OLS-cbdr-u model is selected as the best multiple linear regression model due to the reduced prediction error and AIC value. Additionally, the distance measure is more interpretable and applicable compared to Cartesian coordinates. The model results of the OLS-cbdr-u model are presented in full in Table C.1 of Appendix C and are interpreted in the following section.

6.2.6 Model Interpretation

Residuals and Model Fit

A histogram of the OLS-cbdr-u model residuals is shown in Figure 6.4B. There is apparent negative skewness where large negative residuals are present, corresponding to the overestimation of property prices. A Kolmogorov-Smirnov test results in a p value < 0.05 which is sufficient evidence to reject the null hypothesis that the data follows a normal distribution. Although the normality of residuals is not statistically significant, the residuals appear to be normally distributed with a mean of 0. The residuals are plotted against the corresponding fitted values in Figure 6.4A and appear to have constant variance and show no discernible pattern. Following the approach of James et al. (2023), the residual assumptions of the OLS model are satisfied.

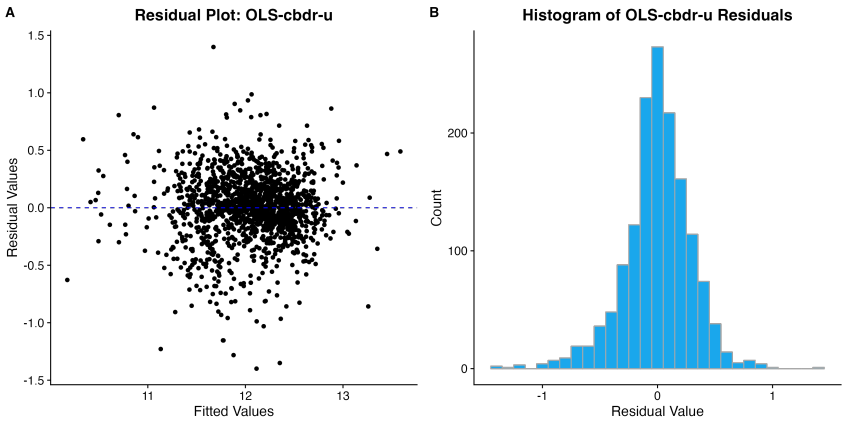

Figure 6.4: OLS-cbdr-u model residuals: A - and fitted values, B - histogram plot.

The residuals of the OLS-cbdr-u model are plotted for Limerick and LCS in Figure 6.5. The spatial pattern is similar to that of the dependent variable (Figure 4.4) suggesting that the model has not effectively accounted for the spatial variation of property price. Compared to the residual pattern of the basic OLS model (Figure 6.2), there is an apparent reduction in the magnitude of observations surrounding LCS and the outer county border.

Contrast Encoding

Through the use of contrast encoding, the coefficients representing the levels of categorical variables are a comparison of each level to the grand mean. In Table C.1, the exponentiated model coefficients are listed, these are the multiplicative effects of each contrast on property price. An exponentiated coefficient greater than 1 relates to an increase in property value from the grand mean, while a value less than 1 signifies a relative decrease. For this reason, a dashed line is plotted at $y = 1$ along with the exponentiated coefficients in Figure 6.6. The

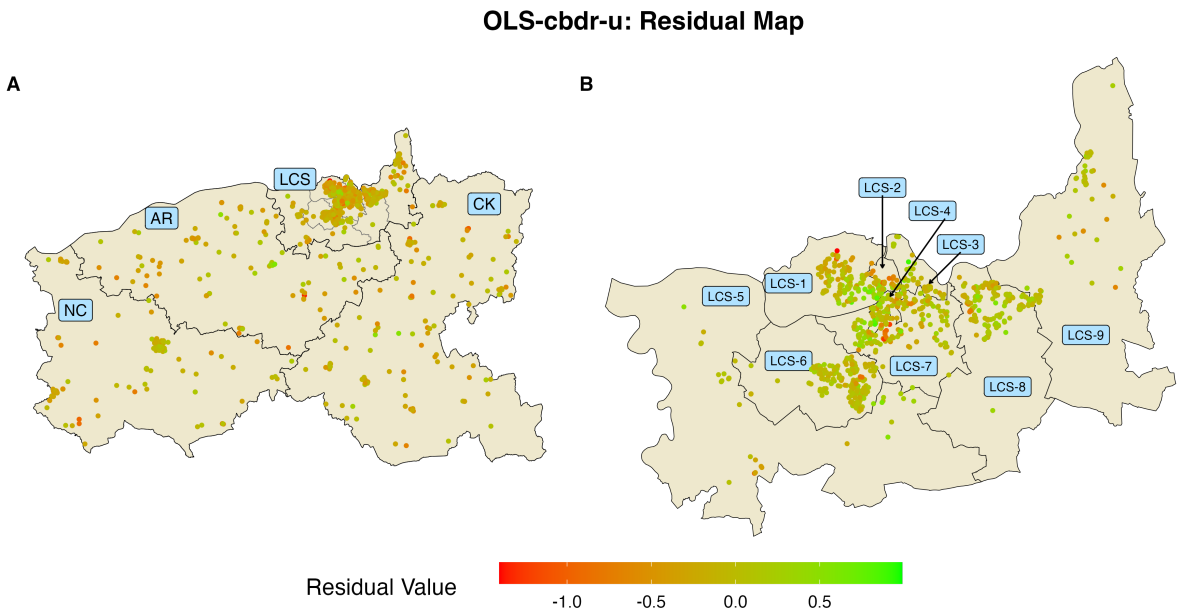

Figure 6.5: Map of residual values from the OLS-cbdr-u in: A - County Limerick, and B - Limerick city and suburbs.

relative scalings of property types in Figure 6.6 demonstrate the increased value of a detached property, where property price increases by 40% on average compared to the grand mean of price across property types, ceteris paribus. A semi-detached property type has a positive effect on property value, while apartment, duplex, and terraced properties have a negative effect on value compared to the grand mean. With an $(1 - 0.87) \times 100 = 13\%$ average decrease for apartments, holding all else constant. Using the approach of Hurley & Sweeney (2022), the relative scalings of coefficients are utilised to compare property types and conditions. The average relative increase in price for end-of-terrace properties over terraced properties can be calculated as $(1-0.90/0.97)\times 100 = 7.22\%$, similarly, for semi-detached and detached properties the average relative increase in price is $(1 - 1.12/1.40) \times 100 = 20\%$, this large increase is likely capturing other effects in the data, such as the differences in site size between a semi-detached property in LCS and a detached home in the countryside. In Figure 6.6, the confidence intervals of the effect of a townhouse and end-of-terrace property each intersect the horizontal unitary line, meaning the intervals contain the value 1, this suggests that no change in price compared to the grand mean is supported.

Similarly, the effects of property condition on the grand mean of property price display an expected pattern in Figure 6.6, there is a clear distinction between the positive effects of new and second-hand properties, and the negative effects of derelict properties and those in need of renovation. A new property condition has the greatest effect on price, with a 95% confidence interval of [1.39, 1.64] and a mean value of 1.51, suggesting a mean increase of 51% from the grand mean across property condition, ceteris paribus. The decrease in property value

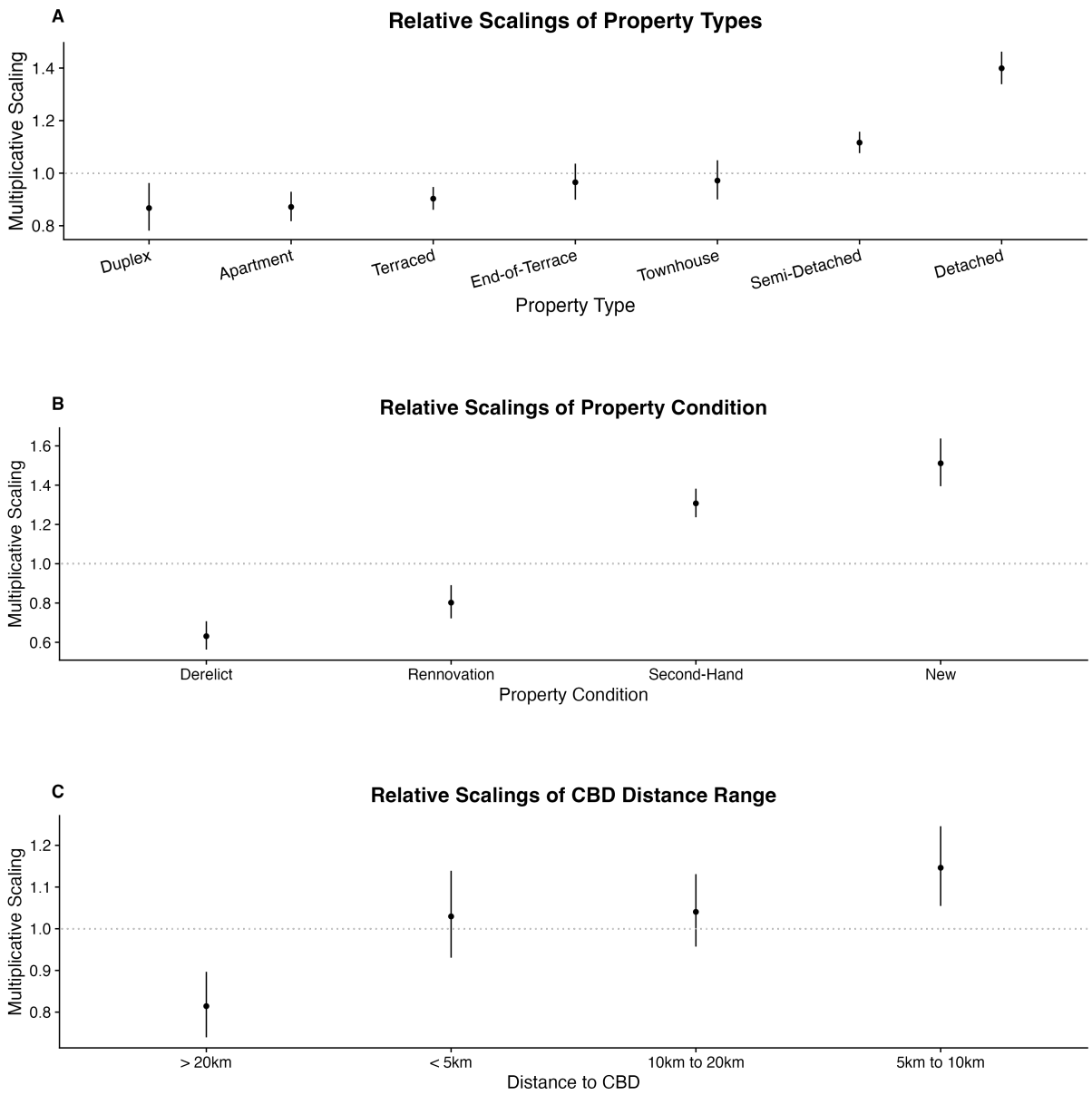

Figure 6.6: Relative scalings of 'Property Type', 'Property Condition' and 'CBD Distance Range' coefficients from the OLS-cbdr-u model.

for properties with a derelict condition is between [1-0.56,1-0.71] = [0.29, 0.44] with a mean decrease of 0.37 or 37% in property price from the grand mean, holding all else constant. The relative increase in property price for new properties over second-hand properties is calculated as $(1 - 1.31/1.51) \times 100 = 13.2\%$. The large gap between second-hand properties and properties in need of renovation demonstrates the concerns from Section 3.3, where a lack of well-defined property conditions is outlined leading to a generalisation of many property conditions as second-hand.

The relative scalings of CBD distance ranges display an expected pattern, similar to that of Figure 4.5B. Properties within 5km and 10-20km of Limerick city centre have a property price approximately equal to the grand mean across CBD distance ranges. The greatest difference is observed between properties within 5-10km of the CBD and properties over 20km from the CBD. Properties within 5-10km of Limerick city centre are associated with an average price increase from the grand mean of 15% with a confidence interval of [5%, 25%], with all else being equal. The average relative increase in price for properties within 5-10km of the CBD over properties within 5km of the CBD is $(1 - 1.03/1.15) \times 100 = 10.4\%$.

Dummy Variable Coefficients

The dummy variables from $\mathbf{Z}_i$ in Equation 6.3 are presented in Table C.1 with the corresponding 95% confidence intervals. The exponentiated coefficient for 'Period' has a 95% confidence interval of [1.67, 2.21] and a mean value of 1.92, supporting an average increase of 92% in price for period properties compared to those that are not, holding all else constant. This large effect is likely capturing the effects of affluent addresses in the city, where period properties are prominent, along with large land holdings for country houses that are not accounted for in the model.

Polynomial Coefficients

The 'Baths' polynomial holds different interpretations such that the effect of an additional bathroom on price varies based on the initial number of bathrooms. There is an average increase of 29% in property price for the addition of a second bedroom, holding all else constant, with a surrounding confidence interval of [28.3%, 30.2%]. This value falls to 9% for the addition of a 5th bathroom, capturing the effects of diminishing returns as seen in Figure 6.1. However, there is unusual behaviour at the upper scale, where the addition of a 7th bathroom is associated with a 40% average increase in price holding all else constant, the surrounding confidence interval of this estimation is [-28%, 289%], suggesting great uncertainty. This is likely a consequence of the data quality, where there is a lack of observations with a greater number of bathrooms and generally many outliers in price for a greater number of bathrooms.

Interaction Terms

There is a significant interaction between 'Beds' and 'Urban Area', suggesting variation in the effect of a bedroom across space. Separate confidence intervals for the effect of an additional bedroom in different urban areas are calculated, holding all else constant and the most notable results are given. Holding all else constant, the multiplicative effect of an 4th bedroom in the City South (LCS-4) lies between [0.88, 1.0.8] with an average value of 0.98, these values are [1.06, 1.31] and 1.18 for the Outer City South (LCS-7), respectively. For Newcastle West (NW), these values are [1.04, 1.27] and 1.15, respectively. There are differences between the average

effect of a 4th bedroom on property price across these regions, notably, there is support for a negative effect in the inner south of the city, a positive effect in the outer city south, and a reduced positive effect in the municipality of Newcastle West. The confidence interval for LCS-4 includes 1, suggesting no change from the grand mean. This is possibly due to the lack of properties with 3 or 4 bedrooms in this area, where the median number of bedrooms is 2. In comparison, the average multiplicative effect of a 2nd bedroom on property price in LCS-4 is entirely positive, with a confidence interval of [1.04, 1.13] and an average value of 1.08, holding all else constant. The reduced effect of a 4th bedroom on property price in Newcastle West compared to the Outer City South aligns with the price differences between the areas as seen in Table 4.2.

There is a significant interaction between 'Fireplace' and 'Urban Area', suggestive of spatial variation in the effect of the property feature on price. The greatest differences are between City North (LCS-2) and North (LCS-1). The multiplicative effect on property value for the addition of a fireplace in LCS-2 is between 0.87 and 1.19 with an average value of 1.02, this suggests the addition of a fireplace has a negligible effect on property value in LCS-2. One reason for this may be the presence of lower-priced houses in the area, with a median property price of €180,000. The interaction term may be acting as a proxy for social and economic effects such as the presence of social housing. On the other hand, LCS-1 has a median property price of €225,000, and the multiplicative effect of a fireplace on price has a confidence interval of [1.06, 1.42] and an average value of 1.22, reflective of a 22% average increase in property price for the addition of a fireplace in LCS-1, holding all else constant. Both of these confidence intervals are wide, suggesting uncertainty of parameter estimation.

6.3 Geographically Weighted Regression

6.3.1 Model Selection

The package *GWmodel* is used to implement the following GWR models and graphically represent the results. The variables of interest for the following GWR models include all of those from Table 4.1, with the use of a 'Year' dummy variables to account for annual differences rather than 'Month' to encourage parsimony in the local model estimation. Additionally, the number of properties from certain months within each local estimation kernel is expected to be low compared to that of the global estimation, which could contribute to bias. No locational variables are included in the models because the GWR model operates at a local geographic scale. Dummy variables representing areas or municipalities are not included as there is no expected dichotomy of property prices at the regional borders.

Comber et al. (2023) recommend initially exploring the model diagnostics with a global MLR model. An OLS regression model, henceforth referred to as the global model, is fitted with additive terms of interest, similar to the basic OLS model in Section 6.2.1. The maximum GVIF statistic is 1.5, highlighting the existence of no concerning multicollinearity. Using a t-test, the effects of a fireplace and garage are found to be insignificant at the 5% level of significance, with p values of 0.07 and 0.19, respectively.

All of the variables of interest are used in the forward selection approach for the GWR model to determine the effect of adding each variable on the corrected AIC (AICc). A model of all variables (GWR-1) and a parsimonious model based on the variable selection results (GWR-2) are chosen. Where GWR-2 contains 'Beds', 'Baths', 'Property Type', and 'Property Condition'.

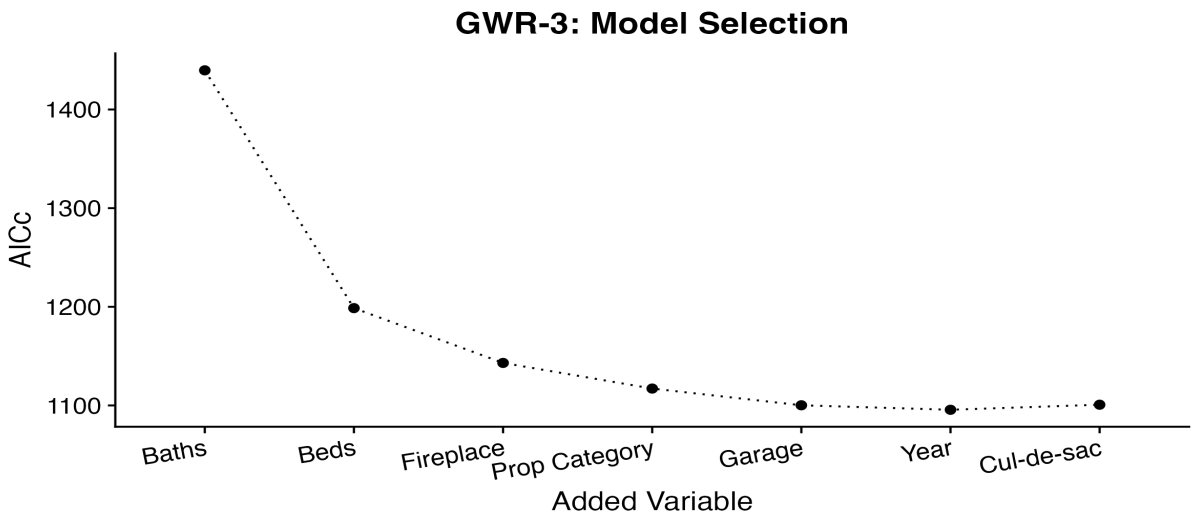

Figure 6.7: Model selection AICc results for the GWR-3 model.

It is expected that there may be relatively few observations for certain property types or conditions among the sparsely distributed areas in the county; hence, for a third model selection process, 'Property Type' and 'Property Condition' are replaced with 'Property Category', which classifies entries as either houses or apartments. The results of this model selection are shown in Figure 6.7, there are diminishing improvements to the AICc following the addition of 'Beds'. Although the "elbow", representing a low AICc and parsimony, is clearly at 'Beds', 'Fireplace' is included in this model. Throughout Section 6.2, the 'Fireplace' variable has demonstrated importance, having a significant global effect on property price and demonstrating significant interaction effects with location variables, the spatial variation of this coefficient is explored with the GWR model.

Table 6.3: Model calibration results for the kernel functions of the GWR-1, GWR-2 and GWR-3 models.

Model	Kernel	Bandwidth	Adjusted R^2	AICc
GWR-1	boxcar	1037	0.54	1132
	bi-square	1037	0.62	887
	Gaussian	271	0.58	1004
	exponential	87	0.65	810
GWR-2	boxcar	669	0.50	1269
	bi-square	669	0.60	958
	Gaussian	104	0.60	966
	exponential	30	0.66	823
GWR-3	boxcar	36	0.65	946
	bi-square	73	0.68	826
	Gaussian	18	0.64	942
	exponential	18	0.61	1024

6.3.2 Model Calibration

As outlined in Section 5.2, the kernel function and bandwidth must be chosen and calibrated at each regression point. The box-car kernel function is discontinuous whereas the bi-square, Gaussian and exponential kernels are continuous functions of distance. For each kernel, an adaptive bandwidth, representing the number of nearest neighbours used to calibrate the model, is optimised. A comparison of kernel estimations for the three models is displayed in Table 6.3, where the greatest adjusted R^2 values for the GWR-1 and GWR-2 models are obtained using the exponential kernel. However, the number of local observations in the respective kernels are approximately 5% and 2% of the overall data, this could cause more pronounced collinearity effects (Fotheringham & Oshan 2016). Additionally, small bandwidths could lead to overfitting the local variations in the price variable and the suggestion of geographical patterns when none exists (Comber et al. 2023). The bandwidth for all kernel functions of the GWR-3 model is considerably lower than that of GWR-1 and GWR-2 models, this is likely due to the reduced number of variables in the model, thus, requiring less data for parameter estimation. The bi-

square kernel is used for each approach, this selection is based on the number of observations included in the model calibration and the reduced AICc value compared to the other functions. This function uses a distance decay weighting for the 1037, 669 and 73 nearest neighbours for the GWR-1, GWR-2 and GWR-3 models, respectively. All other observations receive a weight of 0.

6.3.3 Model Diagnostics

For the GWR-1 model, the exponential and bi-square kernels have 768 and 71 observations with local condition numbers over 30, respectively. The increased local sample size for the bi-square kernel is likely the cause of this reduction in collinearity, as there is a negative relationship between the bandwidth and condition number (Fotheringham & Oshan 2016). In Table 6.4 the number of VIF values over 10 and condition numbers over 30 are listed for various bandwidth specifications of the bi-square kernel of the GWR-2 model, this is suggested as a potential method by Fotheringham & Oshan (2016) for identifying problematic variables and exploring possible bandwidth values. The number of local observations required for model calibration is similar to that of a global model once the diagnostic values are satisfactory. Additionally, 'Property Type' and 'Property Category' are identified as the problematic variables.

Table 6.4: The number of condition numbers over 30 and parameter VIFs over 10 for various bandwidth values for the GWR-2 model with a bi-square kernel function.

BW	CN > 30	Variance Inflation Factor > 10										
		Baths	PC.1	PC.2	PC.3	PT.1	PT.2	PT.3	PT.4	PT.5	PT.6	Year
669 (45%)	591	0	86	139	43	0	29	70	0	14	1	0
700 (47%)	566	0	40	26	30	0	21	67	0	4	0	0
800 (53%)	465	0	8	4	6	0	10	53	0	0	0	0
1,000 (67%)	59	0	0	0	0	0	0	18	0	0	0	0
1,200 (80%)	7	0	0	0	0	0	0	0	0	0	0	0
1,400 (94%)	0	0	0	0	0	0	0	0	0	0	0	0

Note: BW = Bandwidth, the number of nearest neighbours (percentage of overall observations); CN = Condition number; PC.x = Property condition level x; PT.y = Property type level y.

The locations with high condition numbers are plotted in Figure 6.9A, collinearity predominantly occurs in the surrounding areas of Limerick city. This may be due to the homogeneity of observations in the housing estates of the area. Collinearity in the south CK region may be due to the quality of data influencing the parameter estimation of regression points in the area.

The GWR-1, GWR-2 and GWR-3 models of this section take the form of

$$\log(y_i) = \beta_{i0} + \sum_{j=1}^{p} \beta_{ij} x_{ij} + \sum_{k=1}^{q} \gamma_{ik} z_{ik} + \sum_{l=1}^{r} \delta_{il} t_{il} + \varepsilon_i \qquad i = 1, .., n \tag{6.5}$$

where y_i represents the price of property i with an iid residual value $\varepsilon_i \sim N(0, \sigma^2)$. Here, β_{i0} is the intercept term for regression point i representing the mean $\log(Price)$. In the case of the GWR-3 model, the continuous effects of observation i are grouped as $\mathbf{X}_i = \{Beds_i, Baths_i\}$ and z_i is a dummy variable for the presence of a 'Fireplace'. The temporal effects for observation i are modelled as a dummy variable represented by t_i. In general, when more dummy and temporal variables are included, the independent variables at regression point i are represented in vector form as $\boldsymbol{X}_j$, $\boldsymbol{Z}_k$ and $\boldsymbol{T}_l$. These variables have corresponding regression scaling coefficients $\boldsymbol{\beta}_j$, $\boldsymbol{\gamma}_k$ and $\boldsymbol{\delta}_l$, respectively.

Fotheringham & Oshan (2016) suggest removing problematic variables to resolve collinearity in the GWR model. 'Property Type' and 'Property Condition' are replaced by 'Property Category', which contains two levels: house and apartment. The various property types and conditions are not uniformly distributed around the geographical space, thus, local estimations of a certain property type based on their nearest neighbours may be biased compared to a global approach which is calibrated using all observations. Models with all variables are fitted using the same kernel functions as in Table 6.3, and the collinearity diagnostics similar to that of the GWR-1 model are observed. A parsimonious model, GWR-3, is selected based on Figure 6.7, this includes 'Beds', 'Baths' and 'Fireplace', along with 'Years' to account for temporal effects. The GWR-3 model shows no signs of extreme multicollinearity, with a maximum condition number of 14.89, and a maximum VIF value of 7.89 for 'Beds'. The median for local condition numbers is 1.36, similarly, the median values of the VIFs of each parameter do not exceed 1.35. The GWR-3 model does not exhibit any concerning multicollinearity and is a favoured parsimonious model. Thus, the GWR-3 model is selected for further analysis and interpretation. A corresponding global model for GWR-3 is fitted using all observations for model calibration.

The residuals of the GWR-3 model are plotted against their associated fitted values in 6.8A. The residuals appear to have no discernible pattern and constant variance. There is a possible influential point for further investigation. This value is on the lower end of the price scale and has a negative residual value suggesting it is a low-priced property that is overestimated by the model. This observation is likely influenced by local irregularities in the model calibration. The histogram of residual values in 6.8B appears to be relatively normally distributed. In a similar case to the OLS model, the Kolmogorov-Smirnov test results in a p value < 0.05, which means there is sufficient evidence to reject the null hypothesis that the residual values are normally distributed. The model assumptions outlined in 5.2 are satisfied by observation.

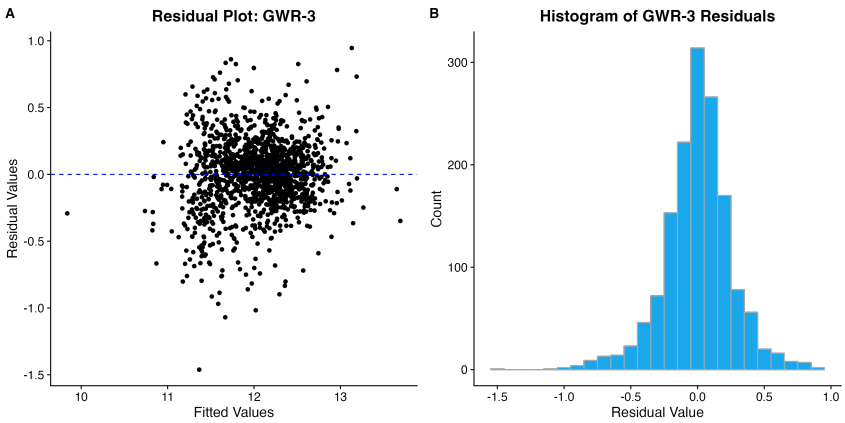

Figure 6.8: GWR-3 model residuals: A - and fitted values, B - histogram plot.

6.3.4 Results and Model Interpretation

In Table 6.5, summary statistics of the parameter values and exponential parameter values are compared to the global OLS model parameters. The median GWR model coefficients align with the global estimates in magnitude and sign. The ranges of exponentiated GWR model coefficients are compared, considering a value above 1 contributes to an increase in the price estimate while a value below 1 contributes to a decrease, ceteris paribus. Both 'Baths' and 'Beds' display changes in signs across the local models. 'Baths' has the lowest minimum value, or largest negative multiplicative effect, of the two while 'Beds' has the highest maximum value, or positive multiplicative effect, of the two.

Table 6.5: Coefficients of the GWR-3 model and the corresponding global model with p values for the global model and a Monte Carlo test of spatial heterogeneity.

Term	GWR-3 Model			Global Model		Monte Carlo
	Min	Median	Max	Estimate	p value	p value
Intercept	10.82	11.93	12.50	11.90	-	<0.01
	(€50,171)	(€152,257)	(€268,417)	(€147,018)		
Baths	-0.16	0.16	0.4716	0.22	<0.001	0.02
	(0.85)	(1.17)	(1.6)	(1.25)		
Beds	-0.11	0.17	0.79	0.16	<0.001	<0.01
	(0.9)	(1.19)	(2.19)	(1.17)		
Fireplace	-0.28	0.09	0.67	0.11	<0.001	0.52
	(0.76)	(1.09)	(1.95)	(1.11)		
Year.2018	-0.21	-0.03	0.13	-0.03	<0.01	0.99
	(0.81)	(0.97)	(1.14)	(0.97)		

Note: Exponential values for coefficients are given in parentheses under the coefficient values.

These results may seem counter-intuitive as one would expect additional bedrooms to increase property value. However, previous studies from Oust et al. (2020) and Sisman & Aydinoglu (2022), for example, demonstrate that sign changes are commonly seen in GWR modelling, and these occurrences are best explored geographically. Similarly, the coefficients for 'Fireplace' and 'Year' change sign throughout the regression points, suggesting spatial variation in these effects.

The use of the GWR model as a predictive tool has increased in popularity, where the dependent variable is estimated based on a weighted combination of surrounding values (Harris, Fotheringham, Crespo & Charlton 2010). Following the approaches of Bitter et al. (2007) and Harris, Fotheringham, Crespo & Charlton (2010), the cross-validation method described in Section 5.2 and formulae from Section 5.3 are used to analyse the results. The GWR-3 model is compared to its global alternative in Table 6.6. The GWR model outperforms the global approach across all metrics, except for the proportion of actual values within their 95% predictive interval suggesting a wider interval and possible uncertainty of the global model.

From the Monte Carlo simulation of coordinate perturbations, the differences in the observed and simulated model coefficient variances are tested to explore the spatial heterogeneity of coefficients. The p values for each variable are listed in Table 6.5. At the 95% level of significance, the null hypotheses of stationarity for the intercept, 'Baths' and 'Beds' are rejected. The variation in the intercept, which is the mean price of a 3-bedroom, 2-bathroom property without a fireplace sold in 2017, likely captures exogenous effects, such as location and neighbourhood effects. While differences in the coefficient terms suggest that the effects vary through space. For 'Fireplace' and 'Year', the p value is > 0.05, there is insufficient evidence to conclude that there is spatial variation in these coefficients at the 5% level of significance, it appears these effects are fixed across space. This is expected for the temporal effect but contradicts the expectation of spatial variation in the effect of a fireplace. It is likely that the previously seen spatial variation of the effect of a fireplace is accounted for by the variation of all other coefficients in the model.

Table 6.6: 10-fold cross-validation results for the GWR-3 model and the corresponding global model.

Model	Adj. R^2	AIC(c)	RMSE	Median APE %	Within 10% of Price	Within 20% of Price	Within 95% Prediction Interval	Moran's I
Global	0.38	1420	€85,115	23.9%	22.9%	42.3%	93.7%	0.22
GWR-3	0.73	826	€72,008	16.4%	32.6%	57.9%	92.6%	0.01

10-fold cross-validation is performed on models to calculate the adjusted (Adj.) R squared, (corrected) Akaike information criterion (AIC(c)), root mean squared error (RMSE), median absolute percentage error (APE), the proportion of observations within 10% and 20% error intervals around actual price and within their 95% prediction intervals. Moran's I is specified for each model fitted to all observations.

The intercept values are plotted in Figure 6.9B. There are clusters of similar values which somewhat align with the municipal boundaries, this supports the conclusions of Bitter et al. (2007) who identify the potential of the GWR model as a submarket identification tool. NC has the lowest intercept values of approximately €90,000, reducing westward, this aligns with the fact that NC has the lowest median property value. There is an apparent "spillover" of intercept values from NC to AR, this occurs since the kernel is not restricted to the municipal boundaries. In comparison to CK and NC, AR contains the most heterogeneous distribution of intercept values suggesting differences in the mean price within the municipality. The highest intercept values are along the intersection of AR and LCS. This conforms with the residual pattern of the basic OLS model in Figure 6.2. The intercept values are heterogeneous in LCS, where high values are apparent in the North City, and lower intercept values are observed in the surrounding areas. These high values likely represent the high-priced areas of the Ennis Road and the North Circular Road in LCS-2.

In Figure 6.9C, the coefficient of 'Beds' displays a different spatial pattern. Most values are unitary with slight increases in LCS, south CK, and east NC. There are obvious outliers in south NC, these are likely caused by the presence of a 5-bedroom property worth €560,000 in the area. Since some of the points are located in relative proximity, this high-priced property receives a greater weight for the calibration of surrounding regression points. Additionally, the lack of other near-neighbours possibly introduces bias when estimating the value of an additional bedroom on property price. This highlights the sensitivity of the GWR model to the quality of the local observations included in model calibration. The exploratory t-test results for the effect of an additional bedroom are plotted in Figure 6.9D, where the test statistics within the range of ± 1.96 are highlighted as significant. There is an alignment between the lower coefficient values in Figure 6.9C and the significant observations in Figure 6.9D. The relationship between the number of bedrooms and property price is likely stronger for lower coefficients. This follows the approach of Fotheringham et al. (1998) for understanding the strength of the relationship, however, conforming to the the model assumptions outlined by Harris, Fotheringham & Juggins (2010), it does not support an inference.

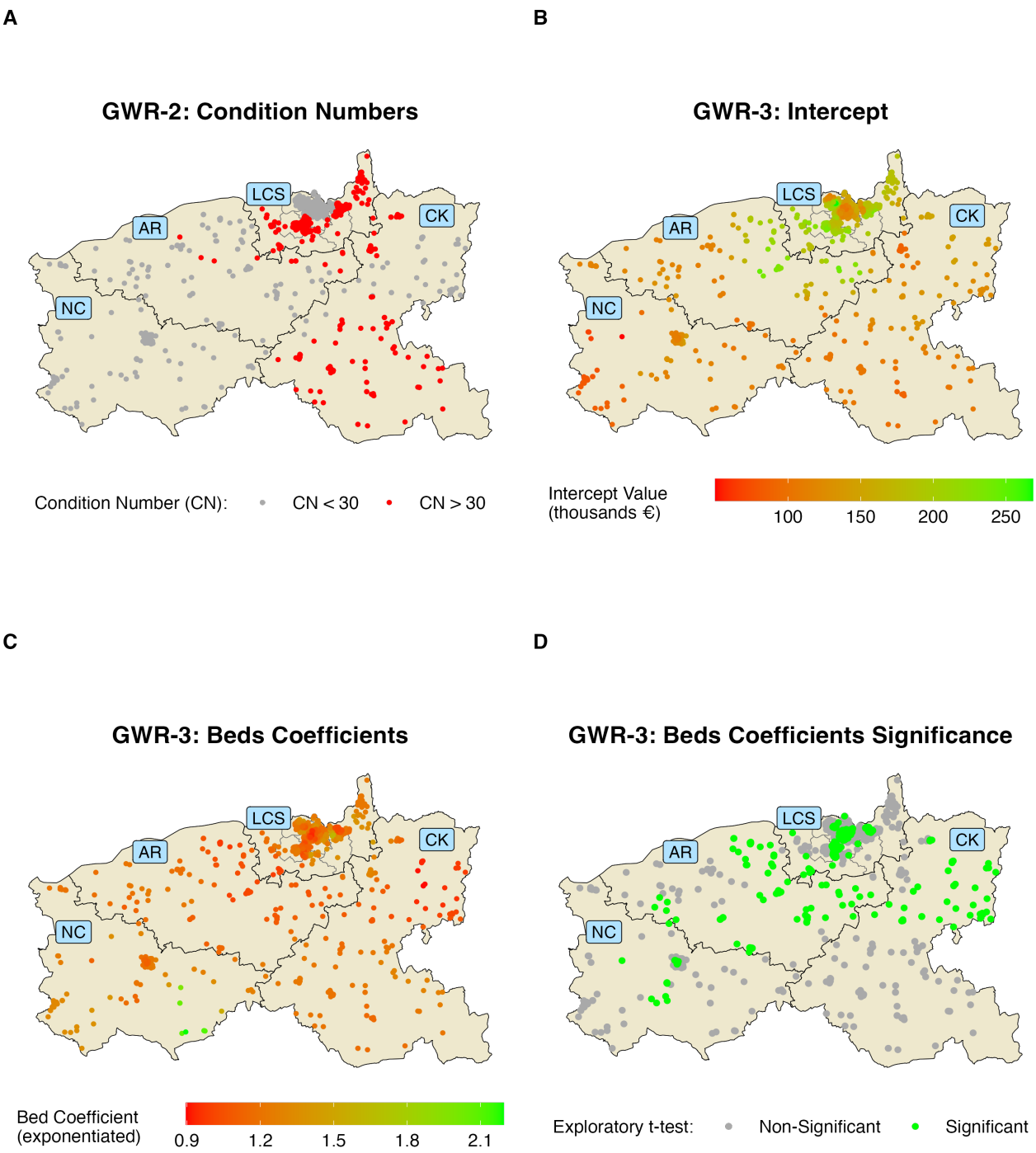

Figure 6.9: Selection of GWR modelling maps.

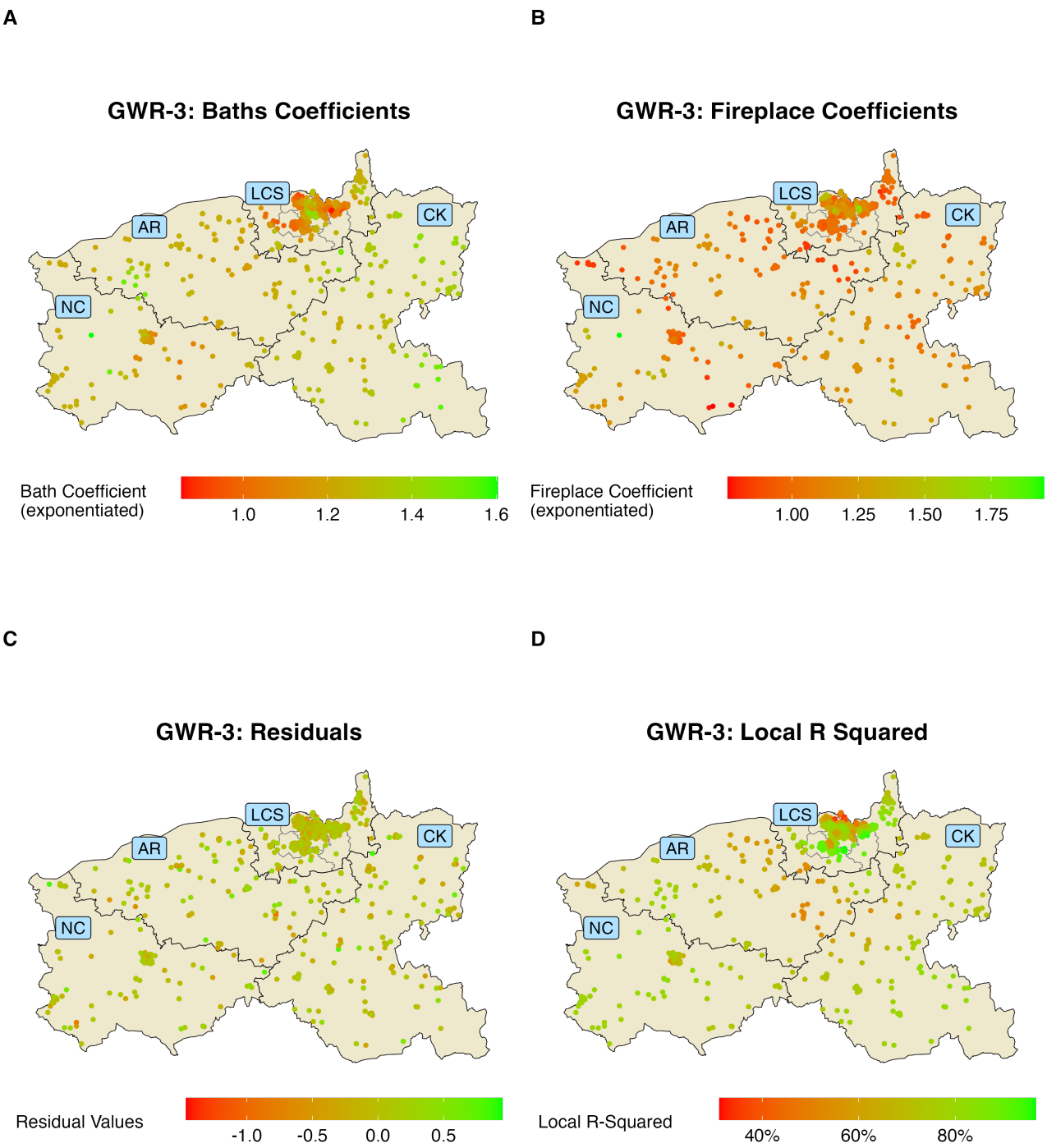

Figure 6.10: Selection of GWR modelling maps continued.

Observing the exponentiated values of the 'Baths' coefficient in Figure 6.10A. There are clear differences between LCS and the surrounding county. This aligns with the test for non-stationarity in table 6.5. Bitter et al. (2007) identify similar patterns between the central city and surrounding areas. There is a negative effect in the areas surrounding the city of LCS and a cluster of positive coefficients in the city centre. This is likely due to the homogeneous properties in the suburbs surrounding the city, where most properties have 3 bedrooms and 2 bathrooms (Table A.2), which aligns with the intercept in this case. The distribution of coefficients in the surrounding county is similarly homogeneous with many positive values increasing eastward. This may be due to the presence of larger detached houses in the countryside.

The fireplace coefficient (Figure 6.10B) displays no obvious pattern, there is a high valued coefficient in the centre of NC, likely due to the lack of proximal observations to this regression point which do not strongly influence the coefficient estimation. Values are relatively unitary signifying a homogeneous effect of fireplace on property price across the county, this aligns with the stationarity relationship identified by the Monte Carlo significance test in Table 6.6. The values are slightly higher in LCS and surrounding areas compared to south CK and NC; 69% and 72% of observations in CK and NC contain fireplaces, respectively, compared to 58% of LCS properties (Table A.2). The increased occurrences of fireplaces in such regions may reduce the associated monetary effect. It is also possible that 'Fireplace' is acting as a proxy for urbanisation or population density as discussed in Section 6.2.6.

The local R^2 values are plotted in Figure 6.10D, these are the coefficients of determination for each regression point and can be used as an exploratory tool for model fit. The goodness of fit appears to be lowest in LCS and north AR. The spatial pattern of the local R^2 resembles the intercept value, where the areas with higher local R^2 appear to have a lower intercept value and low local R^2 values are in similar locations to high intercept values. This is suggestive of differences in model fit for different price strata.

Chapter 7

Results

In Table 7.1, the best-performing OLS and GWR models are compared. The OLS-cbdr-u model is compared to the GWR-3 model, thus, demonstrating that the GWR-3 model outperforms all other OLS models from Table 6.2. This aligns with the results of Oust et al. (2020), who conclude the GWR model outperforms the non-spatial OLS model and the OLS model with dummy variables for regions. Similarly, Bitter et al. (2007) conclude that the GWR model is superior to the OLS model with interactive absolute location terms. The GWR-3 model has a higher adjusted R^2 compared to the OLS model, suggesting improved goodness of fit. This aligns with the results of Sisman & Aydinoglu (2022), who report R^2 values of 0.55 and 0.57 for the OLS and GWR models, respectively. This could be attributed to the benefits of local model calibrations where the parameters are estimated based on the 73 nearest neighbours rather than all observations. The OLS model has lower values for the AIC, RMSE and median APE. This is likely due to the reduced quality of local data for the GWR model calibration and the additional parameters for each local regression model. The predicted price of an observation depends

Table 7.1: 10-fold cross-validation results of the best OLS and GWR models.

Model	Adj. R^2	AIC(c)	RMSE	Median APE %	Within 10% of Price	Within 20% of Price	Within 95% Prediction Interval	Moran's I
OLS-cbdr-u	0.66	667	€63,663	16.3%	32.7%	57.4%	93.9%	0.08
GWR-3	0.73	826	€72,008	16.4%	32.6%	57.9%	92.6%	0.01

10-fold cross-validation is performed on models to calculate the adjusted (Adj.) R squared, (corrected) Akaike information criterion (AIC(c)), root mean squared error (RMSE), median absolute percentage error (APE), the proportion of observations within 10% and 20% error intervals around actual price and within their 95% prediction intervals. Moran's I is specified for each model fitted to all observations.

on the surrounding neighbours of the regression point, and thus the quality of data (Harris, Fotheringham, Crespo & Charlton 2010). The effects of local data quality on model accuracy are demonstrated in Figure 6.10D where the local R^2 value varies throughout the geographical space. Additionally, local irregularities in the data result in outliers in the coefficient values in Figures 6.9C, 6.10A and 6.10B, this aligns with the identification of extreme coefficients by Farber & Yeates (2006). The OLS model considers all observations in the estimation of model coefficients representing the effects of property attributes, contributing to more accurate predictions and reduced prediction errors.

The proportion of estimates within 10 and 20 percent of price are similar for each model, the GWR model has an increased predictive error but maintains a similar level of accuracy to the OLS model. These results may be due to the local irregularities at regression points, influencing the overall model results. The GWR model has a considerably lower Moran's I value, allowing spatial variation of parameters appears to be an effective solution for reducing the spatial autocorrelation of residuals. This is a similar result to Farber & Yeates (2006), who note Moran's I values of 0.2 and approximately 0 for the basic OLS and GWR models, respectively. Similarly, Oust et al. (2020) report Moran's I values of 0.14 and 0.06 for the OLS model with submarkets and GWR model, respectively. The variation of the intercept of the GWR model in Figure 6.9 is the likely reason for this reduction in spatial autocorrelation, as it captures spatial variation in the price variable. Applying a local GWR model compared to the OLS model represents the variation in price as "valleys and hills" rather than a dichotomy of regions in space created by dummy variables in the OLS model (Brunsdon et al. 1996).

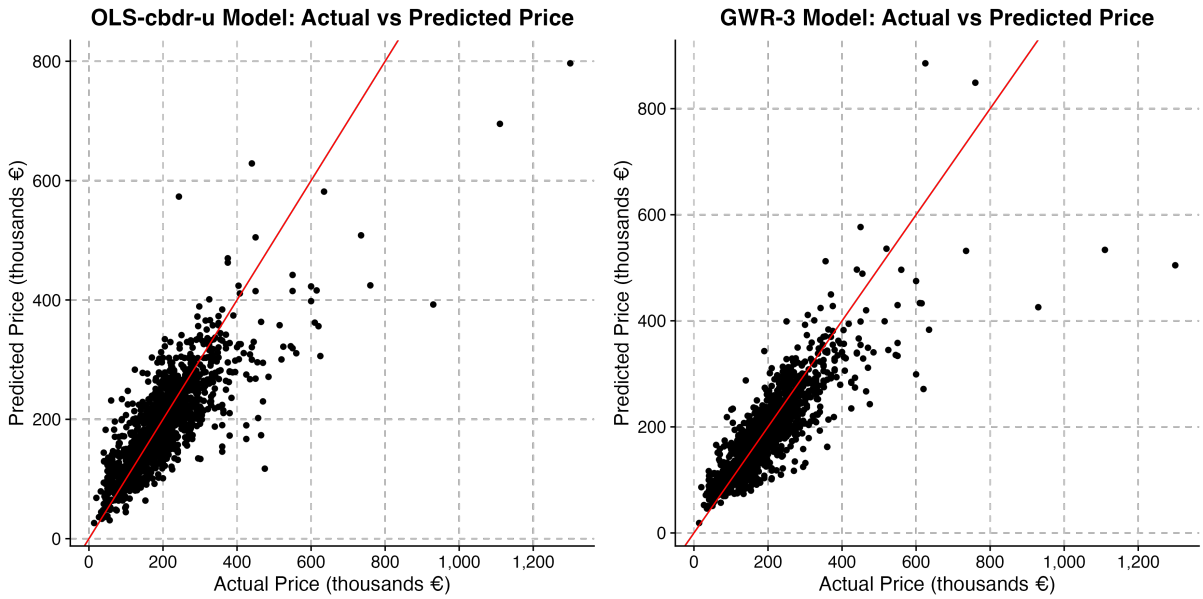

Figure 7.1: Predicted vs actual price values for the OLS-cbdr-u and GWR-3 models.

The exponentiated fitted values are plotted against the observed price for the OLS-cbdr-u and GWR-3 models in Figure 7.1. Both models appear most accurate at predicting prices for properties below €400,000. The residuals increase as the price increases, this is indicative of heteroscedasticity. The pattern is less apparent for the residuals on the log scale in Figures 6.4A and 6.8A and does not contradict the assumption of homoscedasticity. For properties below €400,000, the OLS model appears to underestimate the property price more than the GWR model, while both models do not overestimate the price at the same magnitude. The OLS model is likely underestimating price due to the lack of measures representing properties of superior quality or those on large sites, which compromises the predictive quality for such properties.

There are outliers with prices over €700,000, one such outlier is a 6-bedroom, 6-bathroom house with no description of design features or site size, as a result, this property condition is modelled as "second-hand" leading to the underestimation of price by the OLS model. The GWR model underestimates such outliers to a greater extent due to the lack of similar neighbours for each calibration point. Similarly, outliers on the lower end consist of properties priced at approximately €40,000 with no mention of quality, these are also modelled as in "second-hand" condition. This aligns with the outlying point of the GWR-3 residuals in 6.8A which was overestimated due to a lack of similar calibration points. It appears the OLS model generalises properties that are not either derelict, in need of renovation, or new, and thus fails to correctly predict their values. The GWR model fails to accurately predict the more expensive properties that are sparsely distributed around the geographical space.

The GWR-3 model is used to create a spatial surface over County Limerick. A grid of equally spaced regression points is created for estimation using the GWR-3 model using all observations as the calibration data. The predicted price is obtained for each value on this spatial grid, and using a grid raster of smaller resolution, the values between these regression points are interpolated using inverse distance weighting. In a similar manner to the kernel functions in Section 6.3, this function interpolates the missing values based on the weighted prices of all of the regression points. The resulting surface is mapped in Figure 7.2, where blue and red correspond to low and high predicted price values, respectively. There are apparent "hotspots" in the areas surrounding Limerick city. These areas correspond to Castletroy in the east and Dooradoyle in the south-west. The lower priced properties in the inner LCS areas could be thought of as a "valley" between these two areas resembling "hills". The high predicted values in the northwest of Limerick city are likely influenced by the high-priced local observations on the Ennis Road in LCS-1. The values in CK and AR appear to be relatively uniform and resemble the intercept value in Figure 6.9B. There is a cluster of more expensive properties in AR resembling the town of Adare and a cluster of more expensive properties in NW near the town of Newcastle West. The location of the lower priced properties along the south border of NW aligns with that of the local irregularities of the 'Beds' coefficients in 6.9C.

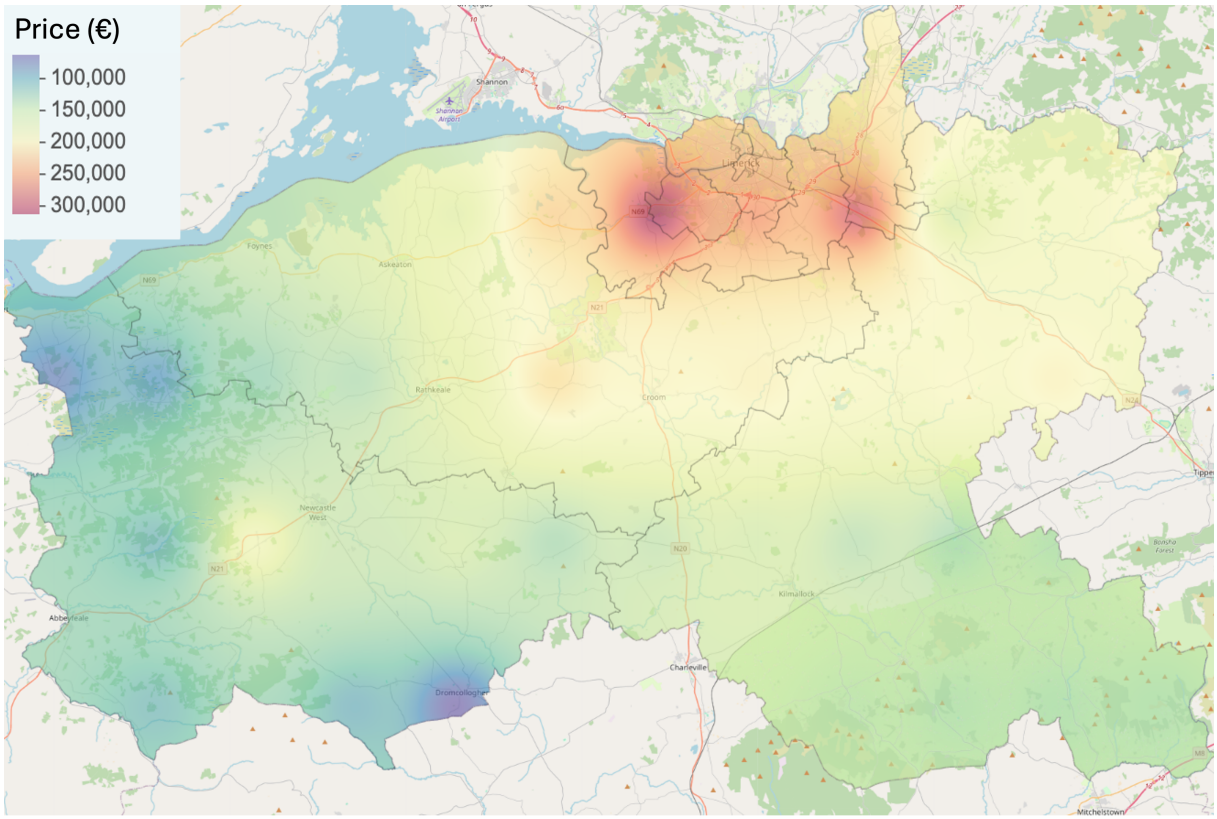

Figure 7.2: Predicted price surface in County Limerick using the GWR-3 model.

Chapter 8

Discussion, Conclusions and Further Work

8.1 Discussion and Conclusions

Due to the large proportion of missing values amongst key predictor variables in the data, the cleaning and filtering of data formed a large part of the initial project work. The cleaning approach involved multiple stages of decision-making and further inspection of values. The imputation of missing values was not explored as an implication of the resulting time constraints. Overall, the data cleaning and text mining resulted in a complete dataset for the Limerick region with minimal missing values, no irrelevant entries and no extreme outliers based on observation. The data appears to be representative of the property market in Limerick.

The application of geocoding to correct property locations and reverse geocoding to extract regions allowed for the development of locational attributes, such as distance from the city. The lack of postcode definitions in Limerick city, identified by O'Hanlon (2011) as the main obstacle to modelling property prices outside of Dublin, was overcome by combining electoral divisions to represent regional areas. Correct locational attributes allowed for the comparison of properties across the area and the implementation of spatial elements in hedonic models.

Through an exploration of the outliers in price, key terms common to the low-priced and high-priced properties were identified. Examples include "derelict" and "renovation" for low-priced houses, and "period home" for outliers on the upper end. These terms contributed towards explaining price variation, and thus, demonstrate the benefits of extracting value from the data through text-mining. Extracting such property attributes from property descriptions has many challenges. Information on properties is given at the realtor's discretion and the quality

of data depends entirely upon whether property conditions and features are correctly specified. Quantifying the language used to describe property listings is a key challenge, for example, terms such as "wonderfully restored", "supreme condition" and "beautiful condition" are difficult to differentiate. Due to the non-existent scale for quantifying the condition of a property, many properties are categorised as "second-hand" despite having varying levels of quality.

The exploratory analysis focused on examining the property types, property prices, and property categories between municipalities and urban areas. The lack of various property types, such as apartments, across Limerick county posed challenges to the development of local regression models. Notable differences were observed in property price and structural attributes between municipalities, where the distribution of such characteristics differs between Limerick city and Limerick county. Apparent differences were observed between property price and distance from the city centre, these results reflected the price distributions between municipalities. In LCS, the property price appears to be lower for inner-city properties compared to those in the surrounding areas. This aligns with the research of McCafferty (2011), who identifies a spatial concentration of poverty in the city centre.

The modelling approaches described in Chapter 2 are compared across many empirical studies. There is difficulty in directly comparing the results of these studies due to differences in the sample data. Oust et al. (2020) and Bourassa et al. (2010) compare relevant approaches using a common dataset and test model accuracy on common out-of-sample data. The benefits of using well-defined submarkets and geographically weighted regression models are identified in each case. These findings align with the results of this study, where such models are used to capture spatial relationships within submarkets.

The multiple linear regression model facilitates an exploration of the methods of accounting for location in property price prediction. Dubin (1988, 1992) outlines a risk of measurement error when including spatial location in the model where submarkets are arbitrarily defined, and spatially autocorrelated residuals are a result of the differences between the true neighbourhoods and submarket definitions. The preceding analysis contradicts these concerns, where the benefits of modelling submarkets through the use of urban area dummy variables are apparent from increased predictive accuracy and reduced residual autocorrelation across all locational models. Including urban areas and distance to the city centre in bands resulted in the best OLS model overall. Although the residual autocorrelation reduced with increased spatial complexity for the OLS models, the concerns raised by Basu & Thibodeau (1998) and Dubin (1988), who note the potential bias of test statistics of the OLS model coefficients, are addressed. This is due to the dependence among observations, apparent from spatially autocorrelated residuals, and can result in mistaken inference.

Residual autocorrelation from the global model approach may be due to the model missing a spatial process where global models discard large amounts of potentially valuable information on spatially varying relationships, Farber & Yeates (2006) suggest using a local approach, such as the GWR model to account for this. The fitted GWR models demonstrate the importance of data quality for model calibration, where local irregularities have a strong effect on the coefficients and model fit of each local model. Problems with collinearity are apparent with the GWR model, and the correct selection of parameters and bandwidth is essential to counter this challenge. Each model has demonstrated some degree of spatial autocorrelation among residuals. This may be due to omitted variable bias, where nearby properties share similar omitted variables and thus have similar residuals. There is an awareness of the omitted variables in this analysis, where neither model includes the social and economic attributes of regions nor information on property size and energy rating. While the GWR model results allow for the visualisation of parameter behaviour, there is a lack of consensus on the most effective way to test for the local significance of relationships, thus, the model has not been used as an inferential tool.

The results of the OLS modelling approach are as expected, where the new property condition and detached property type have the greatest effects on property price. Similarly, end-of-terrace properties are associated with greater property prices compared to terraced properties. Period houses, which represent the majority of high-priced observations, are associated with a large increase in property price. The data quality has affected the parameter estimation, for example, the confidence interval of the effect of a 7th bedroom is wide due to a lack of such properties. The spatial variation of the effect of an additional bedroom is apparent from the OLS model, where the added value of a 4th bedroom varies between urban areas, resulting in a greater increase in Outer City South compared to NW, for example. Similarly, the GWR model results demonstrate significant variation in the bedroom coefficient, with greater effects in Castletroy and City North compared to the inner city. The spatial variation of the fireplace coefficient apparent in the OLS model is non-significant in the GWR model. It is expected that the fireplace coefficient captures exogenous effects, such as socioeconomic factors, in the OLS model, and these are accounted for by the spatially varying intercept in the GWR model. The spatial pattern of this intercept term resembles that of the municipalities in Limerick, likely capturing the underlying submarkets possibly resulting from such exogenous effects.

The GWR model has increased spatial enhancements resulting in increased predictive accuracy, while the OLS model has a reduced predictive error. The effectiveness of the GWR model as an exploratory tool for spatial relationships is apparent. Following the advice of Comber et al. (2023) it is most effective to use these models collectively to understand the effects of spatial heterogeneity and autocorrelation.

8.2 Further work

Property Characteristics and Submarket Identification

The data cleaning methods described in Chapter 3 could be enhanced to extract more complete data on each property. The main considerations are outlined as follows:

- There is potential to mine text descriptions for property size, this was not completed in the initial data cleaning due to many obstacles encountered and resulting time constraints. Many property sizes are given in terms of *length* x *width* for each room, these would need to be multiplied and added together to give the overall area. There are also issues surrounding the units of measurement, meters and feet are used alternatively in property descriptions.
- More information on property design (bungalow, dormer, etc.) and number of floors could be obtained from property descriptions. This further information has the potential to specify differences in property conditions of second-hand properties.
- The missing values for BER, size and property attributes could be investigated by exploring the properties with valid entries for these variables. O'Hanlon (2011) imputes missing values for structural attributes with the median values from previous years. This raises the potential risk of biased regression coefficients, and a preferred multiple imputation method would involve predicting the missing values based on the known values (Donders et al. 2006).

The use of aggregated local electoral divisions potentially introduced bias and measurement errors into this analysis (Dubin 1988). There is potential to apply a clustering method to understand the true submarkets in LCS. Oust et al. (2020) implement a K-means clustering method using the distance between property coordinates to construct artificial market districts and such models with the K-means submarkets outperform those with the a priori regional boundaries.

Modelling Methods

As outlined in Section 2.2, many spatial models aim to account for spatial autocorrelation. One such method of overcoming this is by dropping the assumption of independent error terms. Thus, the error covariance structure could be modelled to implement the regression Kriging methods. There is potential to enhance the GWR models by implementing a multi-scale GWR model, this was not implemented due to computational constraints. The identification of non-linear and spatially varying relationships between covariates and property price suggests a generalised additive model (GAM) could be a suitable modelling method to implement. In a similar manner to Hurley & Sweeney (2022), non-linear relationships between CBD distance, the number of bedrooms and price could be modelled using smoothing splines. Spatial effects could be accounted for by modelling the Cartesian coordinates using a Gaussian process.

Further Research Questions

As part of the extensive research into the social segregation outlined in Section 2.2, McCafferty defines social area types within Limerick city. There is potential to use these social area types in combination with present census statistics on family status and socioeconomic status to include socioeconomic variates, aggregated by the defined urban areas, in the model.

Properties located in the areas surrounding the Ennis Road, the North Circular Road and Castletroy have demonstrated considerably higher sales prices among the observations and the majority of outliers on the upper end of the price scale are observed in these areas. There is potential to investigate the misspecification of addresses in LCS and quantify the added value of having a "Castletroy" or "Ennis Road" address. Hurley & Sweeney (2022) have investigated the impact of mislabelling Dublin postcodes, and conclude that more favourable postcodes are associated with increased property values. As an example, the Ennis Road is highlighted for the region of LCS-1 and properties that mention "Ennis Road" in the text address are plotted in Figure 8.1. The points are coloured according to whether the observations are located on the Ennis Road based on visual inspection. There is a clear dispersion of properties with "incorrect" addresses around the Ennis Road highlighting the potential to investigate the additional monetary value of listing such dwellings as being located on the Ennis Road.

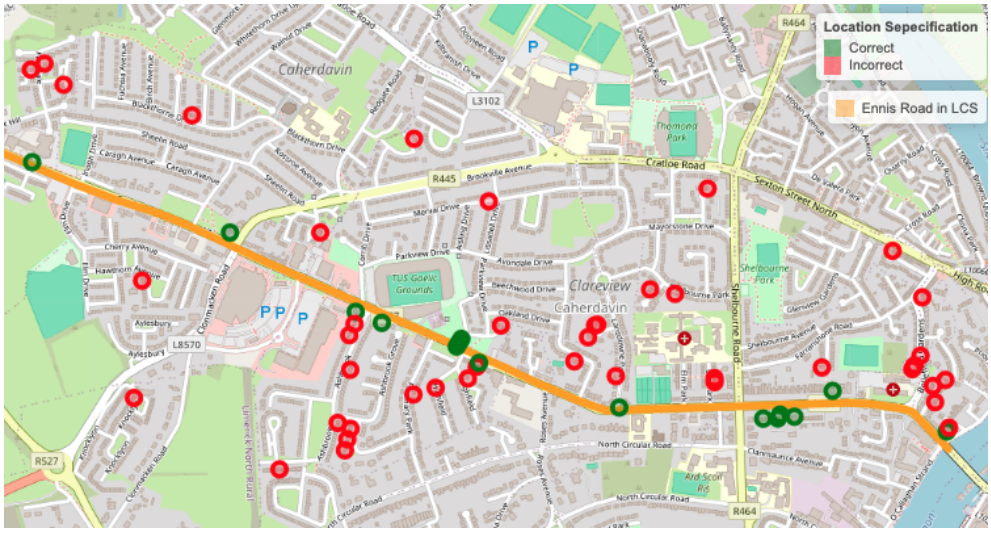

Figure 8.1: Map of the Ennis Road in LCS-1, with Ennis Road properties coloured by the correctness of address specification.

Appendices

A Exploratory Data Analysis

Table A.1: Summary statistics of Limerick houses by property type.

Property Type	Count	Median (€)	IQR (€)	Min (€)	Max (€)
Limerick city and suburbs (LCS)					
Detached	153	316,000	137,500	60,250	1,300,000
Semi-Detached	450	212,500	69,300	40,000	520,000
Terraced	136	162,000	109,500	20,000	735,000
End-of-Terrace	55	168,000	1147,49	52,000	930,000
Townhouse	40	192,000	72,500	57,000	255,000
Limerick county					
Detached	273	176,000	128,000	14,000	450,000
Semi-Detached	144	136,500	39,190	27,000	470,000
Terraced	43	85,000	617,500	37,500	450,000
End-of-Terrace	6	101,500	49,375	53,000	177,499
Townhouse	12	87,500	115,000	57,500	475,000

Table A.2: Summary statistics for property attributes and features across municipalities and urban areas.

Area	Count	Mean Values			Count (Proportions)											
		Beds	Baths	CBD	House		Apartment		Garage		Cul-de-sac		Fireplace		Period	
Municipality																
AR	124	3.53	2.09	21.56	124	(1.00)	0	(0)	64	(0.52)	8	(0.06)	87	(0.7)	0	(0)
CK	163	3.48	1.98	24.25	163	(1.00)	0	(0)	81	(0.5)	10	(0.06)	113	(0.69)	2	(0.01)
NW	198	3.36	2.07	42.38	190	(0.96)	8	(0.04)	81	(0.41)	12	(0.06)	145	(0.73)	0	(0)
LCS	1012	3.30	2.15	3.34	834	(0.82)	178	(0.18)	351	(0.35)	227	(0.22)	591	(0.58)	19	(0.02)
Urban Areas																
LCS-1	141	3.40	2.12	2.39	121	(0.86)	20	(0.14)	72	(0.51)	29	(0.21)	88	(0.62)	5	(0.04)
LCS-2	47	2.87	1.64	0.90	34	(0.72)	13	(0.28)	12	(0.26)	6	(0.13)	28	(0.6)	3	(0.06)
LCS-3	68	3.01	2.06	1.29	50	(0.74)	18	(0.26)	21	(0.31)	20	(0.29)	32	(0.47)	0	(0)
LCS-4	98	2.65	1.55	0.66	43	(0.44)	55	(0.56)	7	(0.07)	3	(0.03)	30	(0.31)	7	(0.07)
LCS-5	33	3.21	2.45	8.85	33	(1.00)	0	(0)	9	(0.27)	7	(0.21)	14	(0.42)	1	(0.03)
LCS-6	209	3.32	2.32	4.17	187	(0.89)	22	(0.11)	83	(0.4)	69	(0.33)	139	(0.67)	1	(0)
LCS-7	152	3.17	1.74	1.83	136	(0.89)	16	(0.11)	63	(0.41)	34	(0.22)	93	(0.61)	1	(0.01)
LCS-8	196	3.69	2.60	4.32	166	(0.85)	30	(0.15)	53	(0.27)	48	(0.24)	119	(0.61)	0	(0)
LCS-9	68	3.71	2.50	8.20	64	(0.94)	4	(0.06)	31	(0.46)	11	(0.16)	48	(0.71)	1	(0.01)

Note: CBD distance is given in km. Period refers to period properties.

B Data Preparation

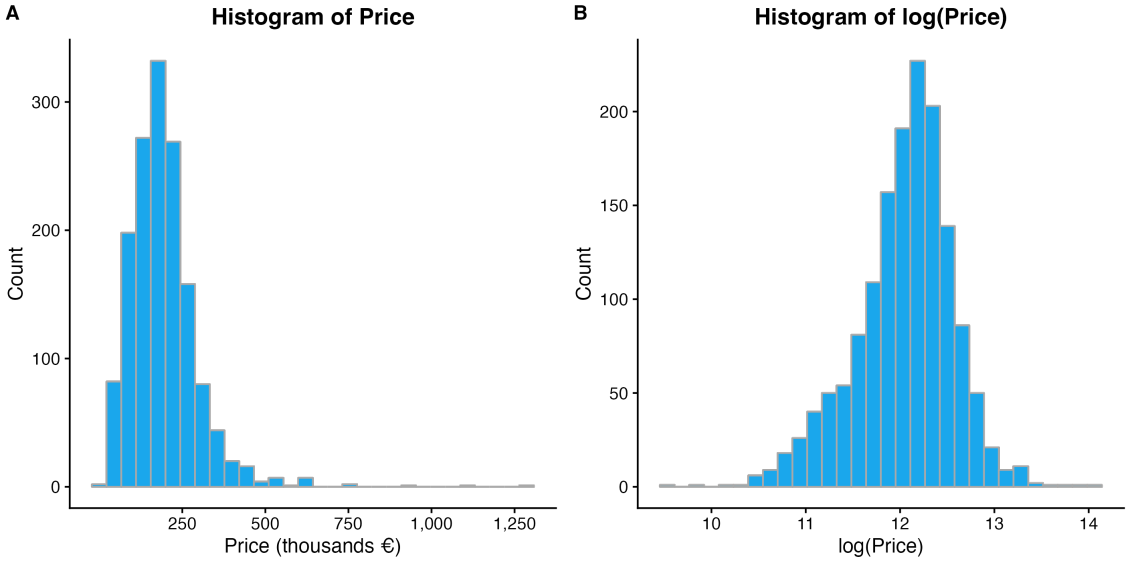

Figure B.1: Histogram of: A - property price, and B - log($Price$).

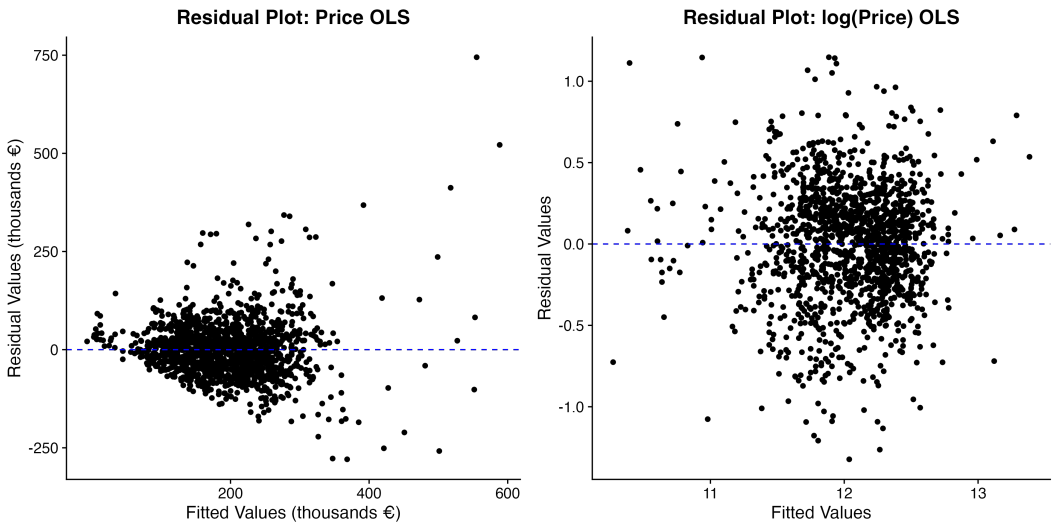

Figure B.2: Scatter plot of fitted values and residuals for the OLS models of price and log($Price$).

C Results

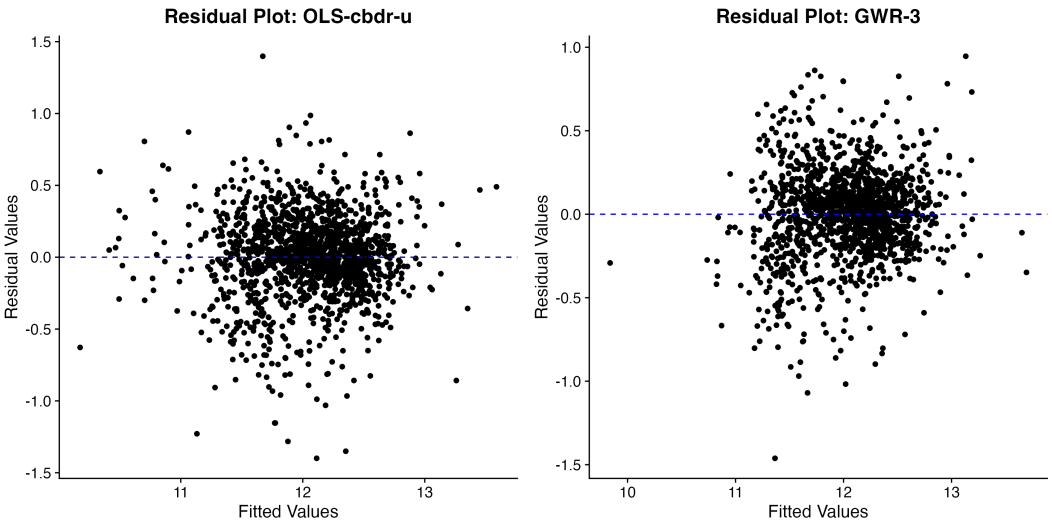

Figure C.1: Scatter plot of fitted values and residuals for the OLS-cbdr-u and GWR-3 models.

Table C.1: OLS-cbdr-u model coefficients and confidence intervals in standard and exponential format.

Variable	Standard Format Estimate	Standard Format 95% Confidence Interval	Exponential Format Estimate	Exponential Format 95% Confidence Interval
Intercept	11.66	[11.58, 11.73]	€115,427	[€106,661, €124,913]
Baths	0.18	[0.16, 0.2]	1.20	[1.17, 1.23]
Baths^2	-0.06	[-0.09, -0.04]	0.94	[0.92, 0.96]
Baths^3	0.01	[0, 0.02]	1.01	[1, 1.02]
Beds	0.12	[0.09, 0.15]	1.13	[1.1, 1.16]
Beds^2	-0.025	[-0.04, -0.01]	0.98	[0.96, 0.99]
Period	0.65	[0.51, 0.79]	1.92	[1.67, 2.21]
Fireplace	0.06	[0.02, 0.1]	1.06	[1.02, 1.1]
Areas				
(AR) Adare-Rathkeale	-0.07	[-0.2, 0.07]	0.93	[0.81, 1.07]
(CK) Cappamore-Kilmallock	-0.19	[-0.33, -0.06]	0.83	[0.72, 0.95]
(NW) Newcastle West	-0.12	[-0.27, 0.02]	0.88	[0.77, 1.02]
(LCS) Limerick city and suburbs				
(LCS-1) North	0.20	[0.12, 0.27]	1.22	[1.13, 1.31]
(LCS-2) City North	0.11	[-0.01, 0.22]	1.11	[0.99, 1.25]
(LCS-3) Corbally	-0.01	[-0.1, 0.09]	0.99	[0.91, 1.09]
(LCS-4) City South	-0.10	[-0.2, -0.01]	0.90	[0.82, 0.99]
(LCS-5) West	-0.01	[-0.13, 0.1]	0.99	[0.88, 1.1]
(LCS-6) Ballycumin	0.15	[0.08, 0.22]	1.17	[1.09, 1.25]
(LCS-7) Outer City South	-0.01	[-0.09, 0.06]	0.99	[0.91, 1.06]
(LCS-8) Ballysimon	0.08	[0.01, 0.15]	1.08	[1.01, 1.16]
(LCS-9) East	-0.02	[-0.11, 0.08]	0.98	[0.9, 1.08]

(continued)

Variable	Standard Format Estimate	Standard Format 95% Confidence Interval	Exponential Format Estimate	Exponential Format 95% Confidence Interval
Property Condition				
Derelict	-0.46	[-0.57, -0.35]	0.63	[0.56, 0.71]
New	0.41	[0.33, 0.49]	1.51	[1.39, 1.64]
Renovation	-0.22	[-0.33, -0.12]	0.80	[0.72, 0.89]
Second-hand	0.27	[0.21, 0.32]	1.31	[1.24, 1.38]
Property Type				
Semi-Detached	0.11	[0.07, 0.15]	1.12	[1.08, 1.16]
Apartment	-0.14	[-0.2, -0.07]	0.87	[0.82, 0.93]
Detached	0.34	[0.29, 0.38]	1.40	[1.34, 1.46]
Duplex	-0.14	[-0.25, -0.04]	0.87	[0.78, 0.96]
End-of-Terrace	-0.04	[-0.11, 0.04]	0.97	[0.9, 1.04]
Terraced	-0.10	[-0.15, -0.05]	0.90	[0.86, 0.95]
Townhouse	-0.03	[-0.11, 0.05]	0.97	[0.9, 1.05]
CBD Distance Range (Km)				
Within 5	0.03	[-0.07, 0.13]	1.03	[0.93, 1.14]
Between 5-10	0.14	[0.05, 0.22]	1.15	[1.05, 1.25]
Between 10-20	0.04	[-0.04, 0.12]	1.04	[0.96, 1.13]
Over 20	-0.21	[-0.3, -0.11]	0.81	[0.74, 0.9]
Interaction Terms				
Beds×AR	-0.03	[-0.08, 0.03]	0.98	[0.92, 1.03]
Beds×CK	0.00	[-0.06, 0.05]	1.00	[0.94, 1.05]
Beds×NW	0.04	[-0.01, 0.1]	1.04	[0.99, 1.1]
Beds×LCS-1	0.01	[-0.05, 0.07]	1.01	[0.95, 1.08]
Beds×LCS-2	-0.04	[-0.16, 0.07]	0.96	[0.85, 1.08]
Beds×LCS-3	0.05	[-0.04, 0.13]	1.05	[0.96, 1.14]
Beds×LCS-4	-0.12	[-0.18, -0.06]	0.89	[0.84, 0.94]
Beds×LCS-6	-0.06	[-0.12, 0]	0.94	[0.89, 1]
Beds×LCS-7	0.07	[0, 0.13]	1.07	[1, 1.14]
Beds×LCS-8	-0.02	[-0.07, 0.03]	0.98	[0.93, 1.03]
Beds×LCS-9	0.02	[-0.05, 0.09]	1.02	[0.95, 1.1]
Fireplace×AR	-0.01	[-0.12, 0.11]	0.99	[0.89, 1.12]
Fireplace×NW	-0.12	[-0.22, -0.02]	0.89	[0.8, 0.98]
Fireplace×CK	0.00	[-0.11, 0.1]	1.00	[0.9, 1.11]
Fireplace×LCS-1	0.14	[0.04, 0.25]	1.15	[1.04, 1.28]
Fireplace×LCS-2	-0.19	[-0.36, -0.02]	0.83	[0.7, 0.98]
Fireplace×LCS-3	0.05	[-0.09, 0.19]	1.05	[0.92, 1.21]
Fireplace×LCS-4	0.17	[0.03, 0.31]	1.19	[1.03, 1.37]
Fireplace×LCS-6	-0.05	[-0.14, 0.04]	0.95	[0.87, 1.04]
Fireplace×LCS-7	0.01	[-0.09, 0.11]	1.01	[0.91, 1.11]
Fireplace×LCS-8	0.12	[0.03, 0.21]	1.13	[1.03, 1.24]
Fireplace×LCS-9	-0.03	[-0.2, 0.13]	0.97	[0.82, 1.14]

Bibliography

1. Adair, A., Berry, J. & McGreal, W. S. (1996), 'Hedonic modelling, housing submarkets and residential valuation', *Journal of property Research* **13**(1), 67–83.

2. Alfiyatin, A. N., Taufiq, H., Febrita, R. E. & Mahmudy, W. F. (2017), 'Modeling house price prediction using regression analysis and particle swarm optimization', *International Journal of Advanced Computer Science and Applications* **8**, 323–326.

3. Allen-Coghlan, M. & McQuinn, K. M. (2021), 'The potential impact of covid-19 on the irish housing sector', *International Journal of Housing Markets and Analysis* **14**(4), 636 – 651.

4. Anselin, L. (1988), 'Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity', *Geographical analysis* **20**(1), 1–17.

5. Anselin, L. & Griffith, D. A. (1988), 'Do spatial effecfs really matter in regression analysis?', *Papers in Regional Science* **65**(1), 11–34.

6. Asadov, A., Ibrahim, M. & Yildirim, R. (2023), 'Impact of house price on economic stability: Some lessons from oecd countries', *J Real Estate Finan Econ* .

7. Basu, S. & Thibodeau, T. (1998), 'Analysis of spatial autocorrelation in house prices', *The Journal of Real Estate Finance and Economics* **17**, 61–85.

8. Belsley, D. A., Kuh, E. & Welsch, R. E. (2005), *Regression diagnostics: Identifying influential data and sources of collinearity*, John Wiley & Sons.

9. Benjamini, Y. & Hochberg, Y. (1995), 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *Journal of the Royal statistical society: series B (Methodological)* **57**(1), 289–300.

10. Bitter, C., Mulligan, G. F. & Dall'erba, S. (2007), 'Incorporating spatial variation in housing attribute prices: a comparison of geographically weighted regression and the spatial expansion method', *Journal of geographical systems* **9**, 7–27.

11. Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V. & Pebesma, E. J. (2008), *Applied spatial data analysis with R*, Vol. 747248717, Springer.

12. Boarnet & Chalermpong (2010), 'New highways, house prices, and urban development: A case study of toll roads in orange county, ca', **8**, 575–605.

13. Bogin, A. N. & Shui, J. (2020), 'Appraisal accuracy and automated valuation models in rural areas', *The Journal of Real Estate Finance and Economics* **60**, 40–52.

14. Bourassa, S. C., Hamelink, F., Hoesli, M. & MacGregor, B. D. (1999), 'Defining housing submarkets', *Journal of Housing Economics* **8**(2), 160–183.

15. Bourassa, S. C., Hoesli, M. & Peng, V. S. (2003), 'Do housing submarkets really matter?', *Journal of Housing Economics* **12**(1), 12–28.

16. Bourassa, S., Cantoni, E. & Hoesli, M. (2010), 'Predicting house prices with spatial dependence: a comparison of alternative methods', *Journal of Real Estate Research* **32**(2), 139–160.

17. Bowes, D. R. & Ihlanfeldt, K. (2001), 'Identifying the impacts of rail transit stations on residential property values', **50**, 1–25.

18. Brueckner, J. K. (1986), 'A modern analysis of the effects of site value taxation', *National Tax Journal* **39**, 49–58.

19. Brunnermeier, M. K. & Julliard, C. (2008), 'Money illusion and housing frenzies', *The Review of Finanical Studies* **21**, 135–180.

20. Brunsdon, C., Fotheringham, A. S. & Charlton, M. E. (1996), 'Geographically weighted regression: a method for exploring spatial nonstationarity', *Geographical analysis* **28**(4), 281–298.

21. Brunsdon, C., Fotheringham, S. & Charlton, M. (1998), 'Geographically weighted regression', *Journal of the Royal Statistical Society: Series D (The Statistician)* **47**(3), 431–443.

22. Burrough, P. & McDonnell, R. (1998), 'Principles of geographical information systems oxford: Oxford university press'.

23. Byrne, G., Charlton, M. & Fotheringham, S. (2009), Multiple dependent hypothesis tests in geographically weighted regression, *in* 'Proceedings of the 10th international conference on GeoComputation', University of New South Wales.

24. Campbell, J. Y. & Cocco, J. F. (2007), 'How do house prices affect consumption? evidence from micro data', *Journal of Monetary Economics* **54**(3), 591–621.

25. Can, A. (1990), 'The measurement of neighborhood dynamics in urban house prices', **66**(3), 254–272.

26. Can, A. (1992), 'Specification and estimation of hedonic housing price models', *Regional science and urban economics* **22**(3), 453–474.

27. Can, A. & Megbolugbe, I. (1997*a*), 'Spatial dependence and house price index construction', *The Journal of Real Estate Finance and Economics* **14**(1-2), 203–222.

28. Can, A. & Megbolugbe, I. (1997*b*), 'Spatial dependence and house price index construction', *The Journal of Real Estate Finance and Economics* **14**(1-2), 203–222.

29. Cannon, S. E. & Cole, R. A. (2011), 'How accurate are commercial real estate appraisals? evidence from 25 years of ncreif sales data', *Journalof Portfolio Management 2011* **35**(5), 1–45.

30. Case, K. E. & Shiller, R. J. (1994), 'A decade of boom and bust in the prices of single-family homes: Boston and los angeles, 1983 to 1993', *New England Economic Review* **5**, 40–51.

31. Casetti, E. (1972), 'Generating models by the expansion method: applications to geographical research', *Geographical analysis* **4**(1), 81–91.

32. Central Statistics Office (CSO) (2023*a*), 'Press statement census 2022 results profile 2 - housing in ireland limerick', https://www.cso.ie/en/csolatestnews/pressreleases/2023pressreleases/pressstatementcensus2022resultsprofile2-housinginirelandlimerick/. Accessed: 20-02-2024.

33. Central Statistics Office (CSO) (2023*b*), 'Residential property price index january 2023', https://www.cso.ie/en/releasesandpublications/ep/p-rppi/residentialpropertypriceindexjanuary2023/. Accessed: 20-02-2024.

34. Cervero, R. & Landis, J. (1993), 'Assessing the impacts of urban rail transit on local real estate markets using quasi-experimental comparisons', **8**, 13–22.

35. Cesa-Bianchi, A. (2013), 'Housing cycles and macroeconomic fluctuations: A global perspective', *Journal of International Money and Finance* **37**, 215–238.

36. Charlton, M., Fotheringham, S. & Brunsdon, C. (2009), 'Geographically weighted regression', *White paper. National Centre for Geocomputation. National University of Ireland Maynooth* **2**.

37. Clapham, D., Buckley, K., Mackie, P., Orford, S. & Stafford, I. (2010), *Young people and housing in 2020: identifying key drivers for change*, Joseph Rowntree Foundation York.

38. Collins, M. L. & Larragy, A. (2011), A site value tax for ireland: approach, design and implementation, Technical report, Trinity College Dublin, Department of Economics.

39. Comber, A., Brunsdon, C., Charlton, M., Dong, G., Harris, R., Lu, B., Lü, Y., Murakami, D., Nakaya, T., Wang, Y. et al. (2023), 'A route map for successful applications of geographically weighted regression', *Geographical Analysis* **55**(1), 155–178.

40. Comber, A., Wang, Y., Lü, Y., Zhang, X. & Harris, P. (2018), 'Hyper-local geographically weighted regression: extending gwr through local model selection and local bandwidth optimization', *Journal of Spatial Information Science* (17), 63–84.

41. Corcoran, C. & Liu, F. (2014), 'Accuracy of zillow's home value estimates', *Real Estate Issues* **39**(1), 45–49.

42. Costello, G. J. (2000), 'Price sizing effects in housing markets', *Journal of Property Research* **17**, 203–219.

43. Cronin, D. & McQuinn, K. (2021), 'House prices and the credit-driven household demand channel: The case of the irish economy', *Credit and Capital Markets* **54**, 199–221.

44. da Silva, A. R. & Fotheringham, A. S. (2016), 'The multiple testing issue in geographically weighted regression', *Geographical Analysis* **48**(3), 233–247.

45. Daft.ie (2017), 'The daft.ie house price report', https://ww1.daft.ie/report/2017-Q4-dafthouseprice-report.pdf?d_rd=1. Accessed: 20-02-2024.

46. Daft.ie (2018), 'The daft.ie house price report', https://ww1.daft.ie/report/2018-Q4-houseprice-daft-report.pdf?d_rd=1. Accessed: 20-02-2024.

47. de Haan, J. & Diewert, E. (2013), '"hedonic regression methods"', *Handbook on Residential Property Price Indices, OECD Publishing, Paris* .

48. Debrezion, G. & Rietveld, E. P. . P. (2007), 'Impact of railway station on dutch residential housing market', **8**, 1–20.

49. Dettling, L. J. & Kearney, M. S. (2014), 'House prices and birth rates: The impact of the real estate market on the decision to have a baby', *Journal of Public Economics* **110**, 82–100.

50. Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T. & Moons, K. G. (2006), 'A gentle introduction to imputation of missing values', *Journal of clinical epidemiology* **59**(10), 1087–1091.

51. Dubin, R. (1992), 'Spatial autocorrelation and neighborhood quality', *Regional science and urban economics* **22**(3), 433–452.

52. Dubin, R. A. (1988), 'Estimation of regression coefficients in the presence of spatially autocorrelated error terms', *The Review of Economics and Statistics* **70**(3), 466–474.

53. Dunne, T. (2012), 'Household formation and the great recession', *Economic commentary* (2012-12).

54. Dupré, D. (2020), Urban and socio-economic correlates of property prices in dublin's area, *in* '2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)', IEEE, pp. 556–562.

55. Dye, R. F. & England, R. W. (2010), 'Assessing the theory and practice of land value taxation', *Policy focus report: Lincoln Institute of Land Policy* pp. 0–32.

56. Eckert, J. K., Gloudemans, R. J. & Almy, R. R. (1990), *Property appraisal and assessment administration*, International Association of Assessing Officers.

57. Egan, P. & McQuinn, K. (2023), 'Regime switching and the responsiveness of prices to supply: The case of the irish housing market', *The Quarterly Review of Economics and Finance* **87**, 82–94.

58. Engstrom, G. & Gren, A. (2017), 'Capturing the value of green space in urban parks in a sustainable urban planning and design context: pros and cons of hedonic pricing', *Ecology and Society* **22**, 94–106.

59. Eurostat (2013), Handbook on residential property prices indices (rppis), Technical report, European Commission.

60. Eurostat (2021), 'House or flat – owning or renting', https://ec.europa.eu/eurostat/cache/digpub/housing/bloc-1a.html. Accessed: 20-02-2024.

61. Farber, S. & Yeates, M. (2006), 'A comparison of localized regression models in a hedonic house price context', *Canadian Journal of Regional Science* **29**(3), 405–420.

62. Fik, T. J., Ling, D. C. & Mulligan, G. F. (2003), 'Modeling spatial variation in housing prices: a variable interaction approach', *Real Estate Economics* **31**(4), 623–646.

63. Fotheringham, A. S., Brunsdon, C. & Charlton, M. (2003), *Geographically weighted regression: the analysis of spatially varying relationships*, John Wiley & Sons.

64. Fotheringham, A. S., Charlton, M. E. & Brunsdon, C. (1998), 'Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis', *Environment and planning A* **30**(11), 1905–1927.

65. Fotheringham, A. S. & Oshan, T. M. (2016), 'Geographically weighted regression and multicollinearity: dispelling the myth', *Journal of geographical systems* **18**, 303–329.

66. Fox, J. & Monette, G. (1992), 'Generalized collinearity diagnostics', *Journal of the American Statistical Association* **87**(417), 178–183.

67. Fox, J. & Weisberg, S. (2011), 'Multivariate linear models in r', *An R Companion to Applied Regression. Los Angeles: Thousand Oaks* .

68. Gabrielli, L., Giuffrida, S. & Trovato, M. R. (2017), 'Gaps and overlaps of urban housing sub-market: hard clustering and fuzzy clustering approaches', *Appraisal: From Theory to Practice: Results of SIEV 2015* pp. 203–219.

69. Gaetan, C., Guyon, X. et al. (2010), *Spatial statistics and modeling*, Vol. 90, Springer.

70. Gallet, C. A. (2004), 'Housing market segmentation: An application of convergence tests to los angeles region housing', *The Annals of Regional Science* **38**, 551–561.

71. Geary, R. C. (1954), 'The contiguity ratio and statistical mapping', *The incorporated statistician* **5**(3), 115–146.

72. Gelfand, A. E., Ecker, M. D., Knight, J. R., & Sirmans, C. F. (2004), 'The dynamics of location in home price', *The Journal of Real Estate Finance and Economics* **29**, 149–166.

73. Gelman, A. & Hill, J. (2006), *Data analysis using regression and multilevel/hierarchical models*, Cambridge university press.

74. Getis, A. (2009), Spatial autocorrelation, *in* 'Handbook of applied spatial analysis: Software tools, methods and applications', Springer, pp. 255–278.

75. Gollini, I., Lu, B., Charlton, M., Brunsdon, C. & Harris, P. (2013), 'Gwmodel: an r package for exploring spatial heterogeneity using geographically weighted models', *arXiv preprint arXiv:1306.0413* .

76. Goodman, A. C. & Thibodeau, T. G. (1998), 'Housing market segmentation', *Journal of housing economics* **7**(2), 121–143.

77. Goodman, A. C. & Thibodeau, T. G. (2003), 'Housing market segmentation and hedonic prediction accuracy', *Journal of Housing Economics* **12**(3), 181–201.

78. Gray, D. (2013), 'House price diffusion: An application of spectral analysis to the prices of irish second-hand dwellings', *Housing Studies* **28**(6), 869–890.

79. Gupta, R. & Miller, S. M. (2012), 'The time-series properties of house prices: A case study of the southern california market', *The Journal of Real Estate Finance and Economics* **44**, 339–361.

80. Haixiao, P. & Ming, Z. (2008), 'Rail transit impacts on land use', **8**, 16–25.

81. Harris, P., Fotheringham, A., Crespo, R. & Charlton, M. (2010), 'The use of geographically weighted regression for spatial prediction: an evaluation of models using simulated data sets', *Mathematical Geosciences* **42**, 657–680.

82. Harris, P., Fotheringham, A. S. & Juggins, S. (2010), 'Robust geographically weighted regression: A technique for quantifying spatial relationships between freshwater acidification critical loads and catchment attributes', *Annals of the Association of American Geographers* **100**(2), 286–306.

83. Hee, C., Bae, C., Myung-Jin, J. & Hyeon, P. (2002), 'The impact of seoul's subway line 5 on residential property values', **8**, 85–94.

84. Helbich, M., Brunauer, W., Hagenauer, J. & Leitner, M. (2013), 'Data-driven regionalization of housing markets', *Annals of the Association of American Geographers* **103**(4), 871–889.

85. Hengl, T., Heuvelink, G. B. & Rossiter, D. G. (2007), 'About regression-kriging: From equations to case studies', *Computers & geosciences* **33**(10), 1301–1315.

86. Hess, D. B. & Almeida, T. M. (2007), 'Impact of proximity to light rail rapid transit on station-area property values in buffalo, new york', **44**, 1–54.

87. Heyman, A. V. & Sommervoll, D. E. (2019), 'House prices and relative location', *Cities* **95**.

88. Hillebrand, M. & Kikuchi, T. (2015), 'A mechanism for booms and busts in housing prices', *Journal of Economic Dynamics and Control* **51**, 204–217.

89. Hoolachan, J., McKee, K., Moore, T. & Soaita, A. (2017), "generation rent' and the ability to 'settle down': economic and geographical variation in young people's housing transitions', *Journal of Youth Studies,* **20**(1), 63–78.

90. Hu, T., Fan, J., Wu, Q. & Zhou, D.-X. (2015), 'Regularization schemes for minimum error entropy principle', *Analysis and Applications* **13**(04), 437–455.

91. Hurley, A. K. & Sweeney, J. (2022), 'Irish property price estimation using a flexible geo-spatial smoothing approach: What is the impact of an address?', *Journal of Real Estate Finance & Economics* .

92. Hwang, S. & Thill, J.-C. (2009), 'Delineating urban housing submarkets with fuzzy clustering', *Environment and Planning B: Planning and Design* **36**(5), 865–882.

93. Irandoust, M. (2019), 'House prices and unemployment: an empirical analysis of causality', *International journal of housing markets and analysis* **12**(1), 148–164.

94. Islam, K. S. & Asami, Y. (2009), Housing market segmentation: A review, *in* 'Review of Urban & Regional Development Studies: Journal of the Applied Regional Science Conference', Vol. 21, Wiley Online Library, pp. 93–109.

95. Jahanshiri, E., Buyong, T. & Shariff, A. R. M. (2011), 'A review of property mass valuation models', *Pertanika Journal of Science & Technology* **19**(1), 23–30.

96. James, G., Witten, D., Hastie, T. & Tibshirani, R. (2023), *An Introduction to Statistical Learning*, 2 edn, Springer.

97. Jing, Y. & Fotheringham, S. A. (2016), 'Local spatiotemporal modeling of house prices: A mixed model approach', *The Professional Geographer* **68**(2), 189–201.

98. John, B. & G.Stacy, S. (2020), 'Mass transportation, apartment rent and property values', **8**, 1–8.

99. Kennedy, G., Myers, S. et al. (2019), 'An overview of the irish housing market'.

100. Kestens, Y., Thériault, M. & Des Rosiers, F. (2006), 'Heterogeneity in hedonic modelling of house prices: looking at buyers' household profiles', *Journal of Geographical Systems* **8**, 61–96.

101. Khanani, R. S., Adugbila, E. J., Martinez, J. A. & Pfeffer, K. (2020), 'The impact of road infrastructure development projects on local communities in peri-urban areas: the case of kisumu, kenya and accra, ghana', **8**, 1–21.

102. Kiel, K. A. & Zabel, J. E. (2008), 'Location, location, location: The 3l approach to house price determination', *Journal of Housing Economics* **17**(2), 175–190.

103. Kitchin, R. (2013), 'Making informed decisions on future housing policy', *Housing Ireland: A Journal for Irish Housing Professionals* **3**, 36–41.

104. Kobylińska, K. (2021), 'The application of spatial autoregressive models for analyzing the influence of spatial factors on real estate prices and values', *Real Estate Management and Valuation* **29**(4), 23–35.

105. Lee, C. L. (2009), 'Housing price volatility and its determinants', *Internation Journal of Housing Markets and Analysis* **2**, 293 – 308.

106. Lennartz, C., Arundel, R. & Ronald, R. (2016), 'Younger adults and homeownership in europe through the global financial crisis', *Population, Space and Place* **22**(8), 823–835.

107. Levokich, O., Rounwendal, J. & va Markwijk, R. (2010), 'The effects of highway development on housing prices', **8**, 1–27.

108. Limerick County Council (LCC) (2022*a*), 'Limerick city facts', https://www.limerick.ie/business/why-limerick/quick-facts. Accessed: 20-02-2024.

109. Limerick County Council (LCC) (2022*b*), 'Limerick development plan 2022-2028'. Accessed: 20-02-2024. **URL:** *https://www.limerick.ie/sites/default/files/media/documents/2022-07/Strategic-Environmental-Assessment-Environmental-Report.pdf/*

110. Limerick County Council (LCC) (2023), 'Limerick city and county council, residential property price index january 2023', https://www.limerick.ie/council/services/your-council/about-us/information-on-limericks-metropolitan-and-municipal/. Accessed: 20-02-2024.

111. Liu, X. (2013), 'Spatial and temporal dependence in house price prediction', *The Journal of Real Estate Finance and Economics* **47**, 341–369.

112. Local Area Electoral Boundary Committe (LEABC) (2018), 'Limerick metropolitan district', https://www.boundarycommittee.ie/Maps/Limerick%20City%20LEA%20Report%20Map%202018.pdf/. Accessed: 20-02-2024.

113. Lu, B., Harris, P., Charlton, M. & Brunsdon, C. (2014), 'The gwmodel r package: further topics for exploring spatial heterogeneity using geographically weighted models', *Geo-spatial Information Science* **17**(2), 85–101.

114. Luttik, J. (2000), 'The value of trees, water and open space as reflected by house prices in the netherlands', *Landscape and Urban Planning* **48**, 161–167.

115. Lyons, R. C. (2015), 'East, west, boom and bust: the spread of house prices and rents in ireland, 2007–2012', *Property Research* **32**, 77–1010.

116. Maguire, P., Miller, R., Moser, P. & Maguire, R. (2016), 'A robust house price index using sparse and frugal data', *Journal of Property Research* **33**(4), 293–308.

117. Mardia, K. V. & Marshall, R. J. (1984), 'Maximum likelihood estimation of models for residual covariance in spatial regression', *Biometrika* **71**(1), 135–146.

118. Mayor, K., Lyons, S., Duffy, D. & Tol, R. S. (2009), A hedonic analysis of the value of parks and green spaces in the dublin area, Technical report, ESRI working paper.

119. McCafferty, D. (2009), 'Aspects of socio-economic development in limerick city since 1970: A geographers perspective'.

120. McCafferty, D. (2011), 'Divided city: the geography of post-celtic tiger limerick'.

121. McCafferty, D. & O'Keeffe, B. (2009), 'Facing the challenge of change: a spatial perspective on limerick'.

122. McKee, K. (2012), 'Young people, homeownership and future welfare', *Housing Studies* **27**(6).

123. Meen, G. (1999), 'Regional house prices and the ripple effect: a new interpretation', *Housing studies* **14**(6), 733–753.

124. Michalis Doumpos, Dimitrios Papastamos, D. A. & Zopounidis, C. (2020), 'Developing automated valuation models for estimating property values: a comparison of global and locally weighted approaches', *Annals of Operations Research* **306**, 415–433.

125. Micheál L. Collins, A. L. (2011), 'Implementing a site value tax for ireland', *Technical Report, Trinity College Dublin, Department of Economics.* .

126. Miller, N., Peng, L. & Sklarz, M. (2011), 'House prices and economic growth', *The Journal of Real Estate Finance and Economics* **42**, 522–541.

127. Mooya, M. (2011), 'Of mice and men: Automated valuation models and the valuation profession', *Urban Studies* **48**, 2211–2440.

128. Moran, P. A. (1950), 'A test for the serial independence of residuals', *Biometrika* **37**(1/2), 178–181.

129. Morgenroth, E. (2014), 'Modelling the impact of fundamentals on county housing markets in ireland'.

130. Myers, R. H. & Myers, R. H. (1990), *Classical and modern regression with applications*, Vol. 2, Duxbury press Belmont, CA.

131. Mykyta, L. (2012), 'Economic downturns and the failure to launch: The living arrangements of young adults in the us 1995-2011', *US Census Bureau Social, Economic and Housing Statistics Division (SEHSD) Working Paper* **24**.

132. Nguyen, D. (2019), 'Automated valuation models (avms) in commercial real estate market', *Aalto University School of Engineering Master's Programme in Real Estate Economics (REC)* **1**, 63 – 77.

133. Oates, W. E. & Schwab, R. M. (1997), 'The impact of urban land taxation: The pittsburgh experience', *National Tax Journal* **50**, 1–21.

134. O'Hanlon, N. (2011), 'Constructing a national house price index for ireland.', *Journal of the Statistical & Social Inquiry Society of Ireland* **40**.

135. Osborne, J. W. & Waters, E. (2019), 'Four assumptions of multiple regression that researchers should always test', *Practical assessment, research, and evaluation* **8**(1), 2.

136. Oust, A., Hansen, S. N. & Pettrem, T. R. (2020), 'Combining property price predictions from repeat sales and spatially enhanced hedonic regressions', *The Journal of Real Estate Finance and Economics* **61**, 183–207.

137. Pace, K. (1998), 'Appraisal using generalized additive models', *Journal of Real Estate Research* **15**(1), 77–99.

138. Pace, R. & Gilley, O. (1997), 'Using the spatial configuration of the data to improve estimation', *The Journal of Real Estate Finance and Economics* **14**, 333–340.

139. Pace, R. K. & Barry, R. (1997), 'Quick computation of spatial autoregressive estimators', *Geographical analysis* **29**(3), 232–247.

140. Pace, R. K., Barry, R. & Sirmans, C. F. (1998), 'Spatial statistics and real estate', *The Journal of Real Estate Finance and Economics* **17**, 5–13.

141. Pandit, K. & Casetti, E. (1989), 'The shifting patterns of sectoral labor allocation during development: developed versus developing countries', *Annals of the Association of American Geographers* **79**(3), 329–344.

142. Panduro, T. E. & Thorsen, B. J. (2014), 'Evaluating two model reduction approaches for large scale hedonic models sensitive to omitted variables and multicollinearity', *Letters in Spatial and Resource Sciences* **7**, 85–102.

143. Pebesma, E. J. & Bivand, R. (2005), 'Classes and methods for spatial data in R', *R News* **5**(2), 9–13.

144. Peng, Z. & Inoue, R. (2022), 'Identifying multiple scales of spatial heterogeneity in housing prices based on eigenvector spatial filtering approaches', *ISPRS International Journal of Geo-Information* **11**(5), 283.

145. Plantinga, A. J., Détang-Dessendre, C., Hunt, G. L. & Piguet, V. (2013), 'Housing prices and inter-urban migration', *Regional Science and Urban Economics* **43**(2), 296–306.

146. Priemus, H. & Whitehead, C. (2014), 'Interactions between the financial crisis and national housing markets', *Journal of Housing and the Built Environment* **29**, 193–200.

147. Pryce & Gwilym (2013), 'Housing submarkets and the lattice of substitution', *Urban Studies* **50**(13), 2682–2699.

148. R. Webster, M. O. (2001), *Geostatistics for Environmental Scientists Statistics in Practice,*, Wiley, p. 271.

149. Rainer Schulz, M. W. & Werwatz, A. (2014), 'Automated valuation modelling: a specification exercise', *Journal of Property Research* **31**, 131–153.

150. Robusto, C. C. (1957), 'The cosine-haversine formula', *The American Mathematical Monthly* **64**(1), 38–40.

151. Roche, M. J. (2001), 'The rise in house prices in dublin: bubble, fad or just fundamentals', *Economic Modelling* **18**(2), 281–295.

152. Rosen, S. (1974), 'Hedonic prices and implicit markets: product differentiation in pure competition', *Journal of political economy* **82**(1), 34–55.

153. Scheurwater, S. (2017), 'The future of valuations', *RICS Insight Paper* .

154. Schnare, A. B. & Struyk, R. J. (1976), 'Segmentation in urban housing markets', *Journal of Urban Economics* **3**(2), 146–166.

155. Shekhar, S., Jiang, Z., Ali, R. Y., Eftelioglu, E., Tang, X., Gunturi, V. M. & Zhou, X. (2015), 'Spatiotemporal data mining: A computational perspective', *ISPRS International Journal of Geo-Information* **4**(4), 2306–2338.

156. Shimizu, C., Karato, K. & Nishimura, K. (2014), 'Nonlinearity of housing price structure: Assessment of three approaches to nonlinearity in the previously owned condominium market of tokyo', *International Journal of Housing Markets and Analysis* **7**(4), 459–488.

157. Shinnick, E. (1997), 'Measuring irish housing quality', *Journal of Economic Studies* **24**(1/2), 95–119.

158. Simon, J. & Stone, T. (2017), 'The property ladder after the financial crisis: The first step is a stretch but those who make it are doing ok', *Reserve Bank of Australia* .

159. Sisman, S. & Aydinoglu, A. C. (2022), 'A modelling approach with geographically weighted regression methods for determining geographic variation and influencing factors in housing price: A case in istanbul', *Land Use Policy* **119**, 106183.

160. Stefanakis, E. (2017), 'Web mercator and raster tile maps: two cornerstones of online map service providers', *Geomatica* **71**(2), 100–109.

161. Stevenson, S. (2004), 'House price diffusion and inter-regional and cross-border house price dynamics', *Journal of Property Research* **21**(4), 301–320.

162. Straszheim, M. (1974), 'Hedonic estimation of housing market prices: A further comment', *The review of economics and statistics* pp. 404–406.

163. Tailte Éireann (2024), 'National statutary boundaries', https://data.gov.ie/organization/ordnance-survey-ireland. Accessed: 20-02-2024.

164. Taylor, L. O. (2008), Theoretical foundations and empirical developments in hedonic modeling, *in* 'Hedonic methods in housing markets: Pricing environmental amenities and segregation', Springer, pp. 15–37.

165. Theebe, M. A. (2004), 'Planes, trains, and automobiles: the impact of traffic noise on house prices', *The Journal of Real Estate Finance and Economics* **28**, 209–234.

166. Thompson, J., Roecker, S., Grunwald, S. & Owens, P. (2012), 'Digital soil mapping: interactions with and applications for hydropedology'.

167. Tobler, W. R. (1970), 'A computer movie simulating urban growth in the detroit region', *Economic geography* **46**(sup1), 234–240.

168. Troy, A. & Grove, J. M. (2008), 'Property values, parks, and crime: A hedonic analysis in baltimore, md', *Landscape and urban planning* **87**(3), 233–245.

169. Tu, Y., Sun, H. & Yu, S.-M. (2007), 'Spatial autocorrelations and urban housing market segmentation', *The Journal of Real Estate Finance and Economics* **34**, 385–406.

170. van den Berg, S. M. (2021), 'Analysing data using linear models'.

171. W. Fuller, G., Johnston, A. & Regan, A. (2020), 'Housing prices and wealth inequality in western europe', *West European Politics* **43**(2), 297–320.

172. Wang, X., Wen, J., Zhang, Y. & Wang, Y. (2014), 'Real estate price forecasting based on svm optimized by pso', *Optik* **125**, 1439–1443.

173. Watkins, C. A. (2001), 'The definition and identification of housing submarkets', *Environment and Planning A* **33**(12), 2235–2253.

174. Weinstein, T. L. C. B. L. (2002), 'The impact of dallas (texas) area rapid transit light rail stations on taxable property valuations', **8**, 1–54.

175. Weisberg, S. & Cook, R. D. (1982), 'Residuals and influence in regression'.

176. Wheeler, D. C. & Calder, C. A. (2007), 'An assessment of coefficient accuracy in linear regression models with spatially varying coefficients', *Journal of Geographical Systems* **9**, 145–166.

177. Wheeler, D. C. & Páez, A. (2009), Geographically weighted regression, *in* 'Handbook of applied spatial analysis: Software tools, methods and applications', Springer, pp. 461–486.

178. Wood, S. N., Bravington, M. V. & Hedley, S. L. (2008), 'Soap film smoothing', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **70**(5), 931–955.

179. Yunjong Kim, S. C. & Y, M. Y. (2020), 'Applying comparable sales method to the automated estimation of real estate prices', *Sustainability* **12**.

180. Zhou, C. (2021), 'House price prediction using polynomial regression with particle swarm optimization', *Journal of Physics: Conference Series* **1802**.