A Comparative Analysis of Machine Learning Methods and Spatial Statistical Methods for Areal Unit Scottish Property Price Data

Cara Margaret MacBride

School of Mathematics and Statistics

University of Glasgow

A thesis submitted for the degree of

Master of Statistics by Research

September 2023

Abstract

Spatial areal unit data are a type of spatial data which consist of a set of contiguous non-overlapping areal units in space, one example being Data Zones (DZ) in Scotland. A special feature about these data is that they are spatially correlated. This means that pairs of areal units that are close to each other in space have more similar data values and structure to one another than areal units that are further apart. In general, spatial data are modelled using classical spatial statistical methods that account for spatial correlation within the data. One widely established spatial method being the conditional autoregressive (CAR) model where spatial correlation is modelled through a set of random effects. However, in recent years, the application of machine learning (ML) methods to spatial data in order to generate predictions has risen in popularity. Unlike spatial methods, machine learning methods can account for non-linear effects. This results in two important questions of interest: (i) Are classical spatial statistical methods or a-spatial machine learning methods best for prediction of spatial areal unit data? and (ii) Can machine learning methods and spatial methods be combined as one to improve predictive performance compared to using the two methods in isolation? By partitioning the data into training and test sets and evaluating predictions using prediction metrics, this MSc addresses these questions in the context of property prices at the Data Zone level in Scotland. In general, I found that there was little difference between spatial methods and machine learning methods in terms of prediction and the combination of both also had a very similar predictive performance.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In statistics, as well as interpreting and analysing data, a common goal is also to make predictions of unknown quantities using that data. A statistical model that is able to make accurate predictions with appropriate uncertainty quantification can play an important role in a variety of day-to-day scenarios, allowing people to predict unknown values, potentially in the future, and prepare or plan accordingly for different eventualities. Many industries rely on predictions made by statisticians in order make effective decisions that will lead to improved business prospects and profits. The predictive models that are built by statisticians can use a variety of complex of statistical methods, in order to make accurate predictions on missing observations in a data set. As there is such a wide range of methods and data sets, it is essential to compare different methods and evaluate which method is most suitable for each particular type of data set.

One industry where predictive models are very beneficial is the real estate industry. It is the job of an estate agent to assess and value a property as accurately as possible based on it's characteristics and spatial location before it is put on the market either to rent or for sale. They play a crucial role in making sure that properties are sold to buyers at realistic prices and also that sellers receive a fair price for what their property is worth (Rightmove, 2023). Therefore, it is important that statisticians investigate and understand spatial patterns in property price data so they can identify which methods most accurately predict property sub-markets and estate agents are able to carry out their job successfully.

Traditionally, when presented with areal unit data and faced with the challenge of making predictions, the general approach would be to model the data using spatial methods. Spatial methods rely on using the spatial structure of the areal units to model corre-

lation by a set of random effects (Lee and Mitchell, 2013). Examples of spatial methods can include simultaneous autoregressive (SAR) models and the popularly used conditional autoregressive (CAR) models (Dormann et al., 2007). CAR models can be fitted within a Bayesian setting using both Markov Chain Monte-Carlo simulation (MCMC) and Integrated Nested Laplace Approximation (INLA) (Rue et al., 2009), the latter being possible due to the invention of the INLA package in R (R INLA Project, 2020).

Another tool which has become increasingly popular in recent years to generate predictions is the use of tree-based machine learning methods, with some examples being bagging, random forest and gradient boosting (Boehmke and Greenwell, 2019). These are all ensemble methods with different strengths which are based on a set of basic decision trees. A single decision tree is an algorithm which splits the data into various subgroups which have similar response values through the use of splitting rules. When multiple decision trees are combined through the ensemble methods mentioned above, they usually perform much better than single trees in prediction. Contrary to the spatial methods, these machine learning methods do not account for any spatial correlation in the data but instead are effective in fitting complex non-linear relationships between features and a target (Boehmke and Greenwell, 2019).

It must be noted that there has not been much research carried out on spatial and machine learning methods in the context of predicting areal unit data, however as mentioned above these methods have different strengths. Therefore, a good idea would be to fuse the strengths of both methods together as this could produce an optimal method for prediction. There are various possible approaches in which these methods could be combined, however the most popular current approach is the geographically weighted random forest (GWRF) method proposed in Georganos et al. [2021]. This consists of constructing a unique random forest for each known areal unit, and using these random forests to make predictions on nearby areal unit with unknown response values.

To test the comparative predictive performance of the methods from the three paradigms, spatial, machine learning and the GWRF, I am going to use data published by the Scottish Government on residential property transactions in 2018 (Scottish Government, 2021). This data set consists of 6,881 observations on 30 variables. Of the 30 variables, there is one target variable, property price, and 29 other variables are features which relate to the spatial location and property characteristics. The data will be partitioned into training and test sets and predictions will be made on the test set using the training set. By using prediction metrics such as root mean square error (RMSE) and median absolute

error (MAE), the performance of each of the methods will be measured and compared so the overall best method for prediction in the context of property price will be able to be determined.

1.1 Aims and Objectives

In this study we will use statistical methods to model spatial areal unit data on property prices at the Data Zone level in Scotland. There are three primary aims of this study:

- To quantify how well average property prices in small areas can be predicted in Scotland.
- To explore if classical spatial modelling approaches or a-spatial machine learning methods provide better prediction of property prices.
- To find out if the spatially adjusted GWRF method can improve prediction compared to the simpler approaches outlined above.

In addition, within these aims I answer the following important questions of interest:

- Through spatial statistical modelling, is there spatial correlation between property prices in neighbouring Data Zones in Scotland after the effects of the features have been accounted for?
- Which features of the data set are the most important for predicting property prices?

1.2 Thesis Structure

Firstly, before constructing any prediction models, Chapter 2 will explore the characteristics of the data set and determining if there are any interesting relationships between the covariates and the property price variable. I will also use data splitting techniques to partition the data into training and test splits to maintain consistency throughout the thesis. On the training set, I will use a cross validation approach to tune each of the models and determine the optimal tuning parameter combinations for each method I investigate. Then, I will construct a normal linear model and assess its predictive performance to see the impact that covariates have on determining property price. This model will be used as a baseline for the more complex models to outperform.

In Chapter 3, the concept of spatial prediction will be introduced and property price predictions will be made using spatial CAR models. The predictive ability of the CAR model will be evaluated using prediction metrics and it will be compared against the linear model to show the impact that accounting for spatial structure has on a model's accuracy.

Chapter 4 will discuss various tree-based machine learning methods starting with the foundation of tree-based methods, the basic decision tree, and following on with three ensemble methods; bagging, random forest and gradient boosting. Property price predictions will be made using each of these methods allowing their respective predictive abilities to be compared with one another and also with the spatial CAR model and linear model.

After studying the different methods in isolation in Chapters 3 and 4, Chapter 5 will explore the combination of spatial and machine learning methods through a method recently proposed, the geographically weighted random forest (Georganos et al., 2021). Again, property price predictions will be made and the results will show whether spatial and machine learning methods perform better alone or if combining them as one method will produce an optimal model for prediction. At the end of this chapter, the final conclusion will be made on whether classical spatial models or a-spatial machine learning methods are better at predicting property prices in Scotland or if using spatially adjusted machine learning methods can outperform both of these methods.

Finally, Chapter 6 will provide an overall summary of each of the methods of prediction applied to the data set and their results. Furthermore, I will give a short critical evaluation of my thesis and propose some improvements that could be made in future research of this topic if time was not limited.

Chapter 2

Data and exploratory analysis

Firstly, I describe the data set and undertake exploratory analysis to identify the key patterns in the data. All analysis uses the R statistical software system (R Core Team, 2022). The data set has 6,881 observations on 30 variables.

2.1 Study Region

This study is based in Scotland, United Kingdom in 2018. The year 2018 was chosen as it is the most recent publication of small-area average property price data in Scotland. Scotland has been split up into 6,976 small non overlapping areal units called Data Zones. These are essentially spatial footprints of small geographical polygons of around 500-1,000 people of similar socioeconomic backgrounds which all fit together to create the map of Scotland (Scottish Government, 2004). They were initially generated by the Scottish Government in 1991 then finalised and formally introduced in 2002 after the 2001 Census (Scottish Government, 2004), and further updated in 2011. Each of these Data Zones belongs to one of the 32 Local Authorities in Scotland. Local Authorities are small local governments who are in charge of public services such as schools, transport, roads etc within their boundary (Scottish Government, 2017). This would suggest that Local Authorities may play an important role in impacting property prices. For this study, due to their small numbers of Data Zones (which makes prediction difficult) and the fact that they do not have a physical land border with the rest of Scotland (which makes spatial modelling difficult), the island local authorities of Shetland, Orkney and Western Isles have been removed from the analysis. This concludes the final data set as 6,881 Data Zones from 29 Local Authorities, covering mainland Scotland.

As previously mentioned there is a possibility that local authority could have an ef-

fect on property price. This is because of certain factors such as education, council tax, amenities and geographical location that are local authority specific. Local authorities that boast successful exam results and provide good schools leading to pupils having better job prospects in the future will have higher property prices as the demand from parents to get their children into a good school is very high (Marshall, 2013). Furthermore, the better the transport links to cities, the more likely people are to purchase a property as it makes their daily life more accessible, thus an area with good transport will be more desirable to live in than an area further out in the countryside (Brown, 2022).

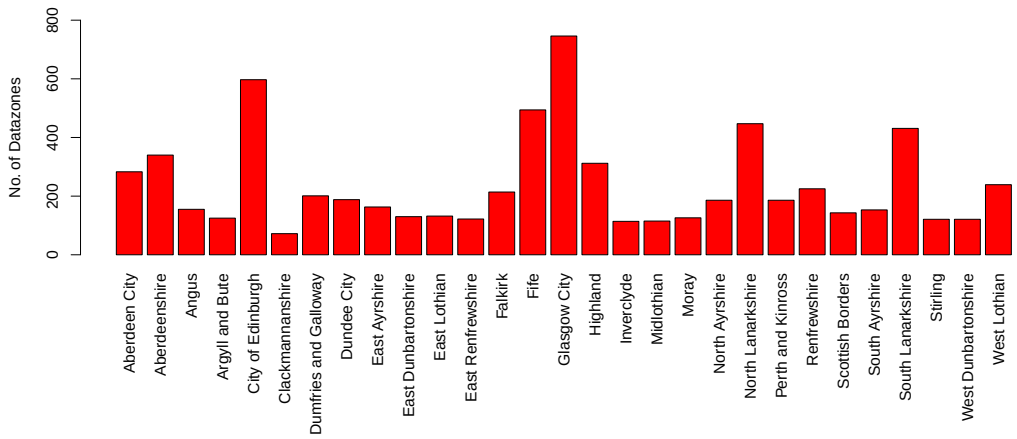

Figure 2.1: Bar plot showing the number of Data Zones within each local authority in Scotland.

Figure 2.1 shows that there is sizeable variation in the numbers of Data Zones in each local authority across Scotland, with the majority of local authorities having between around 100 to 250 Data Zones. Glasgow City has the highest number of Data Zones (746) followed by the City of Edinburgh which has 597. These areas despite not having very large areas in square kilometres (Scottish Government, 2012), have the highest numbers of Data Zones in them due to how densely populated they are. On the other hand, Clackmannanshire has the least with only 72 Data Zones, this is over 10 times less the number in Glasgow City.

2.2 Property Price

The prediction variable for this study is the average *Property Price* of all sold properties in each Data Zone in 2018 which was extracted from a 2018 Scottish Government publication on residential property transactions recorded by Registers of Scotland (Scottish

Government, 2021). This publication provides various different types of information regarding property prices. For this study the median price of all properties sold in 2018 in each Data Zone was used rather than the mean, as it prevented skewness due to small numbers of high or low priced properties. Furthermore, Data Zones with less than 5 properties sold were suppressed from the data set because of the risk of breaching confidentiality and also so that there were at least 5 different property prices used to calculate the average. There was however, one Data Zone which had an average sold property price of £600. This value was deemed as unknown due to the likelihood of there being an error in its calculation.

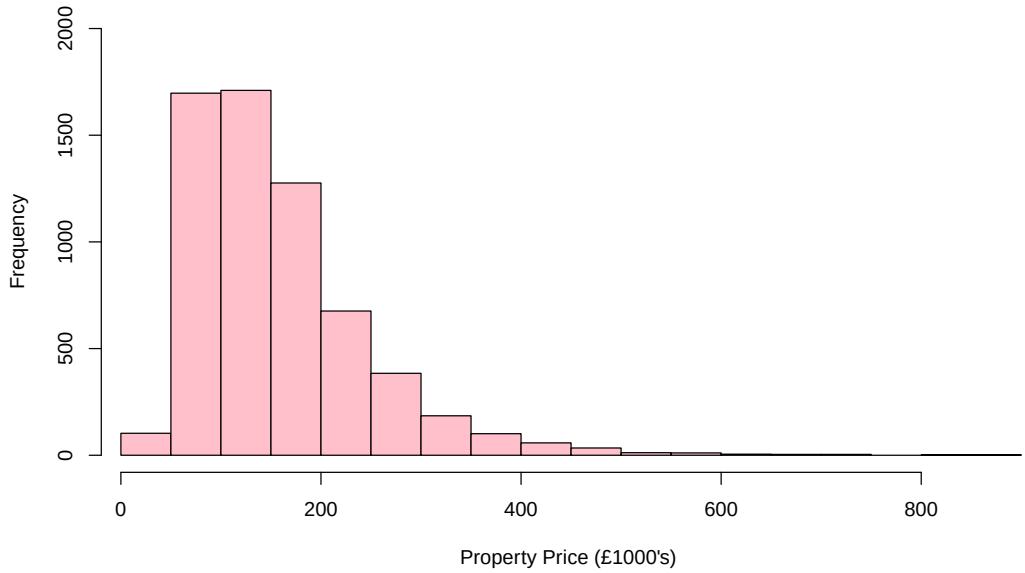

Figure 2.2: Histogram of sold property prices across Scotland in 2018. Along the x axis is the value of the property price in £'s and along the y axis is the number of data zones whose average sold property prices corresponds to this.

Looking at the whole of Scotland, the minimum average property price for a Data Zone is £19,500 while the maximum is £878,000. The mean property price is £159,056 while the median is lower at £139,282. Figure 2.2 is a histogram of the distribution of average property prices across Scotland and gives a visualisation of the data distribution. It shows obvious right skewness and no symmetrical pattern. This suggests that it is very unusual to have high sold property prices and that most properties are valued at between around £50,000 and £250,000. There is a peak at around £150,000 indicating that this is the most common property price in Scotland, the mode. It is important to note that these data include all different kinds of properties from flats to detached houses and from

cities to rural areas.

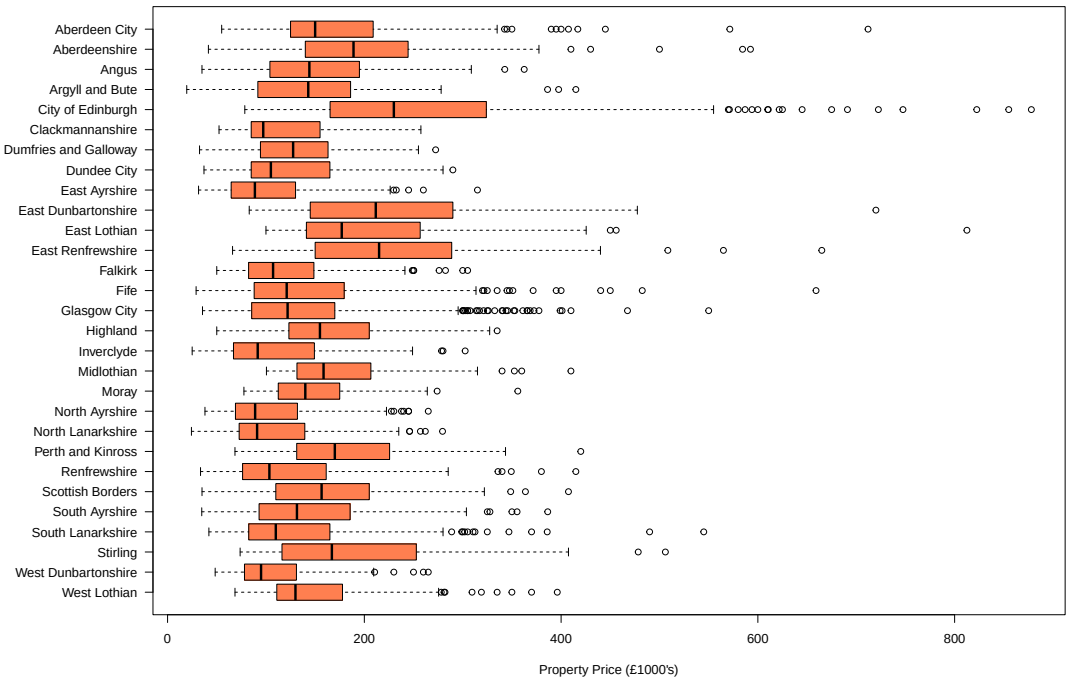

Figure 2.3: Boxplots of the sold property prices across all 29 Local Authorities in mainland Scotland in 2018.

When looking closely at Figure 2.3, it is clear that 26 of the 29 local authorities across Scotland have a median property price of under £200,000. There are only 3 local authorities- City of Edinburgh, East Dunbartonshire and East Renfrewshire where the median is over £200,000. However, these three local authorities have the largest interquartile ranges - £158,834, £143,500 and £138,822 respectively. This indicates that the property prices are more dispersed in these areas. Figure 2.3 also shows that there are a sizeable number of outliers on the right hand side in nearly every local authority. This highlights the variation in property prices within every local authority in Scotland. The highest property values seem to be from local authorities situated in or around Glasgow and Edinburgh, which are the two major cities in Scotland where there will be lots of flats and apartments in the city centres valued for a lot more than a house in a rural community for example. Clackmannanshire is the only local authority that has no outliers. It also has a low interquartile range of £70,000 in comparison to Scotland, which is £100,508, emphasising that the majority of it's properties are probably of similar prices. Contrary to this, the City of Edinburgh and Glasgow City local authorities have a large number of outliers. When looking at the spatial map of Glasgow City in Figure 2.4 we see it is mainly purple hence lower property prices and the corresponding boxplot in Figure 2.3 further supports this. However, there are some areas such as Merrylee, Newlands,

Pollokshields and the West End which are in orange meaning they have higher average property prices. As these are such small areas in comparison to the rest of Glasgow City, this could be why these areas could be considered outliers. On the other hand, the City of Edinburgh seen in Figure 2.5, which still has a good number of purple shaded areas, has slightly more higher priced areas than Glasgow City. This explains why there are lots of outliers towards the more expensive end of the property ladder - The Grange and Calton Hill are home to some of the most expensive properties in the city. Similar to Glasgow City, these are small areas in comparison to the big picture – one Data Zone only consisting of 3 streets yet the most expensive – hence this is why they could be considered outliers. All in all, when looking in more detail at the Data Zones within local authorities, the spatial pattern in property prices can be visualized in the spatial map of Scotland in Figure 2.6, where we see the vast majority of Data Zones in Scotland having average sold property prices of £300,000 or less.

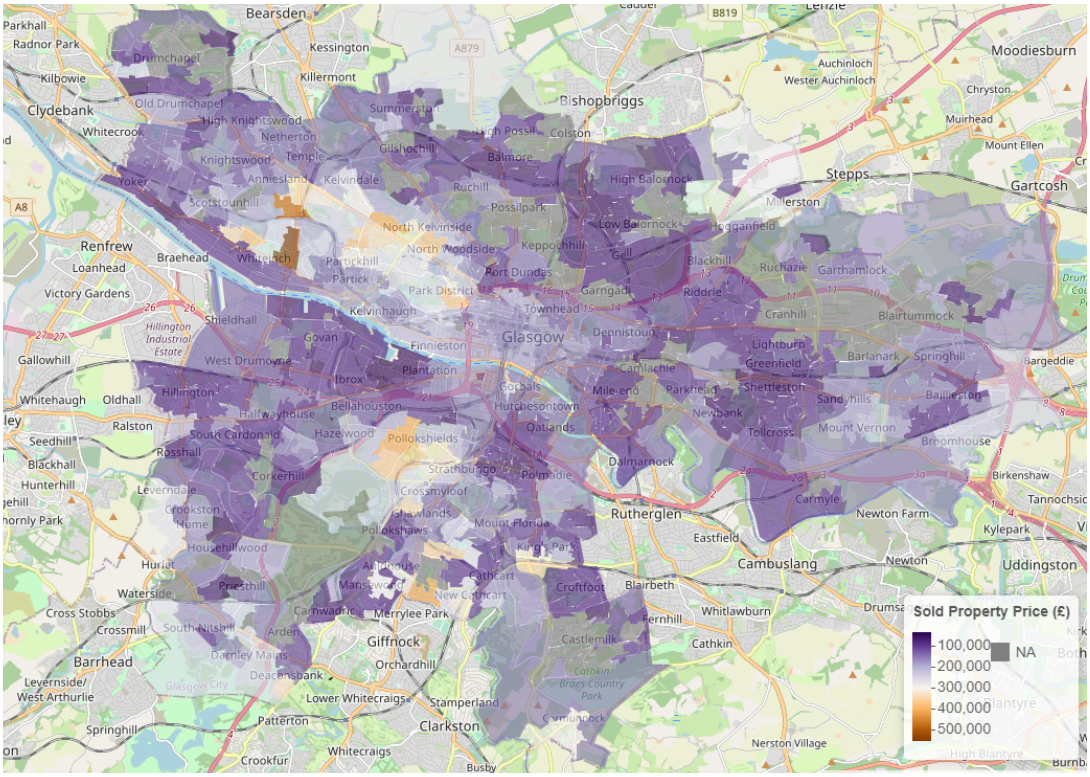

Figure 2.4: Spatial map of the average property price in the Glasgow City local authority that consists of 746 Data Zones. Along the right hand side there is a scale which represents the value of the average sold property price.

Altogether of the 6,881 Data Zones, we have 617 Data Zones that have missing property price values. This means that in 617 Data Zones there were either no properties sold or less than 5 properties sold in 2018. The latter is the threshold used for suppressing

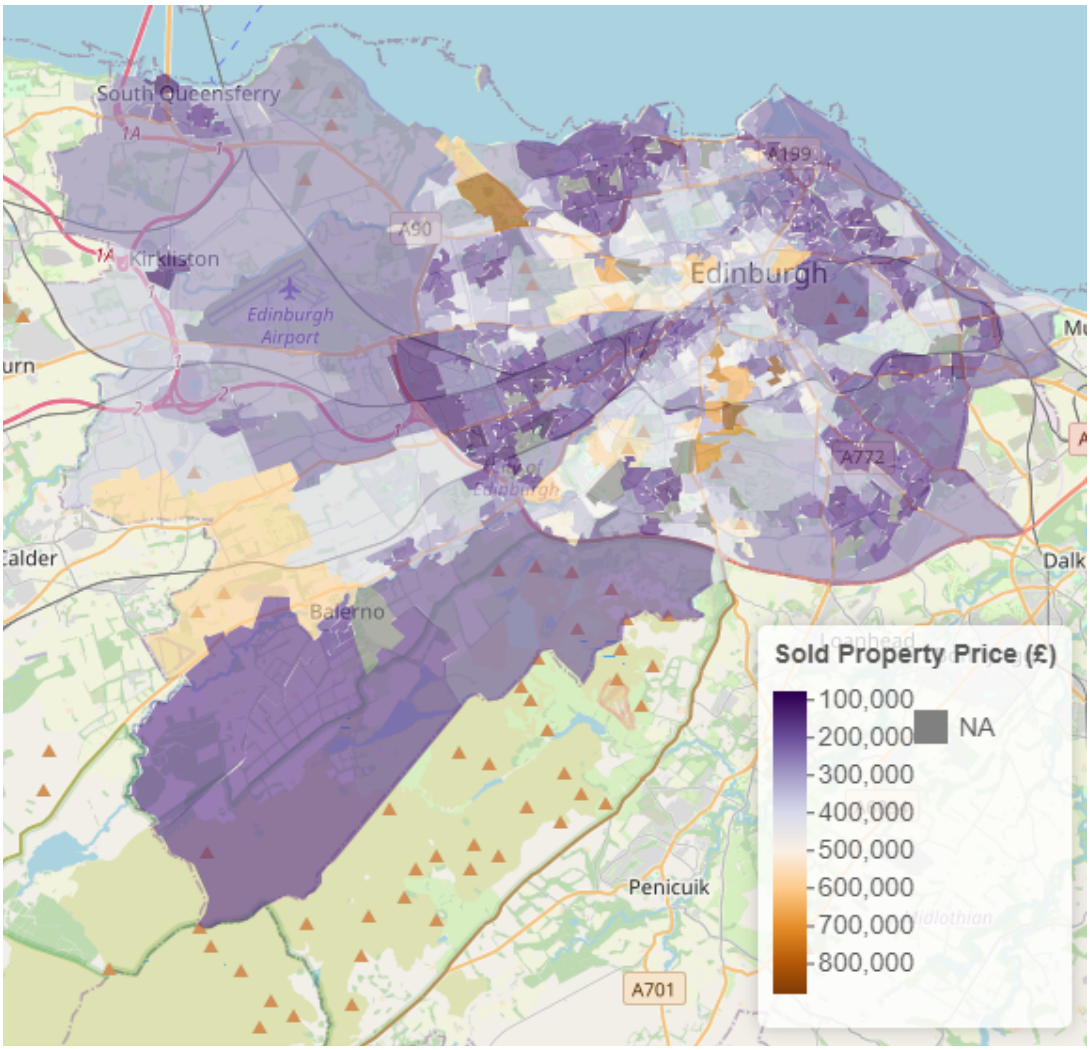

Figure 2.5: Spatial map of the average property price in the City of Edinburgh local authority that consists of 597 Data Zones. Along the right hand side there is a scale which represents the value of the average sold property price.

the data due to the risk of confidentiality by being able to identify individual property sales. The average percentage of missing values in a local authority is approximately 9%. There are 8 local authorities whose percentages are above this value including 3 out of 4 of the local authorities with the largest number of Data Zones.

Although Glasgow City has the most missing values, 125, and Clackmannanshire the least, 3, when finding the missing values in each local authority as a percentage of all values we get slightly different results shown in Table 2.1. We see the largest percentage missing is indeed Glasgow City with around 16.8% while East Lothian has the smallest percentage missing values, approximately 3%. There are many factors which could contribute to this including that Glasgow City takes into account the city centre, this is

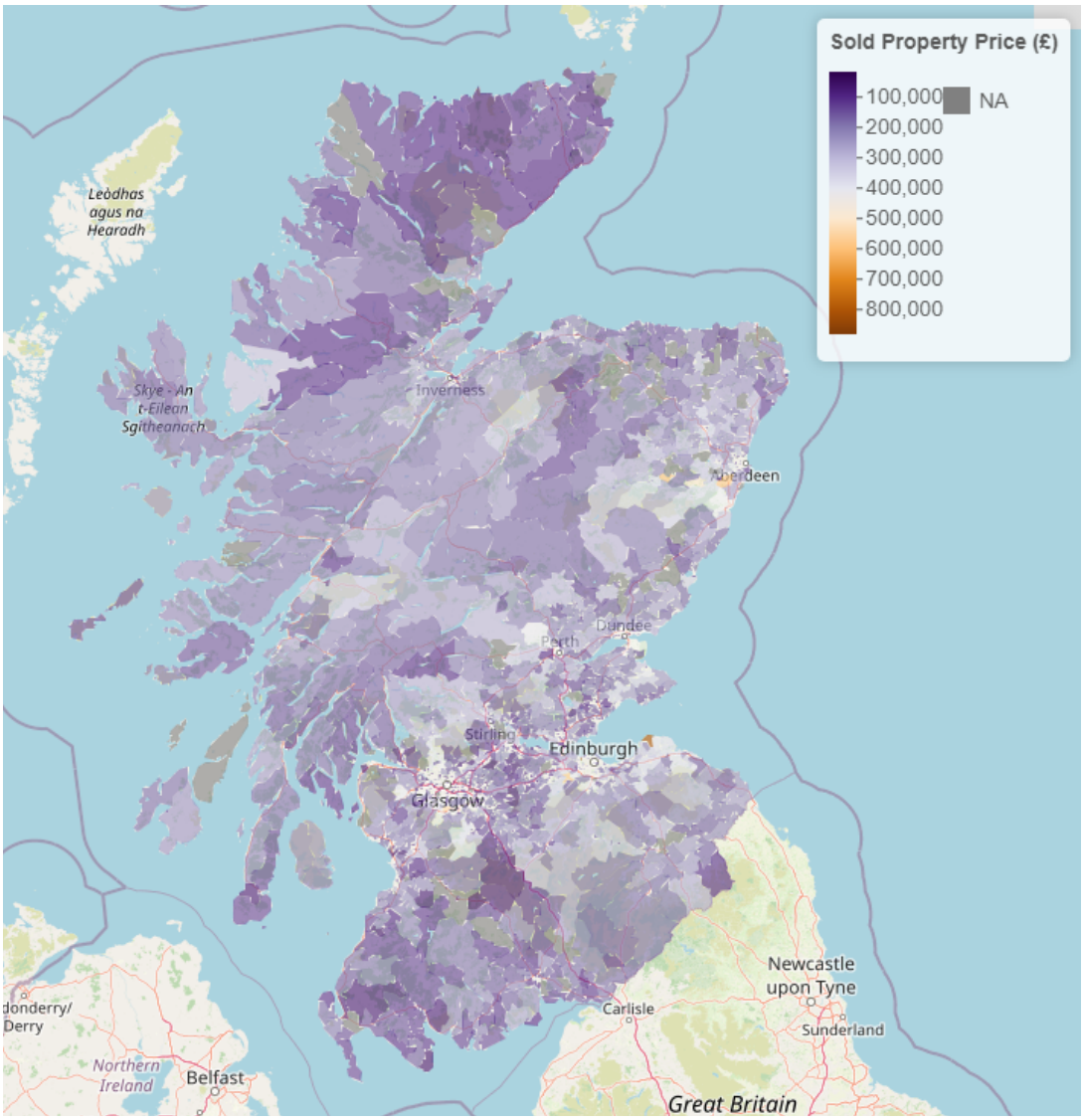

Figure 2.6: This is a spatial map of the average property price in Scotland divided up into the 6881 Data Zones. Along the right hand side there is a scale which represents the value of the average sold property price.

where there are lots of businesses, hotels and education buildings such as Universities. This means fewer properties will be sold in these areas because these buildings tend to be occupied by the same business or landlord for a long period of time. Furthermore, in the city there are many people renting properties as they cannot afford to buy and this renting is not taken into account as the landlord is still owning the property. The deprivation level in Glasgow City is around 44% as of 2020 (Scottish Government, 2012). This was measured using the Scottish Index of Multiple Deprivation (SIMD) (Scottish Government, 2020a) which means that 44% of the residents of Glasgow City Council are living in the bottom 20% of all Data Zones by poverty in Scotland. This is another reason which could explain why the percentage of missing values is the highest, as many of the

Table 2.1: Table of the 29 local authorities and their corresponding numbers of Data Zones, number of missing values and percentage of missing values.

Local Authority	Total Values	Missing Values	Missing Percentage(%)
Aberdeen City	283	26	9.19
Aberdeenshire	340	29	8.53
Angus	155	11	7.10
Argyll and Bute	125	9	7.20
City of Edinburgh	597	54	9.05
Clackmannanshire	72	3	4.17
Dumfries and Galloway	201	16	7.96
Dundee City	188	13	6.91
East Ayrshire	163	11	6.75
East Dunbartonshire	130	4	3.08
East Lothian	132	4	3.03
East Renfrewshire	122	5	4.10
Falkirk	214	18	8.41
Fife	494	28	5.67
Glasgow City	746	125	16.76
Highland	312	24	7.69
Inverclyde	114	16	14.04
Midlothian	115	19	16.52
Moray	126	17	13.49
North Ayrshire	186	13	6.99
North Lanarkshire	447	64	14.32
Perth and Kinross	186	11	5.91
Renfrewshire	225	14	6.22
Scottish Borders	143	5	3.50
South Ayrshire	153	12	7.84
South Lanarkshire	431	18	4.18
Stirling	121	10	8.26
West Dunbartonshire	121	19	15.70
West Lothian	239	19	7.95

people living in these Data Zones cannot afford to buy a property.

2.3 Covariates

This data set consists of 29 covariates describing the characteristics of the area and also the housing stock in each area, and are summarised in the three groups below - Physical Geography, Characteristics of the Area and Property Type Characteristics.

By investigating correlation plots we are able to see if there is a linear (or non-linear)

relationship between each of the covariates and the price variable. Firstly, the correlation coefficient is a number between -1 and 1 which will show how strong or weak a linear relationship between two variables is. A value close to -1 means there is a strong negative linear relationship while if it is close to 1 there is a strong positive linear relationship. In Figures 2.7, 2.8 and 2.9 there are scatterplots in the final row which show the relationship between property prices and each of the other variables. Along the diagonal is the density of each variable while the other scatterplots show collinearity between covariates.

2.3.1 Property Type Characteristics

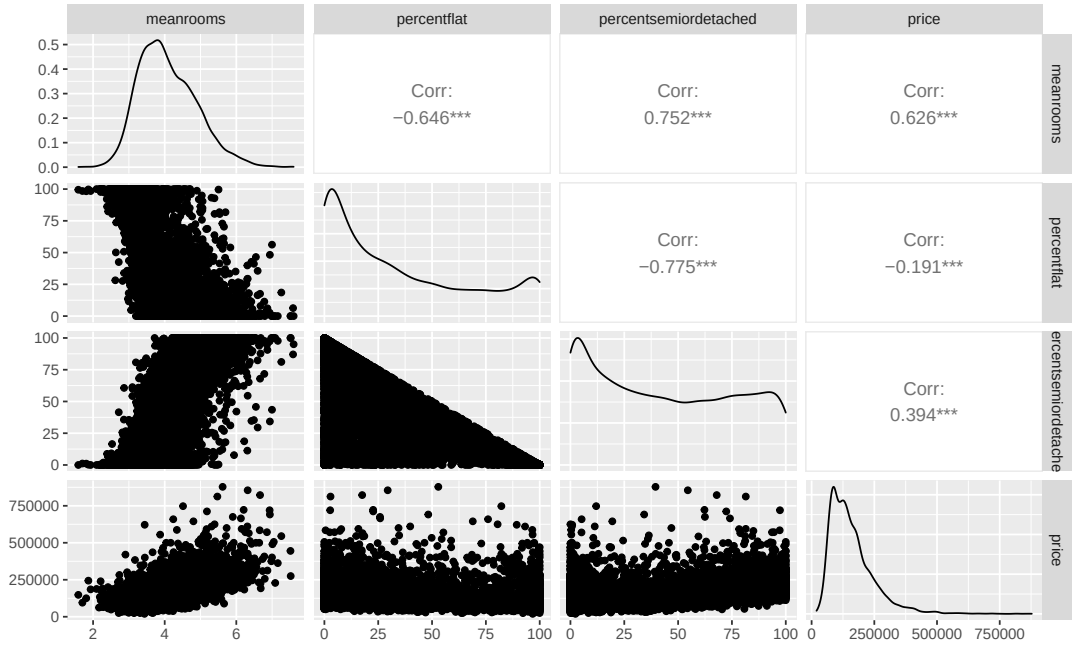

Figure 2.7: Correlation plot showing the relationship between the variables *price*, *mean rooms*, *percentage of flats* and *percentage of semi or detached* using scatterplots, density plots and correlation coefficents.

Firstly, Property Type Characteristics help to give a summary of the properties themselves. Covariates describing the housing stock include the mean number of rooms, percentage of flats and percentage of semi detached and detached properties. Mean rooms calculates an average number of rooms for all properties, excluding kitchens and bathrooms, in each Data Zone. We see in Figure 2.7 that, as expected, as the price of a property increases it seems that the average number of rooms it has increases also. However we do see some lower numbers of mean rooms towards the higher end of the price axis. This would possibly be because of where the property is situated - a smaller house in the west end of Glasgow is a lot more expensive than a bigger house in Queens Park for example. This is due to the fact that the West End of Glasgow is probably overall

a more desirable place to live in than Queens Park. Next, the percentage of flats and the property prices in an area seem to follow a slight U shape. This suggests that there are lots of areas with small percentages of flats that have high property prices but also areas with small percentages of flats with low property prices. Lastly the percentage of semi detached and detached properties seems to follow a positive linear relationship with property price because of the very slight gradient seen in the scatterplot. This means that in general, as the percentage of semi-detached and detached properties increase in an area, the property prices increase.

2.3.2 Physical Geography

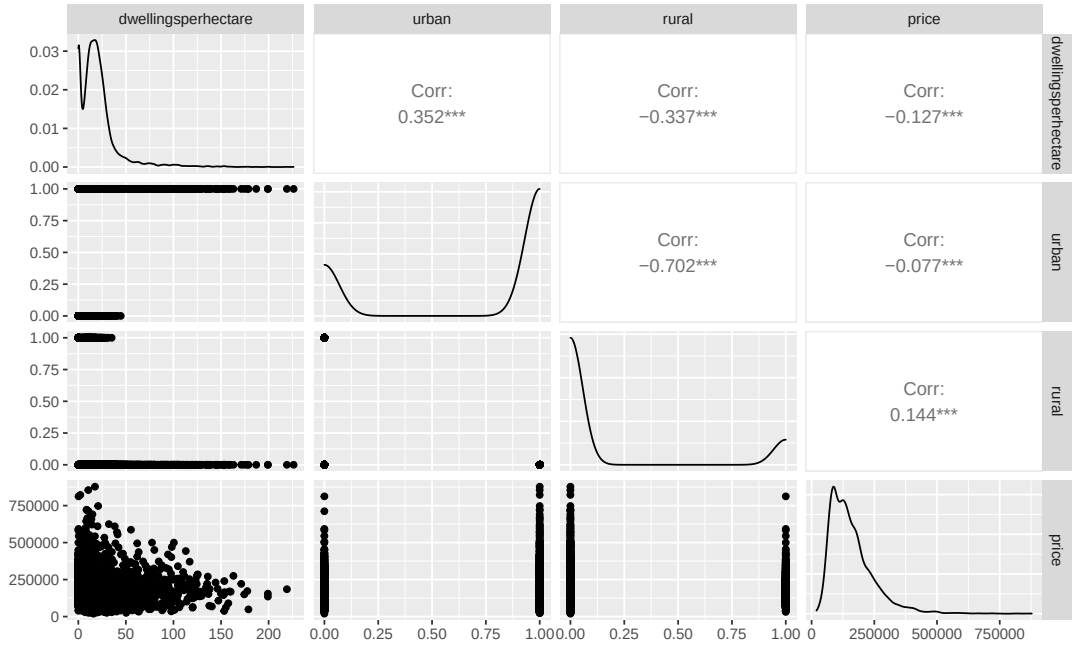

Figure 2.8: Correlation plot showing the relationship between the variables *price*, *dwellings per hectare*, *percentage urban* and *percentage rural* using scatterplots, density plots and correlation coefficients.

A number of variables are available about the geography of the Data Zone, including dwellings per hectare, percentage of urban properties and percentage of rural properties. Dwellings per hectare measures the number of residential properties per hectare of land hence the smaller the number the lower density the properties. This means that there is more room for personal space such as gardens, driveways and garages. In Figure 2.8, although there is no clear relationship between property price and dwellings per hectare, it can be speculated from the scatterplot that as the property price seems to increase, the number of dwellings per hectare decreases. This makes sense as bigger homes tend to take up more land so dwellings per hectare would be smaller. However, we see that there

are some areas with properties of around £150,000 with the same number of dwellings per hectare as areas with properties of around £400,000. This could be because of the local area where it is located and the desirability of it. Also, areas with lots of flats will have more dwellings per hectare and this could contribute to why there is right skewness as in some areas with flats the prices are a lot higher.

Furthermore, other important variables are whether a Data Zone is classified as urban or rural. The Data Zones are classified as either urban or rural through an 8 fold urban-rural classification measuring urbanicity developed by the Scottish Government (Scottish Government, 2020a). This 8 fold classification was simplified into 3 levels , urban, semi-urban small towns, and rural, and the urban and rural categories are retained here for analysis. There is no clear relationship between urbanicity and property prices across Scotland. For example, there are some very expensive properties in urban areas while there are also some very low priced properties in urban areas. The same can be said about rural areas. However, it is important to note that when examining the scatterplots, there is a larger volume of properties of high prices of around £750,000 in urban areas than in rural areas. These expensive properties are more likely to be city townhouses and flats rather than the large mansions of the same price in rural areas.

2.3.3 Characteristics of Data Zones

It was also important to look at the features of the area other than its property characteristics, so the Scottish Index of Multiple Deprivation was used to provide data on the socioeconomic features of each data zone. This looks at how deprived an area is across 7 different domains (Scottish Government, 2020a) - Employment, Income, Crime, Housing, Health, Education and Access. Each of these domains accounts for a different percentage of the final SIMD score, for example Employment accounts for 28% of the overall final score while Housing accounts for only 2%. Crime was removed from this study because its single indicator had 435 missing values, while in the other domains each indicator had at most 14 missing values. These small amounts of missing values were imputed using the k nearest neighbours algorithm with k=5 after scaling each indicator to have mean 0 and standard deviation 1. As the set of indicators within each domain are often highly correlated, principal component analysis was carried out to represent each domain with independent components. Note, this was only done for domains which had more than one indicator. In each case, enough principal components were kept so that the cumulative proportion of variation explained in each domain was over 80%. This resulted in each domain having the following number of indicators- education (2), access (3), employment

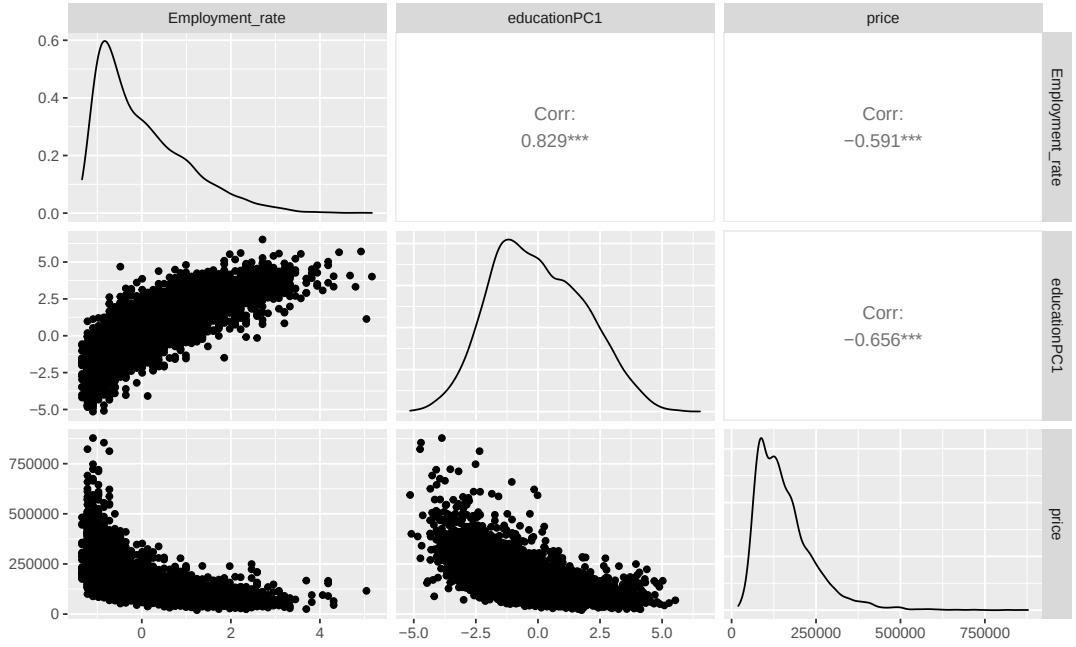

Figure 2.9: Correlation plot showing the relationship between the variables *price*, *employment rate* and *education* using scatterplots, density plots and correlation coefficients.

(1), income (1), health (3) and housing (2).

The final variable is the number of properties in each area that were allocated to a group from A to H corresponding to their council tax band. Band A is the cheapest properties while band H is the most expensive properties, although these bands vary by local authority. Then, a principle component analysis was applied to these 8 bands due to their high collinearity and the top 3 principle components were retained as features for analysis as they explained around 80% of the variation.

Displayed in Figure 2.9 are the correlation plots of two area based measures from the SIMD (Scottish Index of Multiple Deprivation), employment rate and education, with the price variable (Scottish Government, 2020b). At first glance, both scatterplots seem to follow a similar pattern suggesting that employment rate and education are indeed related to one another. This can be seen from the strong positive correlation coefficient of 0.829. Employment rate is a combination of 3 indicators of employment- those receiving Jobseeker's Allowance, those receiving Universal Credit who are unemployed and those who are claiming Incapacity Benefits, Employment Support Allowance or Severe Disablement Allowance (Scottish Government, 2020b). The higher the value of the employment rate value, the more deprived the Data Zone is. Essentially, the employment rate is a measure of unemployment. The scatterplot highlights that the more unemployment in an

area, the lower the property prices, as expected. We see that the majority of lower priced properties have a high unemployment. However, there are a good number who have low property prices yet low unemployment rates. These could perhaps be more affluent Data Zones which have affordable housing. Therefore, it is likely that the levels of deprivation in an area impacts property prices, for example the exact same house could be built in a more deprived area and an affluent area however due to the affluent area having a higher SIMD score thus being more desirable to live in, the price will be much higher here. As mentioned previously, the education covariate is just one of two principal components of education. In the SIMD, Education is measured using 5 indicators- School Pupil Attendance, Attainment of School Leavers, Working age people with no qualifications, 17-21 year olds enrolling in higher education and those age 16-19 not participating in higher education, employment or training (Scottish Government, 2020b). The scatterplot of education against property price shows that the vast majority of less educated people live in properties that are of lower prices. This makes sense as those who are well educated tend to be in better jobs which in turn would provide them with a better income and as a result will allow them to be able to purchase properties of higher prices.

One of the other domains of SIMD measures the geographical access to a variety of services, as measured by the travel time to get there. The better the amenities in an area could also impact the house prices. For example living near a more upmarket supermarket such as Waitrose or M&S can increase property prices (Shaw, 2017) . A study was conducted by Lloyds Bank where the average property prices in postal districts which have chain supermarkets were compared with average property prices in wider towns. The result of this study was that properties situated closer to a supermarket had a higher selling price than those that didn't, of average around £20,000 more than those further out. In fact, this is where the so called "Waitrose Effect" comes into play, as it was found that living nearby a Waitrose can increase the price of your property by around £40,000, compared to Aldi which can reduce the prices by between £3,000-£4,000 (Redhead, 2015). However, due to the cost of living increasing and more affluent areas having Lidl and Aldi stores being built, there is an increase in property price from budget supermarkets. This is very much a causality issue - did the supermarket being built cause the house prices to be higher or were they already high due to other factors? The West End of Glasgow is a good example of this. Furthermore, the number of fast food restaurants or outlets in an area may also have an effect on the property prices. Public Health England recent research has shown that the poorest areas in England have the highest numbers of fast food places and have at least 5 times more than the more affluent areas (Public Health England, 2018). This is because these businesses often

bring with them local litter problems with packaging being dropped and bins overflowing from not being emptied. This creates an eyesore hence making the area less desirable to live. Furthermore, drive thru's, especially with the sky rocketing of demand for fast food during covid times, can lead to traffic problems due to excessive queues during peak times and noise and air pollution from vehicles. Despite this research being carried out in England, it can be expected that Scotland will follow the same trend.

2.3.4 Data Splitting

In order to assess how successful a model is in prediction, we must take the data set and split it into various subgroups for training and then testing the model. This must be done before constructing any spatial or machine learning models so that the same training and test sets are used to compare all models. This is because using the same data on which a model is fit to assess its predictive performance will give biased results and tend to overfit (Robertson and Gray, 2021).

The main goal of this thesis is to create an algorithm which predicts outcomes of property price where there is no data most accurately using the set of covariates. This is otherwise known as generalizability. In total we have 6,881 Data Zones, 617 of which have missing values. Therefore, these missing values are removed from the total because these are the quantities that we want to predict and we cannot validate our models on them as there are no true values, which leaves us with 6,264 observed property price values. These can then be split into training and test sets. The training set will be used to fit the model while the test set, which must not be used before having chosen the final model, will be used as new data to evaluate the success of the subsequent models performance (Robertson and Gray, 2021). If on the other hand, the test set was incorporated into the making of the model, this would mean that the model would have been made in order to suit the test set when it should be independent of it. In this study, an 80-20 randomized split will be used with 80% of the Data Zones forming the training set and the remaining 20% forming the test set. If too much is spent in training, i.e. over 80%, then the model will be unsuitable as there is not enough test data to properly evaluate the performance. Whereas on the other hand, if too much is spent in testing, i.e. over 40%, then it will be unlikely to accurately assess how good the model parameters are as the model will not be fit realistically because there is not enough training data (Boehmke and Greenwell, 2019). Therefore, it is important to have a good split which is why the 80-20 split was selected. Once split the test set consists of 1,253 Data Zones and the training set has 5,011 Data Zones.

Following this, the training set will be split randomly into 10 subgroups of near enough equal size – 9 groups of 501, and 1 group of 502 (due to there being a remainder) - with each group in turn being used as a validation set within a 10 fold cross validation procedure. In this 10-fold cross validation, 9 of the 10 groups are considered training sets and 1 group is considered the validation set, and this procedure is repeated 10 times with a different validation set each time. The number 10 was chosen because this is typically what the value of k is in the majority of k-fold cross validation models. The greater the value of k, the smaller the difference between the estimated and the true performance will be on the test set (Boehmke and Greenwell, 2019), but the longer the cross validation process will take to run.

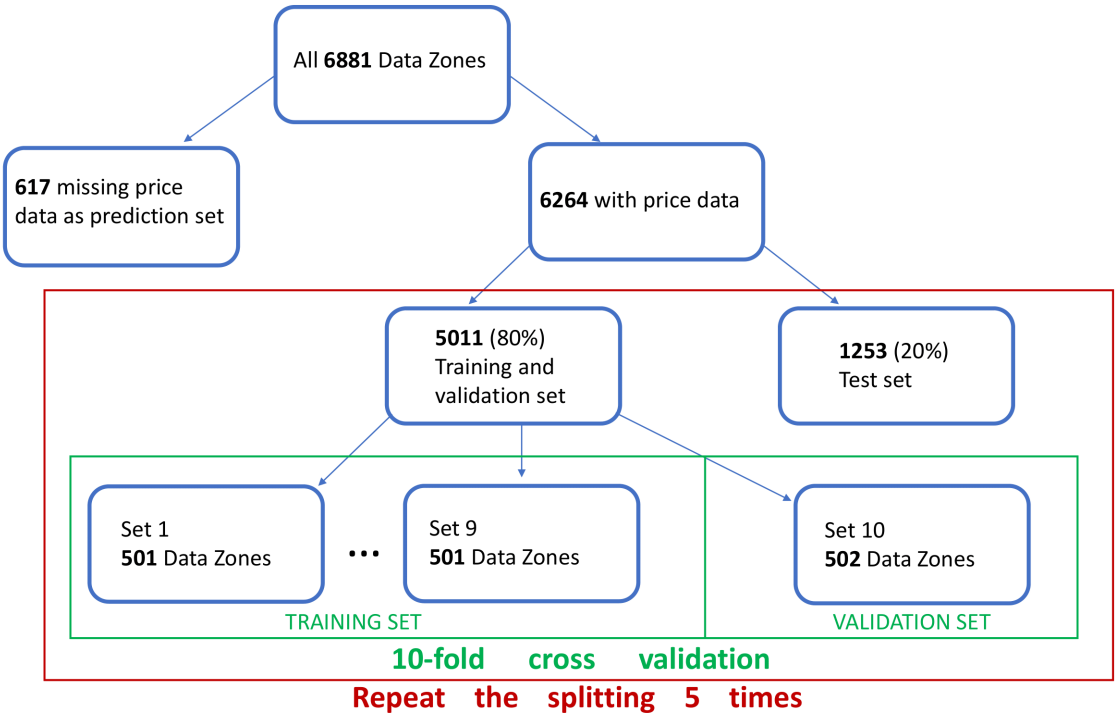

Figure 2.10: This diagram is a visualisation of the data splitting procedure.

This procedure is repeated a further 5 times in order to take into account any biases which may occur in the random data splitting (Boehmke and Greenwell, 2019). Figure 2.10 is a diagram showing how the data are split. Furthermore, the performance of the models are then evaluated with the test set using various different techniques and statistical analysis to come to a conclusion of which model is the most accurate and best for prediction of property prices in Scotland. Some examples of the statistical metrics evaluated are Root Mean Squared Error (RMSE), Median Absolute Error (MAE), Bias and Prediction Interval Coverage.

The RMSE measures the square error in the same units as the price variable so it is easily computable - essentially it is the square root of the Mean Squared Error. It is given by the following equation:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{2.1}$$

where n represents the number of observations in the test set and $(y_i - \hat{y}_i)$ is the difference between the actual value of property price on y_i and the predicted value of property price on $\hat{y}_i$ based on the model developed on the training set. The objective is to minimize the RMSE value as much as possible.

Next, the MAE is a measure of the median absolute error between the actual and predicted values. It is defined as,

$$MAE = \underset{i=1,\ldots,n}{\text{median}}\{|y_i - \hat{y}_i|\}. \tag{2.2}$$

Again, the smaller the value of MAE is the better. The MAE is less affected by outliers than the RMSE, because it does not have a squared term and uses the median rather than the mean.

Bias is a value computed by calculating the difference between the actual and the predicted value of property price. It follows the formula:

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i), \tag{2.3}$$

The aim is to have the bias value as close to 0 as possible as this would indicate the model is successful in prediction on average.

The Prediction Interval Coverage measures the proportion of the 95% prediction intervals that contain the true value, which should be close to 0.95. Thus it measures the accuracy of predictive uncertainty and not point prediction.

2.4 Normal Linear Model

I begin by assessing the predictive ability of a normal linear model as a simple base case. Ten models are compared, two for each of the 5 training/test data splits. The first uses all the covariates, while the second applies a backwards elimination procedure to select important covariates.

Backwards Elimination is when the original full regression model with all the possible covariates is fitted, then the covariate with highest p-value is removed. This process is repeated until all the covariates have a p-value which is significant, so in this case have a value of less than 0.05. By carrying out this process, we end up with 10 models in total - 5 'original' models with all the covariates of the 5 data splits and 5 respective 'backwards' models of the 5 data splits which contain only covariates chosen via backwards elimination. When the final backwards selection models are achieved they are then compared to the original full regression models with all the covariates using the prediction metrics described above and listed in Table 2.2.

However, before making predictions from this model, it is crucial to determine whether the model is valid and satisfies the 4 assumptions. The model assumptions are as follows - data must be normally distributed, residuals must have a mean value of 0, there must be constant variance and residuals must be independent. If these assumptions are not satisfied then transformations could be made.

First of all, the data splits are visualised and examined using plots with all of them producing very similar results. Figure 2.11 examines the normality of the residuals for one of the data splits. By looking at the normal q-q plot we see that the data points do not form a straight line and instead form a line which curves towards the right hand side with lots of outliers. This would suggest that the normality assumption is not validated and the residuals are not normally distributed and are very skewed to the right hand side. This indicates that normality is not present and hence a transformation must be used. In order to achieve normality a log transformation is carried out.

After transforming the average property price variable for all of the 5 data splits using a log transformation, the following plots were achieved of the residuals. Despite all the data splits again showing very similar results, the same data split used previously is used again as an example in Figure 2.12 . This new q-q plot in Figure 2.12, which assesses normality through the distribution of residuals, is an improvement on the q-q

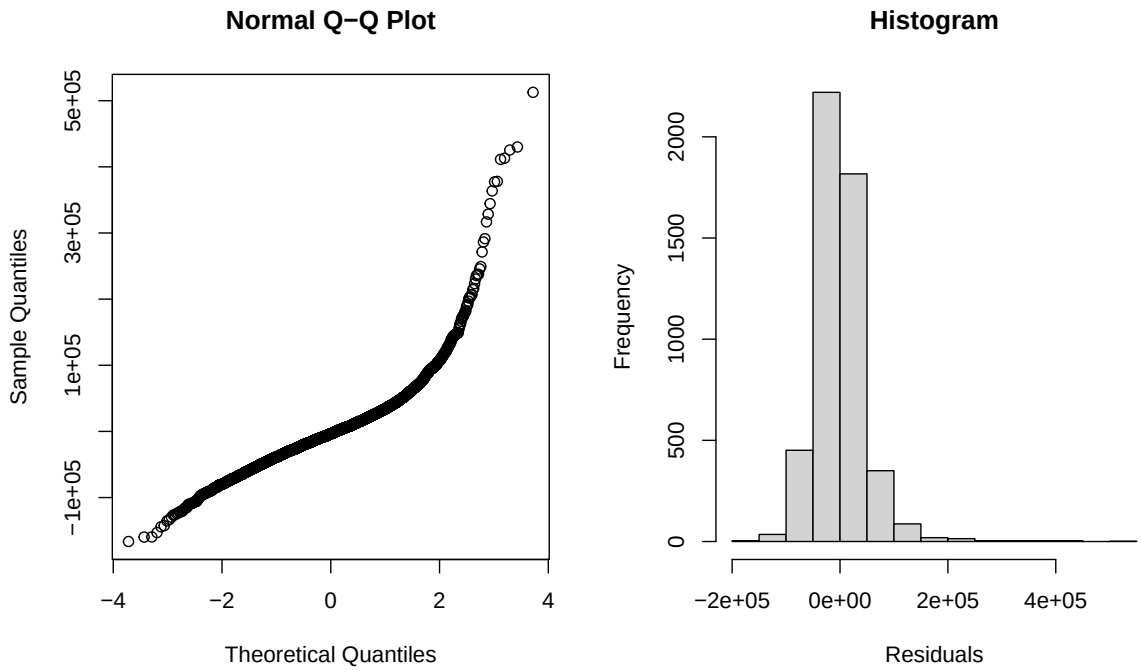

Figure 2.11: Normal Q-Q Plot and Histogram of residuals from the full covariate model before transformations.

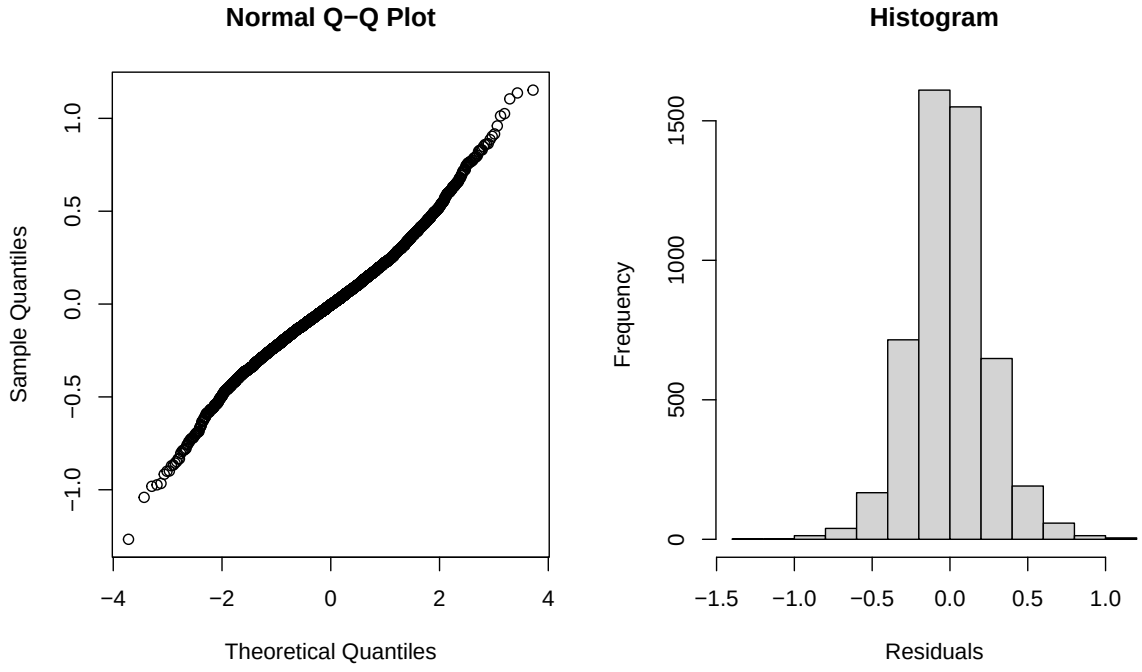

Figure 2.12: Normal Q-Q Plot and the Histogram of residuals from the full covariate model once transformed using a log transformation.

Table 2.2: Table showing different Statistical Model Evaluation Analysis of each of the 5 Original Full Regression Models and their respective Backwards Model.

MODEL	Parameters	RMSE	MAE	Coverage Probability
1 original	23	£47,995	£20,038	0.947
1 backwards	12	£47,933	£20,231	0.945
2 original	23	£48,973	£19,902	0.936
2 backwards	14	£48,926	£20,003	0.935
3 original	23	£45,631	£20,910	0.943
3 backwards	12	£45,754	£20,926	0.941
4 original	23	£43,524	£18,704	0.958
4 backwards	14	£43,446	£19,098	0.956
5 original	23	£45,875	£18,984	0.951
5 backwards	13	£45,802	£19,181	0.951
Average Original	23	£46,400	£19,708	0.947
Average Backwards	13	£46,372	£19,888	0.946

plot in Figure 2.11 and now resembles more of a straight line with no obvious curve at the edges anymore. Furthermore, the histogram is now symmetric and has a peak around 0. This highlights that now the conditions for normality appear reasonable, so all further modelling will be done on the log scale and predictions then back transformed to the original scale.

2.4.1 Variable Selection

Firstly, in Table 2.2, when comparing each of the original models to their respective backwards model, the backwards models all have significantly less parameters - at least 9 but sometimes 11 less parameters. It is clear that these parameters are covariates that are insignificant to the model by use of backwards elimination based on p-values set at 0.05. This means that they have coefficients of around 0 hence they make no real impact whether they are part of the model or not. This can be proved by the very small difference in Mean Error, Median Absolute Error and Bias between the original and the backwards models. Therefore, it is best to select the original models for the spatial modelling and machine learning models. This is because here and in spatial models these models are linear in the covariates. However, machine learning models can be non-linear, so there may be some effects of the covariates that are non-linear and not picked up here hence in the future I keep all covariates in the models.

Despite there being such large values of all the mean errors and median absolute errors, this is not a cause for concern. They are both large because of the large range in property prices. As found earlier in Section 2.2, the average property price is around

£150,000 but there are some clear outliers which have prices of around £800,000. These outliers cause the data to be skewed and hence create large errors in prediction. The median absolute error is lower than the mean error because the median is affected much less by the skewing and the outliers than the mean is. When looking at all the models as a whole the RMSE is around £46,000 with all models within £5,000 of each other while the MAE is around £20,000 with the range being around £2,000. These values being so close to one another indicate similarities between each of the data splits.

Finally, the final column in Table 2.2 is the coverage of the 95% prediction intervals, i.e. the proportion of times they contain the true value. All 10 of the models have a proportion of between 0.93 and 0.96 of the test set which lies within the prediction interval. As these are all close to 0.95, this shows that all models are fairly successful in quantifying predictive uncertainty, with Model 4, both original and backwards models, having the highest proportion lying within the prediction interval - 0.958 and 0.956 respectively.

2.5 Discussion

A general overview of the data has been provided to give an insight into some properties of the data set. Then, as the aim is to create an algorithm which best predicts property price using the covariates, by splitting the data randomly into training-test data splits and transforming to reduce uncertainity has aided to create and assess a best fit model of the data. Further analysis of the models and their predictive performance has been calculated using various prediction metrics.

Overall, despite noticing some key patterns and relationships between the covariates and the price variable, it is clear that there is not one single factor that has a strong relationship with property price as there is a lot of noise in each of the scatterplots. Therefore, this leads to the conclusion that there are multiple factors that contribute to the price of a property, which motivates the use of regression models and machine learning techniques in the analysis.

The use of backwards elimination to test the impact of covariate selection has shown that the original and backwards models are very statistically similar to each other. This occurring in not only 1 or 2 but all 5 of the data splits highlights that the covariates which have been removed have very little to no impact on the normal linear regression models. Through various prediction metrics, it has been shown that there is very little difference between each of the data splits emphasising the consistency between all 5.

Hence, we proceed into further analysis using the full set of covariates on the chance that there are non-linear effects produced by the covariates. This will be conducted by fitting a spatial model to the data splits to evaluate prediction abilities and find a best fit model.

Chapter 3

Property price predictions using spatial conditional autoregressive models

3.1 Introduction

In this chapter, the aim is to utilize the spatial structure in the data to improve the accuracy of the spatial prediction of property prices in Scotland. Firstly, a condition surrounding spatial autocorrelation must be satisfied before developing this spatial model. Spatial autocorrelation must be shown to exist in the residuals from the non-spatial model, otherwise the spatial structure in the data is unlikely to aid the prediction. This is done by evaluating the residuals from the simple linear regression model in the previous chapter using a Moran's permutation test (Moran, 1950). Once it has been determined that spatial autocorrelation does exist, then I can continue to fit a spatial model and use it for predictions.

3.2 Exploratory Analysis

In order to correctly model spatial dependence and show how spatially close Data Zones are to one another, a neighbourhood matrix $\mathbf{W}$ must be created. The neighbourhood matrix $\mathbf{W}$ is the vehicle by which the spatial closeness (proximity) between each pair of areal units is determined. Typically a binary specification is used, where two areas are defined to either be neighbours or not neighbours of each other. Note this is different from point level data, where exact distances between points are used to measure closeness. This is not used here because the distance between two areas is not unique as they are

areas and not single points. This neighbourhood matrix is the basis for defining the spatial autocorrelation structures implied by the models used in this section. There are two commonly used methods that can be used to create the neighbourhood matrix **W**, the K-Nearest Neighbours (KNN) method and the border sharing method (Bivand et al., 2008).

3.2.1 KNN Method

In the KNN method (Bivand et al., 2008), the central points of all Data Zones are identified, and a value of k is selected which represents the number of nearest neighbouring Data Zones used to create **W**. Then if Data Zone B_j is one of k nearest neighbours to Data Zone B_i, this can be represented in the **W** matrix as w_{ij}=1, indicating that these areas are spatially close to each other, with w_{ij}=0 otherwise if this is not the case. This method is applied to all Data Zones in turn to construct **W**. The resulting matrix will be asymmetric because B_i may be one of the k nearest neighbours to B_j, but B_j may not be one of the k nearest neighbours to B_i. To solve this, if $w_{ij} = 0$ and w_{ji}=1 then both are set to be equal to 1 to make sure that **W** is symmetric. Moreover, w_{ii} will always equal 0 because the Data Zones cannot be neighbours of themselves.

3.2.2 Border Sharing Method

The border sharing method is more straightforward than the KNN method and produces a symmetric matrix without needing a correction step. Essentially, if B_i and B_j share a border, an edge where they both meet, then w_{ij}=w_{ji}=1 in the **W** matrix and if not they will equal 0. Again, w_{ii} will always equal zero since a Data Zone cannot be a neighbour of itself.

As a general rule, border sharing is the preferred method for this type of spatial areal unit modelling, due to its simplicity. However, because our training and evaluation processes will split the data into training and test data splits, there will be missing areal units in the study region when fitting the model so that some Data Zones will have no neighbours under the border sharing rule. Therefore in this particular case, it is more beneficial to continue with the KNN method to avoid having isolated Data Zones with no neighbours.

3.2.3 Assessing the presence of spatial autocorrelation

Once the **W** matrix has been produced it can be further used to assess the presence of spatial autocorrelation. Spatial autocorrelation gives a measure of the relationship

between neighbouring observational units based on the similarity of their data values (Shekhar and Xiong, 2007). One method for investigating if spatial autocorrelation is present in the residuals of the simple non-spatial models is a Moran's I test. Developed by statistician Patrick Alfred Pierce Moran, it is an extension of Pearson's correlation coefficient to measure spatial autocorrelation between areal units (Moran, 1950). This test computes a Moran's I statistic, which is a single value that measures the strength of the linear association in areal data with respect to their spatial locations. Supposing we have data $\mathbf{Z}$=(Z_1, ..., Z_n) and neighbourhood matrix $\mathbf{W}$, with $\bar{Z}$ representing the sample mean of $(Z_1, \ldots, Z_n)$. Then Moran's I statistic can be defined as:

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}(Z_i - \bar{Z})(Z_j - \bar{Z})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}(Z_i - \bar{Z})^2}, \tag{3.1}$$

where the value of I will lie between -1 and 1, and gives the spatial autocorrelation in the data. The value of I can be interpreted with regards to spatial autocorrelation as follows:

$$I = \begin{cases} 1, & \text{there is perfect positive autocorrelation} \\ 0, & \text{there is no spatial autocorrelation} \\ -1, & \text{there is negative spatial autocorrelation.} \end{cases} \tag{3.2}$$

When carrying out a Moran's I test, a hypothesis test is carried out to assess the significance of the spatial autocorrelation, where the hypotheses are set as follows:

- H_0= No spatial autocorrelation
- H_1= Positive spatial autocorrelation.

At a 5% significance level the p-value is calculated. It is computed by Monte Carlo permutation, a randomised process where independence is satisfied. The process consists of taking the data points then randomly reallocating them to Data Zones. The value of the I statistic is then obtained, and the process is repeated a large number of times, say 10,000. These 10,000 I statistics have been generated under independence as they are based on a random permutation of the data, and they can then be compared to the Moran's I value from the data. The subsequent p-value from the hypothesis test is the proportion of these randomised values that are bigger than the observed value of Moran's I. If the p-value is less than 0.05 then H_0 is rejected and we conclude that spatial autocorrelation is present. Whereas, if the p-value is greater than 0.05 then H_0 is not rejected and it is concluded that there is not evidence to suggest that there is positive spatial autocorrelation.

When spatial autocorrelation is present in the residuals from a non spatial model, this would indicate that a spatial model may be more appropriate than the linear model (Lee, 2013). One would hope that prediction could improve by taking into account the spatial structure in this situation in the data. The linear model created in the last chapter was applied to data from all Data Zones with non-missing price values, and the residuals were created. Table 3.1 displays the results of Moran's I tests for these residuals, which were conducted with $k = 1, \ldots, 10$. If this test suggests that there is no spatial structure left in the residuals then there is no point in fitting a spatial model as the covariates account for the entire spatial structure in the data.

Table 3.1: Table of Moran's I applied to residuals from a simple linear model for various k.

k	I statistic	P-value
1	0.19901	<0.001
2	0.20409	<0.001
3	0.18683	<0.001
4	0.17715	<0.001
5	0.16671	<0.001
6	0.15947	<0.001
7	0.15610	<0.001
8	0.15053	<0.001
9	0.14462	<0.001
10	0.14225	<0.001

The results from Table 3.1 show that the I statistics from all 10 values of k are positive and all very similar to each other. A pattern present from these results is that as the value of k increases, the I statistic decreases hence there is less and less spatial association. This would make sense as the higher the number of nearest neighbours a Data Zone has, the lower the chance of all of these Data Zones being similar to each other as some are situated further away from the Data Zone in question than others. With the I statistics all being between 0.14 and 0.21, it indicates that neighbouring Data Zones seem to have some association with each other and share some similarities in their linear model residuals. This illustrates that it is not just a particular specific value of k but that there is spatial autocorrelation in the residuals for all values of k considered. The p-values for every value of k are less than 0.001 so since this is less than 0.05, we can reject the null hypothesis which states there is no spatial autocorrelation and we can conclude that there is significant evidence to suggest that there is spatial autocorrelation present between neighbouring Data Zones. This is true for all values of k suggesting this result is robust to the choice of k.

3.3 Spatial modelling of areal unit data

Conditional Autoregressive (CAR) models are the most common way of modelling spatial autocorrelation in areal unit data. The correlation is modelled by a set of random effects, which are forced to be spatially correlated by making their joint prior distribution a CAR type model (Lee and Mitchell, 2013). In what follows the model is fitted in a Bayesian paradigm. The general model begins with a linear model of the form,

$$Y_k \sim \mathrm{N}(\mathbf{x}_k^T \boldsymbol{\beta} + \phi_k, \sigma^2) \quad \text{for } k = 1, \ldots, n. \tag{3.3}$$

This is a similar structure to the linear model described in the previous chapter however with an extra variable, ϕ_k, which is a random effect for the k^{th} Data Zone. In this spatial linear model the vector of random effects $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$, are defined to be spatially correlated and hence the residuals from this model should now be independent. These random effects are modelled with a CAR Model based on the neighbourhood matrix $\mathbf{W}$.

CAR models are a type of Gaussian Markov Random Fields (GMRF), which are a general class of models used to construct dependence amongst random variables (Rue and Held, 2005). Another example of a GMRF are Autoregressive (AR) models which are used in time series. The simplest CAR model is the Intrinsic CAR model (Besag et al., 1991), which states that the random effect ϕ_k depends only on the random effects ϕ_j in a small number of neighbouring areas, as defined by $\mathbf{W}$. It can be written as a set of n univariate full conditional distributions, $f(\phi_k|\phi_1, \ldots, \phi_{k-1}, \phi_{k+1}, \ldots, \phi_n)$ for all k, where n is the total number of Data Zones. These conditional distributions are given by

$$\phi_k|\boldsymbol{\phi}_{-k} \sim \mathrm{N}\left(\frac{\sum_{j=1}^n w_{kj}\phi_j}{\sum_{j=1}^n w_{kj}}, \frac{\tau^2}{\sum_{j=1}^n w_{kj}}\right) \quad \text{for all } k = 1, \ldots, n, \tag{3.4}$$

where $\boldsymbol{\phi}_{-k}= (\phi_1, \ldots, \phi_{k-1}, \phi_{k+1}, \ldots, \phi_n)$. This model has a normal distribution with the mean being the sample mean of the random effects in neighbouring areas and the variance being a single parameter τ^2 divided by the number of neighbours the Data Zone k has. The definition of the variance makes logical sense here as it is inversely proportional to the number of neighbours and therefore the more neighbours a Data Zone has, the less uncertainty the model implies about its random effect. Although this model is intuitively simple, there are some drawbacks. Firstly, the joint distribution $f(\phi_1, \ldots, \phi_n)$ corresponding to the set of $f(\phi_k|\phi_{-k})$ for all k is improper. The precision matrix is singular as its determinant is zero, hence the covariance matrix does not exist. Also, there is no parameter to control the strength of the spatial correlation (Lee, 2023).

The Leroux CAR model (Leroux et al., 1999) is an adaptation of the Intrinsic CAR model with a parameter for spatial dependence, ρ, which addresses the above problems. Its full conditional distribution is given by

$$\phi_k|\boldsymbol{\phi_{-k}} \sim \mathrm{N}\left(\frac{\rho\sum_{j=1}^{n} w_{kj}\phi_j}{\rho\sum_{j=1}^{n} w_{kj}+1-\rho}, \frac{\tau^2}{\rho\sum_{j=1}^{n} w_{kj}+1-\rho}\right). \tag{3.5}$$

It is important to note that ρ only has values within the range 0 to 1 with values close to 0 representing weak correlation, 0.5 representing moderate correlation and close to 1 representing strong correlation. If $\rho = 0$ then the distribution is $\phi_k|\boldsymbol{\phi}_{-k} \sim N(0, \tau^2)$, indicating independence is present, and this means that the neighbouring random effects have no impact on the distribution of ϕ_k. On the contrary if $\rho = 1$ the random effect will follow the Intrinsic CAR model, and so because there is strong correlation the random effect ϕ_k will be explained by the random effects in the neighbouring Data Zones. The joint distribution for the Leroux CAR model corresponding to the above full conditionals can be written as

$$\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n) \sim \mathrm{N}(\mathbf{0}, \tau^2\boldsymbol{Q}(\boldsymbol{W}, \rho)^{-1}), \tag{3.6}$$

where $\boldsymbol{Q}(\boldsymbol{W}, \rho)$ is the precision matrix which is given by,

$$\boldsymbol{Q}(\boldsymbol{W}, \rho) = \rho[\mathrm{diag}(\mathbf{W1}) - \mathbf{W}] + (1-\rho)\mathbf{I}. \tag{3.7}$$

Here, $\mathbf{W}$ is the neighbourhood matrix while $\mathbf{1}$ is an n $\times$ 1 vector of ones, and $\mathbf{I}$ is the n $\times$ n identity matrix. When $i \neq j$, then Q_{ij} = -ρw_{ij} but when $i = j$ this means Q_{ii}=$\rho\sum_{j=1}^{n} w_{ij}$+(1-$\rho$). Finally, if $\rho \in [0, 1)$ then $\boldsymbol{Q}(\boldsymbol{W}, \rho)$ is invertible, however if $\rho = 1$ it will be singular.

3.3.1 Prior distributions

For each of the parameters, τ^2, ρ and $\boldsymbol{\beta}$= $(\beta_1, \ldots, \beta_p)$, prior distributions need to be specified. These are prior beliefs about each of the parameters separately. In this thesis, the prior distribution for each β_j is specified as

$$\beta_j \sim \mathrm{N}(0, \sigma_\beta^2) \quad \text{for } j = 1, \ldots, p, \tag{3.8}$$

where σ_β^2 is the variance which is chosen here to be 100,000 to make the prior weakly informative and give almost no prior information about the values of $\boldsymbol{\beta}$. The variance

parameter τ^2 is assigned a log-gamma distribution on the log precision scale, that is,

$$\ln\left(\frac{1}{\tau^2}\right) \sim Log - Gamma(1, 0.01). \tag{3.9}$$

This distribution is the default family of priors which have been suggested by the INLA programme that is used for inference (Rue et al., 2009).

Finally, as ρ must be in the interval [0,1], its logit is modelled as a normal distribution as follows, which is again an INLA default.

$$\ln\left(\frac{\rho}{1-\rho}\right) \sim \mathrm{N}(0, 10). \tag{3.10}$$

After these prior distributions for the parameters are specified, they are then combined with the likelihood function for the observed data shown in (3.3) and (3.6) to construct the posterior distribution for the model (Gelman et al., 1995).

3.3.2 Spatial prediction

Since the property price data are modelled on the log scale, any predictions for the test set observations $\hat{Y}_i^{(test)}$ must be made on the original scale. Hence to achieve appropriate property price predictions, the final predictions have the form,

$$\hat{Y}_i^{(test)} = \exp\{\mathbf{x}_i^{(test)^T}\hat{\boldsymbol{\beta}} + \hat{\phi}_i^{(test)}\}. \tag{3.11}$$

Here, $\hat{Y}_i^{(test)}$ is the test set observation to be predicted, $\mathbf{x}_i^{(test)^T}$ is the set of covariates corresponding to this observation and $\hat{\boldsymbol{\beta}}$ are regression parameters estimated as the posterior mean values based on the model applied to the training set. Since this is a spatial model, it is also necessary that the random effect, $\hat{\phi}_i^{(test)}$, is estimated.

This is done using the same Leroux CAR model applied to training set. Thus

$$\hat{\phi}_i^{(test)} = \mathbb{E}(\phi_i^{(test)}|\boldsymbol{\phi}) = \frac{\hat{\rho}\sum_{r=1}^{K} w_{ir}\hat{\phi}_r}{\hat{\rho}\sum_{r=1}^{K} w_{ir} + 1 - \hat{\rho}}, \tag{3.12}$$

where r is an observation from the training set and i from the test set. Here, $\hat{\rho}$ and $\hat{\phi}_r$ are the posterior means from the training set and $w_{ir} = 1$ if the r^{th} Data Zone in the training set is one of the k^{th} closest to the i^{th} Data Zone in the test set.

This model is fit to 9 out of the 10 folds of the training set separately for each value of k when constructing $\mathbf{W}$, and is used to predict the 10th (validation set). This process is repeated 10 times each time leaving out a different fold. The RMSE and MAE are then computed for each value of k from 1 through 10 and the 'best' value of k is chosen. Once the best value has been chosen, the model is then refit to the entire training set using this 'best' value of k in order to make predictions on the test set.

3.3.3 Parameter estimation

Two common ways of fitting CAR Models within a Bayesian setting are Integrated Nested Laplace Approximation (INLA)(Rue et al., 2009) and Markov Chain Monte-Carlo (MCMC) simulation (Gómez-Rubio, 2021a). INLA can be implemented via the INLA package in R which is used for approximating Bayesian inference of Latent Gaussian models (R INLA Project, 2020). In recent years, it has become a popular alternative to MCMC simulation because MCMC can often be computationally expensive when dealing with large volumes of data like in this study (Gómez-Rubio, 2021b).

3.4 Choosing the number of neighbours k to construct W

Before the spatial model is fit and predictions are made, it is necessary to determine whether there are covariates that are not needed now due to being represented by spatial terms. These covariates will be removed from the model. Therefore, two covariates – easting and northing – have been removed as linear terms in the model, because the effects of spatial location are instead modelled more flexibly by the random effects as summarised above. Although there are arguments for both keeping these covariates as linear terms or removing them from the prediction model, it has been decided to remove these covariates. This decision is justified in terms of the easting covariate because there could be a relationship that as you move further east across Scotland then property price increases, but this would imply that the whole of Glasgow has lower property prices than Edinburgh. This is untrue because although there may be some very expensive affluent areas in Edinburgh, there are also areas like this in Glasgow, hence some areas in Glasgow will indeed have higher property prices than some areas in Edinburgh. Moreover, the same theory can be applied to the northing covariate – the further north of Scotland you travel, the higher or lower the property prices. This would imply that two places at

similar latitudes have similar property prices, but Fort William and Aberdeen contradict this. Aberdeen is a big city in comparison to Fort William which would suggest that property prices would be higher here as there is better access to amenities. Thus linear relationships between property price and easting/northing seem inappropriate. However, all other covariates from the linear models in the previous chapter are retained.

3.4.1 Validation strategy

As shown previously in Section 2.3.4, 5 different data splits are considered through the process shown in Figure 2.10. These are training-test splits assigned randomly in order to assure that the splitting hasn't impacted the results (e.g. all of Glasgow has ended up in the same data split). In each split, the test set is removed and a model is created with the training set which in turn is separated into 10 more splits within itself. For the purposes of choosing the best k for constructing $\mathbf{W}$, the model is fit to 9/10 of the training set and is used to predict the 10th part of the training set otherwise known as the validation set. This process is repeated 10 times and the validation set is changed each time. The tuning parameter k is selected by applying the model with $\mathbf{W}$ constructed with each value of k from 1 to 10, and predictions made of the validation set in each case. Table 3.2 shows the computed values of the root mean square error (RMSE) for the validation sets on each of the 5 data splits for each of the 10 values of k, while Table 3.3 shows the median absolute error (MAE). Across the 5 splits and 10 values of k, the metrics are computed over all observations in the training set. The averages across the 5 data splits are used to obtain the best value or values of k for prediction, which is the one which produces the lowest RMSE and MAE. From Table 3.2, the best value of RMSE is £42,595 which occurs when $k = 7$, whereas Table 3.3 shows the best value of MAE is £18,007 when $k = 3$. Therefore, since the best values of k are not the same for the RMSE and MAE metrics we will continue test set prediction in the next section with both $k = 3$ and $k = 7$. In general, as the values of RMSE and MAE are very similar for all values of k, it indicates that its value should not have a big effect on the final results.

3.4.2 Test Set Predictions

The spatial model is refit to each entire training set with k equal to 3 and 7, and property prices are predicted for the test set using the fitted model on each of the 5 test splits. The resulting RMSEs and MAEs are shown in Table 3.4, which also includes the results from the linear model from the previous chapter, for comparison. Comparing the spatial models to the original linear model, we see that there is a general improvement as the RMSE values have decreased as have the MAE values for both values of k in each of the

Table 3.2: Table of root mean square error (RMSE) of property price for each value of k when constructing $\mathbf{W}$ applied to the 5 data splits.

k	Split 1	Split 2	Split 3	Split 4	Split 5	Average
1	£44,368	£44,429	£45,698	£45,858	£45,332	£45,137
2	£42,751	£42,786	£44,140	£44,105	£43,701	£43,497
3	£42,361	£42,299	£43,633	£43,539	£43,246	£43,016
4	£42,230	£42,054	£43,394	£43,462	£42,895	£42,807
5	£42,091	£41,910	£43,302	£43,295	£42,943	£42,708
6	£42,007	£41,879	£43,132	£43,142	£42,869	£42,606
7	£41,982	£41,933	£43,108	£43,148	£42,803	**£42,595**
8	£42,017	£41,998	£43,089	£43,124	£42,814	£42,608
9	£42,148	£41,937	£43,131	£43,159	£42,864	£42,648
10	£42,150	£42,015	£43,187	£43,189	£42,904	£42,689

Table 3.3: Table of median absolute error (MAE) of property price for each value of k when constructing $\mathbf{W}$ applied to the 5 data splits.

k	Split 1	Split 2	Split 3	Split 4	Split 5	Average
1	£18,258	£18,438	£18,376	£18,910	£18,806	£18,578
2	£17,945	£18,198	£18,330	£18,860	£18,340	£18,335
3	£17,762	£17,966	£17,968	£18,341	£17,998	**£18,007**
4	£17,968	£17,993	£18,168	£18,170	£18,195	£18,099
5	£17,859	£18,066	£18,113	£18,308	£18,112	£18,092
6	£17,878	£17,838	£18,163	£18,226	£18,207	£18,062
7	£17,985	£17,812	£18,234	£18,111	£18,031	£18,035
8	£17,858	£17,775	£18,139	£18,294	£18,158	£18,045
9	£17,919	£17,809	£18,208	£18,210	£18,000	£18,029
10	£17,840	£17,788	£18,233	£18,128	£18,157	£18,029

5 data splits. The RMSE and MAE differ from each other in both the linear and spatial models because the median is much less affected by skewness and outliers than the mean is. Comparing RMSE and MAE values of the linear model to each of the spatial models for each split, it can be seen that when $k = 3$, the RMSE and MAE decreases by around 9%, and likewise when $k = 7$, the decreases are around 10%. This gives an indication that the spatial models are more accurate in prediction than the original linear model as the lower the values of RMSE and MAE, the better the prediction. The bottom of Table 3.4 shows the average values of RMSE and MAE for the linear model and the spatial models for the two values of k over all 5 data splits. By calculating these averages, this makes it easier to compare the models. Overall, we see that there are very similar results when $k = 3$ and $k = 7$. By using spatial models instead of the linear model, the RMSE has improved by around £5,000 while the MAE has improved by nearly £2,000. When predicting on the test set, $k = 7$ has the lowest RMSE in every split and the lowest

Table 3.4: Table showing root mean square errors (RMSE) and median absolute errors (MAE) of each of the 5 data splits using a linear model, and two spatial models with different k values.

Split	Model	RMSE	MAE
1	Linear	£47,995	£20,038
	KNN($k = 3$)	£44,590	£17,915
	KNN($k = 7$)	£44,068	£18,386
2	Linear	£48,973	£19,902
	KNN($k = 3$)	£44,291	£17,972
	KNN($k = 7$)	£43,775	£18,111
3	Linear	£45,631	£20,910
	KNN($k = 3$)	£40,453	£17,880
	KNN($k = 7$)	£40,122	£17,431
4	Linear	£43,524	£18,704
	KNN($k = 3$)	£39,573	£18,036
	KNN($k = 7$)	£39,439	£17,624
5	Linear	£45,875	£18,984
	KNN($k = 3$)	£40,992	£17,723
	KNN($k = 7$)	£40,861	£17,244
Average	Linear	£46,400	£19,708
	KNN($k = 3$)	£41,980	£17,905
	KNN($k = 7$)	£41,653	£17,759

overall average. However, despite having the lowest average MAE, in 2 out of 5 of the data splits, $k = 7$ has a higher value of MAE than $k = 3$. This shows that both values of k have very similar outcomes as there are very small differences between their RMSE and MAE values. However, due to having lower averages, $k = 7$ is the better value for prediction. Finally, these results suggest that using spatial models has improved the accuracy compared to the simple linear model, because the random effects in the model have taken the remaining spatial structure in the data taken into account.

3.5 Discussion

In conclusion, by constructing a CAR model with spatially correlated random effects, we see that there is a clear improvement in predictive performance from the original linear models, thus suggesting that the spatial modelling has helped create a more accurate prediction model.

By evaluating the original linear models using a Moran's I test to obtain I statistics, we have shown that spatial autocorrelation is present in the residuals of a simple linear

model hence spatial modelling is appropriate for predicting property prices in Scotland. The best value of k when constructing $\mathbf{W}$ was firstly selected by comparing the RMSE and MAE for values $k = 1$ to 10 across 5 data splits, which resulted in k=3 and k=7 being chosen as they had the lowest MAE and RMSE values respectively. Furthermore, by applying these two values of k to predict the test set observations and comparing these spatial models to the original linear model, a significant improvement can be seen and it can be concluded that $k = 7$ is the best value for prediction. Overall, the average RMSE value has decreased by around £5,000 and the MAE by nearly £2,000, emphasising that the spatial model has higher predictive accuracy in predicting property prices across Scotland than the linear model.

However, despite this improvement, a problem with spatial models is that all of the covariate effects are assumed to be linear which could negatively impact the accuracy of the predictions. So, it may be beneficial to explore and consider non-linear methods of prediction such as Machine Learning (ML) methods. In the following chapter, I will investigate whether ML methods such as random forest and gradient boosting machines, which are known to be good at prediction, can outperform the spatial models above in the case of property prices.

Chapter 4

Property price predictions using classical machine learning methods

4.1 Introduction

Previously, the predictive ability of spatial CAR models were compared to linear models in the context of property price data and I learned through prediction metrics that the spatial CAR models produced more accurate predictions than linear models. Another commonly used family of methods which are good for prediction, as they consider both linear and non-linear covariate effects, are machine learning methods, which have been popularly used in various different settings including medical (Cruz and Wishart, 2006), geographical (Georganos et al., 2021) and financial (Bensic et al., 2005).

In this chapter, I will investigate specifically tree-based machine learning methods beginning with a single basic decision tree (Section 4.2) then following on with ensemble tree-based methods such as bagging (Section 4.3), random forest (Section 4.4) and gradient boosting (Section 4.5). Despite originating from the same family of methods, each of these tree-based methods have different strengths, which are outlined later in this chapter. Bagging has been used to help protect computers from threats through an intrusion detection system as it uses the bootstrapping technique to create an ensemble of decision trees then takes an average over this in order to reduce the rate of false-positive threats arising (Gaikwad and Thool, 2015). Random forests, which similarly build an ensemble of trees but encourage more randomness amongst the trees by only using a subset of covariates, have been used in the diagnosis of colon cancer by identifying the genes which have the highest correlation with the cancer (Su et al., 2022). Finally gradient boosting helped to prevent bank failure in the USA through identifying key factors of risk using an

ensemble of small trees where each tree builds upon its predecessor (Carmona et al., 2019).

By investigating these tree-based machine learning methods, this will allow a conclusion to be reached as to whether, in a spatial data context, classical a-spatial machine learning methods can predict better than standard spatial models. This is a very new topic of research which has been rarely approached, some examples being Tehrany et al. [2019] and Knoll et al. [2019] in the context of environmental spatial data, so the results I gather will provide an interesting insight into machine learning methods in a spatial data context. Furthermore, another aim is to see how different tree-based machine learning methods compare in predictive ability to one another, whether one specific method is significantly better at prediction or if all methods have very similar predictive performance.

Predictive performance of the tree-based methods will be measured using the same data splitting techniques from Section 2.3.4 and train-validation-test strategies as Sections 3.4.1 and 3.4.2 in order to maintain consistency throughout. This allows the tree-based methods to not only be easily comparable to one another but also to the spatial CAR model as seen in Section 3.2.1. Like previously, by calculating the root mean square error (RMSE) and median absolute error (MAE) and using these prediction metrics to measure the performance of each of the methods, the overall best method of prediction can be determined. The remainder of this chapter is organised as follows: Section 4.2 discusses basic decision trees whilst Sections 4.3, 4.4 and 4.5 analyse the bagging, random forest, and gradient boosting methods respectively and finally Section 4.6 gives a comparison of each of the machine learning methods to each other whilst also analyses them against previous methods studied such as the linear model and spatial model.

4.2 Decision Trees

Decision trees are a class of non-parametric algorithms that split the data in the training set into numerous subgroups that internally contain similar response values through a set of splitting rules in order to make predictions on the test set. They can be very beneficial in the sense that they are easy to construct and interpret, however alone they typically lack in predictive performance by being biased or imprecise (Boehmke and Greenwell, 2019). In order to fix this problem and achieve a model with the best predictive performance, random forest and gradient boosting methods can be used as they are, in essence, a combination of multiple decision trees. These will be discussed later on in this chapter, but first single decision trees are summarised. Despite there being various methods for building a decision tree, the classification and regression tree (CART) algorithm devel-

oped by Leo Breiman in 1984 is the most renowned (Breiman, 2017). This algorithm caters for two different kinds of problems, regression and classification, which must be distinguished before creating the tree, and in this thesis I focus on regression because the response variable, property price, is numerical rather than categorical. A decision tree splits the training data into various homogeneous subgroups, where the response variables in each subgroup are very similar to each other. Then the average response value of the observations for each subgroup is used to predict any test set observation assigned to that group. These subgroups, also known as nodes, are constructed by partitioning the training data based on values of covariates using simple binary style splitting rules such as $x < c$ or $x \geq c$, where each observation is allocated to one of two subgroups depending on which group it fits in. As this process is repeated, a structure which visually resembles a tree occurs with numerous nodes and branches. The root node at the top of the tree consists of all the training data before any partitioning has begun. Then each node is split into z further nodes until a stopping criteria has been reached and the nodes at the bottom are known as terminal nodes. Any nodes that lie between the root node and the terminal nodes are referred to as internal nodes as can be seen in Figure 4.1 which is taken from Chapter 9 of Hands-on Machine Learning with R textbook (Boehmke and Greenwell, 2019).

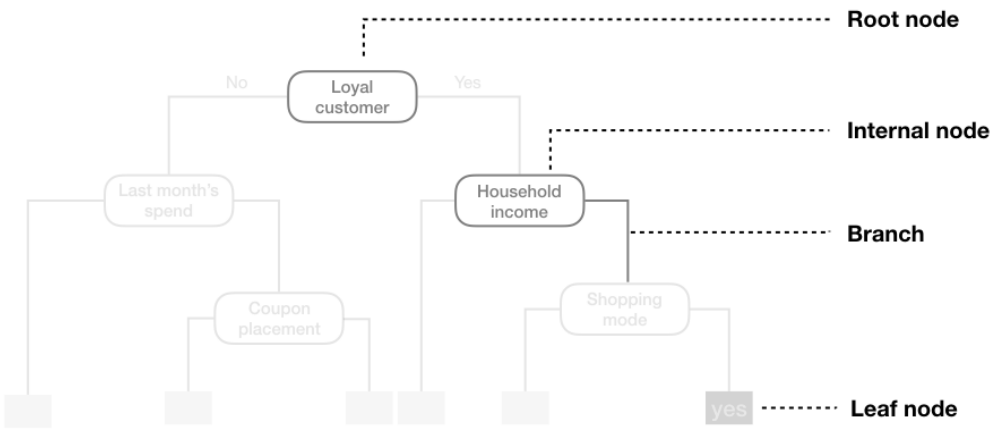

Figure 4.1: Diagram of the structure of a decision tree identifying the 3 different types of nodes - root, internal, and leaf.

4.2.1 Partitioning

The construction of a decision tree requires the training data to be split into subsets, so CART uses binary recursive partitioning which means that the splitting at each node is dependent on how the data are split at the nodes above. Therefore, put simply it

is a process that depends on the results of the previous partitioning. The aim of this procedure is at each split to find the covariate (x_i) which best splits the data into two nodes (R_1 and R_2), so that the overall error between the response variable (y_i) and the prediction (c_i) is minimized as much as possible, where c_i is the mean of the responses that fall in R_i. The following sum of squared errors (SSE) is minimised at each split,

$$SSE = \sum_{i \in R_1} (y_i - c_1)^2 + \sum_{i \in R_2} (y_i - c_2)^2. \tag{4.1}$$

By repeating this process until the tree is 'too complex' (has too many terminal nodes), an overfitted tree is achieved. This tree is then simplified by pruning as discussed below. It is important to note that when dealing with multiple covariates, it is possible that a single covariate can dominate and be used multiple times to partition the data. For example, as the tree grows, this specific covariate can be used repeatedly to find the optimal split in the data.

4.2.2 Creating an optimal tree

Decision trees can be of any size, small or large, which is why it can be complicated to find a structure that provides an optimal prediction. Sometimes trees can be too large and too overcomplicated, which can lead to the data being overfit, leading to non-optimal prediction in the test set. On the contrary, there can be trees which only partition the data once, leading to an inaccurate prediction. In both these circumstances a problem of poor predictive performance occurs. To tackle this problem the complexity/size of the tree must be chosen to obtain the optimal predictive performance. Two methods used to create the best decision tree for prediction are early stopping and pruning.

Early stopping consists of applying different types of growth restrictions to the tree such as, restricting the depth of the tree by implementing a strict rule on the number of levels it has, or limiting the minimum number of observations allowed in a terminal node. By limiting the number of levels this could result in a shallow tree which means there is less variance in the predictions. However, sometimes a tree being too shallow can lead to too much bias and the patterns and interactions within the data cannot be picked up. Similarly, allowing the minimum number of observations in each terminal node to be too small can lead to a high variance which will result in the predictions being good for the training set but not for the test set. Similarly, having too large a minimum number of observations in a terminal node can reduce the variance and like before when restricting the levels, this will result in a shallow tree which is unable to properly encapsulate trends in the data. Although both of these restrictions are independent, they still have

an effect on each other, i.e. setting a minimum value of 10 observations in each terminal node will have an impact on how many levels the tree has. A disadvantage of the early stopping method is that it relies on human decisions as the size and structure of the tree all depends on the values chosen for the restrictions.

Pruning is an alternative method to early stopping, which is when a very large tree is created first, and is then pruned back, removing branches that are not particularly important, in order to generate an optimal subtree for prediction. This subtree is constructed using a cost complexity parameter, α, which penalizes the SSE function in (4.1), by the number of terminal nodes it has, T. That is, one minimises

$$SSE + \alpha|T|, \tag{4.2}$$

which aims to balance out fit to the data via a small SSE (favouring a complex tree) against simpler trees via a small $\alpha|T|$. Then, the best tree is chosen by optimising (4.2) with respect to α. This is done by comparing multiple different trees constructed based on various values of α through cross validation leading to the best tree in a predictive sense. Once these results are compared with one another and the value of α which minimizes (4.2) is identified, predictions on the test set can be evaluated.

4.2.3 Prediction using decision trees

A basic decision tree can be constructed for the property price data, and its predictive performance can be evaluated using the same prediction metrics as earlier, namely root mean square error (RMSE) and median absolute error (MAE). This method is applied to all 5 of the data splits with α chosen through 10-fold cross validation on the training set in order to predict the test set. Then, its predictive ability can be compared with the other prediction models in this thesis, such as the original linear model and the spatial CAR model with $k = 7$, to determine which model is the most accurate for predicting property price in Scotland. It is worthwhile to note that when producing a decision tree, the same set of covariates is used as the linear model i.e. with easting and northing included again unlike in the spatial model. This is because there is not an explicit use of space in decision trees hence it cannot be captured in any other way.

Figure 4.2 is an example of the property price predictions for one of the 5 data splits using a basic decision tree. This decision tree has 8 terminal nodes, otherwise known as prediction regions, with each node being represented by one of the vertical lines on the graph. On inspection, this tree seems unsuitable because there are so few terminal nodes

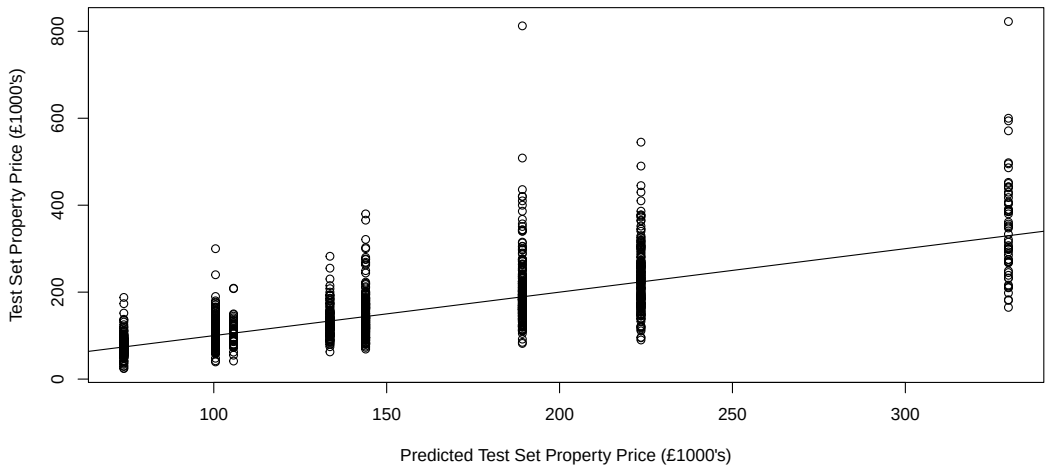

Figure 4.2: Scatterplot of property price predictions (x-axis) vs true property prices (y-axis) for one of the test sets of tree depth 4 with 8 terminal nodes.

despite there being 1253 observations. This means that there could be more than 150 observations in a prediction region, depending on the characteristics of the Data Zone and the path it follows in the tree, which in this case is around $1/8^{th}$ of the total number of observations. A better and more accurate prediction would be for Figure 4.2 to resemble more of a random scatter instead of distinct vertical lines and the tree to have more depth hence more terminal nodes leading to less observations in each prediction region.

Table 4.1 shows how the decision tree predictions compare to the linear model and the spatial model with $k = 7$. In all 5 data splits, it is clear that the RMSE and MAE values are notably higher for the decision tree in comparison to the two other models. This highlights that the linear model and the spatial model both have much better predictive performance than a simple decision tree across all data splits. In order to easily compare the models with each other, the average values of RMSE and MAE for each of the models are calculated and are displayed at the bottom of Table 4.1. In general, the linear and spatial models outperform the decision tree because they have significantly lower average errors. The RMSE of the tree is around £10,000 higher than the linear model and around £15,000 higher than the spatial model. Moreover, the MAE further highlights the difference between the models as the tree has an MAE value £5,000 higher than the linear model and £7,000 higher than the spatial model. As these values are largely different from one another and the aim is to have the lowest RMSE and MAE value possible, this suggests that the basic decision tree has poor predictive performance so adaptations that combine multiple trees are now considered.

Table 4.1: Table showing root mean square errors (RMSE) and median absolute errors (MAE) of each of the 5 data splits using a decision tree, a linear model and a spatial model where $k = 7$.

Split	Model	RMSE	MAE
	Tree	£59,625	£25,504
1	Linear	£47,995	£20,038
	Spatial($k = 7$)	£44,068	£18,386
	Tree	£58,692	£24,581
2	Linear	£48,973	£19,902
	Spatial($k = 7$)	£43,775	£18,111
	Tree	£55,160	£25,259
3	Linear	£45,631	£20,910
	Spatial($k = 7$)	£40,122	£17,431
	Tree	£54,119	£25,195
4	Linear	£43,524	£18,704
	Spatial($k = 7$)	£39,439	£17,624
	Tree	£56,049	£23,683
5	Linear	£45,875	£18,984
	Spatial($k = 7$)	£40,861	£17,244
	Tree	£56,729	£24,845
Average	Linear	£46,400	£19,708
	Spatial($k = 7$)	£41,653	£17,759

4.3 Bagging

Bootstrap aggregating, otherwise known as bagging, is an algorithm which combines multiple decision tree models and takes an average of the predictions from these trees. This assists in increasing the accuracy of the predictions whilst also reducing the variance.

Firstly, b bootstrap samples of the training data of size n are obtained, where sampling is done with replacement. Then a decision tree is constructed for each bootstrapped sample. Thus, the resulting number of decision trees created is equal to the number of bootstrap samples chosen, b. Test set predictions are made for each decision tree separately and a final bagged prediction, $\hat{f_{bag}}$, is generated by combining the individual predictions and taking an average. The process is given by the following equation,

$$\hat{f_{bag}} = \frac{\hat{f}_1 + \hat{f}_2 + \ldots + \hat{f}_b}{b}, \tag{4.3}$$

where $\hat{f}_1 \dots \hat{f}_b$, are the predictions from the decision trees using b bootstrapped data samples. In this setting the decision trees constructed are unpruned, as this allows the variance to be kept high and the bias low, therefore optimal bagging will occur when averaging over the b trees. In turn, bagging will reduce the variance overall because of the averaging process. Generally, the more decision trees that there are, the better the prediction will be, but there will be a point reached where the prediction RMSE will stabilize and even if more trees are added, there will be very little, if any, improvement that is significant.

Table 4.2 shows how the out-of-sample RMSE and MAE for the 5 data splits compare for 50, 100 and 150 bootstrapped samples using the bagging method. Overall, there is not a large difference in the RMSE values between each of the 3 bootstrap values, generally less than £1,000. As between 50 and 150 trees there is very little improvement in the RMSE value, this would imply that the optimal number of bootstrap samples required for the best predictive performance is somewhere towards the larger end of this range. This is because as we reach towards 150 trees, the improvement in RMSE between the trees becomes smaller.

Table 4.2: Table showing the root mean square errors (RMSE) and median absolute errors (MAE) of each of the 5 data splits using bagging and based on 50, 100, and 150 bootstrapped samples.

	Number of Bootstraps					
Split	50		100		150	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
1	£45,740	£17,895	£44,857	£17,912	£44,745	£18,037
2	£45,421	£18,334	£44,711	£17,673	£44,805	£17,487
3	£41,499	£18,447	£41,044	£17,517	£41,175	£18,793
4	£39,601	£16,372	£39,889	£17,054	£39,795	£17,049
5	£43,122	£18,119	£42,252	£17,615	£42,131	£17,490
Average	£43,077	£17,833	£42,551	£17,554	£42,530	£17,771

An important factor surrounding these results is that there is an obvious spread in RMSE values between the 5 data splits. For example, in the case where there are 50 bootstrapped samples, Split 4 has the smallest RMSE of £39,601 while Split 1 has the largest, £45,740, which is a difference of over £6,000. This could be due to the dynamics of each split i.e. Split 1 may have more higher priced areas in the test set, which are generally predicted less well than lower priced areas. On the other hand, the MAE values do not follow this pattern and instead are much closer in value to each other. This may be because any high priced outliers in the test set will have no effect on the median so if

there is an extremely high property price in one of the test set splits, this will only have an effect on the RMSE value and not on the MAE value.

Figure 4.3 is a graphical example of the predicted property prices for one of the 5 data splits using 50 bootstrap samples. When comparing this to Figure 4.1 which only used a single decision tree, there is now a linear relationship which is the desired outcome. Hence, this indicates that the predictive performance is better and more accurate when using the bagging method where multiple trees are combined and averaged over rather than using one single decision tree on its own. This can be seen by comparing Tables 4.1 and 4.2, which show that the bagging method, no matter the number of bootstrap samples, has better predictions of property prices than both a single decision tree and the linear model. However, using the spatial CAR model with $k = 7$ is the best overall so far for prediction, having an RMSE that is lower by £877 as seen in Table 4.6.

Despite the results in Table 4.2 and Figure 4.3 showing that the bagging method is fairly good at prediction, it does have a crucial weakness – namely the individual trees are similar. Even though the data in each bootstrap sample is randomized and no 2 bootstrap samples are the same, the same covariates are still being used to create each tree. Therefore, this means that all of the decision trees will likely follow a very similar structure at the top of the trees and will only differ towards the bottom. This will lead to similar trees being constructed due to the strong impact of certain covariates on the data. Hence, bagging prevents the variance from being further reduced. In order to tackle this problem, the random forest and gradient boosting methods, which are an extension of decision trees, can be used and evaluated to see if they can produce better predictions.

4.4 Random Forests

Random forests are algorithms that offer improved predictive performance over decision trees and potentially bagging. They follow the same fundamentals as bagged decision trees but encourage more randomness in the construction of the trees by reducing the between tree correlations. Due to the between tree correlation being present in bagging, this prevents the overall variance from being further reduced, so to combat this random forests reduce correlation by using less correlated trees. There are many different ways of structuring and creating a random forest but the most popular, which will be used in this thesis, is Leo Breimann's method (Breiman, 2001). This method allows for very few tuning parameters whilst also normally decreasing the error through randomisation.

4.4.1 Structure

Similar to all of the previous models, the random forest algorithm is applied separately to each of the 5 original data splits, being fitted and trained on the training set and subsequently used to predict the test set. The construction of a random forest begins the same as bagging where the desired number of trees is chosen, and a bootstrap sample of the training data is generated to create each decision tree. Then for a given tree in order to split each node into two further nodes, instead of the whole set of covariates being considered, each node is limited to a fraction of the set of covariates, m_{try}, with covariates allocated to this small set randomly. The number of covariates considered for use at each node split is set beforehand, with each node having the same number but a different selection of possible covariates for making the split. The covariate which has the strongest relationship with the data at each node will be used as the variable on which the node split is determined on. This means that the covariates which determine the node split could differ at every split depending on which subset of covariates are considered at that split. The tree will continue to grow until the stopping criteria is reached, and in the case of random forests, the minimum number of observations in each node is fixed by the user. Once this is reached we have a full decision tree, and this same process is repeated for each of the bootstrap data samples until the number of desired trees is obtained. The final prediction can then be calculated by combining the predictions of each individual tree in the random forest and taking an average (Boehmke and Greenwell, 2019).

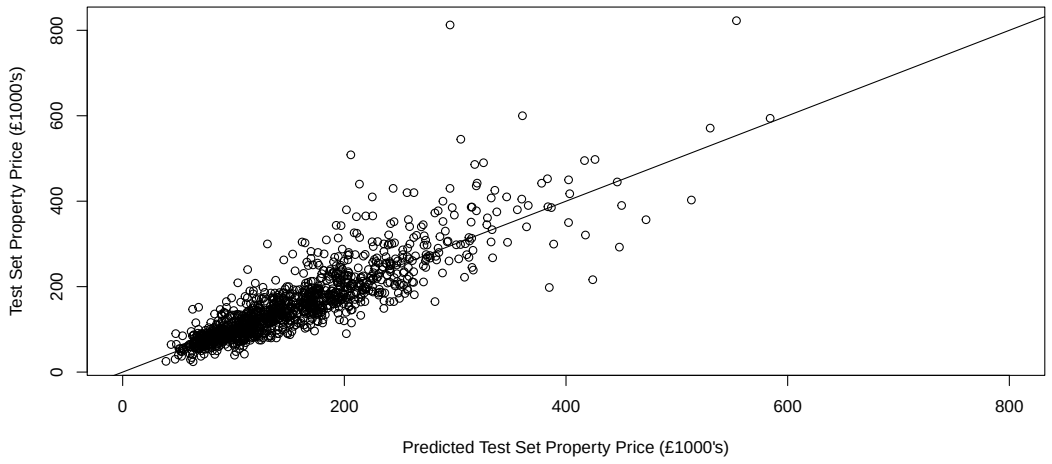

Figure 4.3: Scatterplot of property price predictions for one of the test sets using bagging based on 150 bootstrap samples.

4.4.2 Tuning Parameters

The m_{try} tuning parameter controls the number of variables on which the nodes could be split on, and because of the induced randomization in the covariate selection process at each node, will help to keep the trees decorrelated (James et al., 2013). Typically, a sensible value of m_{try} is $\frac{p}{3}$ where p represents the total number of covariates in the full set (Boehmke and Greenwell, 2019). Hence this means that only one third of the covariates is considered at each split, and of this third the covariate which has the strongest relationship with the response variable will be the variable on which the node split is determined on at each split in order to grow the tree. The value of m_{try} can be altered to suit different data sets. For example, data sets with very few covariates of large importance will generally benefit from using a higher value of m_{try} as this will increase the likelihood of at least one of the covariates of importance being considered as the variable which determines the node split (Boehmke and Greenwell, 2019).

For this property price data set, I will investigate different values of m_{try}, from low to high, in order to see the impact that the value of m_{try} has on predicting property prices. Firstly, the value of m_{try} is $\frac{p}{10}$ with 10% of the full set of covariates being considered, then this value will be increased in 10% increments, rounded to the nearest whole number, until the full set of covariates is reached and it will in turn be equivalent to bagging.

Additionally, the number of trees in a random forest is another tuning parameter that plays a crucial role in determining the predictive ability of the model. As seen previously when comparing the single decision tree versus the bagging method in Section 4.3, where multiple trees are used, as the number of trees increase the RMSE will decrease because the accuracy in prediction has improved due to a decreased error rate (Boehmke and Greenwell, 2019). Therefore, when building random forest models on this property price data set, I will investigate how 50, 100 and 150 trees compare, the same values used in the bagging method. This will allow these machine learning methods to be compared fairly. Thus there are 30 possible tuning parameter combinations, which are all combinations of $b = 50, 100, 150$ and $m_{try} = 2, 4, 6, 9, 11, 13, 16, 18, 20, 23$. The final tuning parameter in the model, the number of minimum observations in each terminal node, is a fixed parameter and will be set at 2 for all combinations.

4.4.3 Choosing the tuning parameter combination

Before making any predictions on the test set, the optimal values of the tuning parameters, namely the number of trees (b) and the number of covariates (m_{try}), must be chosen

by evaluating the training-validation set. Similar to the spatial CAR model shown in Section 3.2.1, the data are randomly split into 5 training-test data splits using the splitting process shown in Figure 2.10. Then, the test sets for each of the 5 splits are removed and random forest models are created using the training sets. For the purposes of choosing the optimal tuning parameter combination, in each split the training set is further partitioned into 10 smaller sub-splits of roughly equal size. The random forest is subsequently fit to 9/10 of the training set for each combination of tuning parameters, and is then used to predict the 10th, the validation set. This process is repeated 10 times, with a different validation set each time, using 10-fold cross validation. In turn, this will lead to 30 different combinations of results being generated per split and Tables 4.3 and 4.4 show the values of root mean square error (RMSE) and median absolute error (MAE) obtained by each combination.

From Table 4.3 the best value of RMSE across the 5 splits is £42,805 which is obtained when the value of m_{try} is 20 and b=150. On the contrary, the least optimal combination has a RMSE of £46,453 and occurs when m_{try}=2 and b=50 creating an overall range of approximately £3,600. This would suggest that increasing both b, the number of trees, and m_{try} will decrease the RMSE value and hence provide a more accurate prediction. Moreover, Table 4.4 shows that the MAE has a slightly different optimal combination, as the lowest MAE value of £17,380 is obtained when m_{try}=13 and b=100. The least optimal combination outputs a MAE of £18,713 which is again when m_{try} is 2 and b=50. These results highlight that it is highly likely that there are certain covariates that have a stronger impact on property price than others, and hence when a small value of m_{try} is used this could lead to unimportant covariates being used as the split variable. Therefore, since the best combination of tuning parameters are not the same for the RMSE and MAE metrics, we will continue to make predictions on the test set in the following section using both (m_{try}=20, b=150) and (m_{try}=13, b=100).

4.4.4 Test Set Predictions

The random forest model is refit to the entire training set firstly with (m_{try}= 20, b=150), and then with (m_{try}=13, b=100), and property prices are predicted for the test set using the fitted model for each of the 5 splits. The resulting RMSE and MAE values can be seen in Table 4.5, where the combinations can be compared with one another and averages are calculated over all 5 data splits. In general, the values of RMSE are very similar to each other across all 5 splits with very little difference between the two combinations. Likewise, the MAE values for the 5 splits are all close in value to one another. In 4 out of 5 of the splits the m_{try}=20 and b=150 combination has a lower RMSE than the m_{try}=13 and

Table 4.3: Table showing the root mean square error (RMSE) of property price for each combination of tuning parameters when building a random forest applied to the 5 data splits.

m_{try}	b	Split 1	Split 2	Split 3	Split 4	Split 5	Average
2	50	£45,840	£45,966	£46,869	£46,706	£46,886	£46,453
4	50	£44,218	£43,844	£44,953	£45,055	£44,404	£44,495
6	50	£43,347	£43,567	£44,201	£44,247	£43,322	£43,467
11	50	£42,881	£42,873	£43,621	£43,706	£43,229	£43,262
13	50	£42,947	£42,833	£43,429	£43,762	£43,589	£43,312
16	50	£42,781	£43,129	£43,591	£43,166	£43,165	£43,167
18	50	£42,643	£42,846	£43,377	£43,418	£43,498	£43,156
20	50	£42,308	£42,941	£43,642	£43,018	£43,199	£43,021
23	50	£42,675	£42,911	£43,425	£43,339	£43,019	£43,074
2	100	£45,619	£45,581	£46,640	£46,444	£46,236	£46,104
4	100	£43,975	£43,656	£44,769	£44,800	£44,258	£44,292
6	100	£43,292	£43,279	£43,972	£44,035	£43,703	£43,656
9	100	£42,902	£42,859	£43,791	£43,566	£43,190	£43,261
11	100	£42,668	£42,602	£43,396	£43,542	£43,030	£43,048
13	100	£42,616	£42,514	£43,431	£43,693	£43,325	£43,116
16	100	£42,617	£42,782	£43,391	£43,182	£43,052	£43,005
18	100	£42,490	£42,698	£43,430	£43,203	£43,240	£43,012
20	100	£42,196	£42,663	£43,451	£43,061	£42,960	£42,866
23	100	£42,403	£42,673	£43,483	£43,253	£43,145	£42,991
2	150	£45,420	£45,525	£46,531	£46,395	£46,144	£46,003
4	150	£43,790	£43,584	£44,780	£44,677	£44,207	£44,208
6	150	£43,300	£43,169	£44,025	£44,056	£43,611	£43,632
9	150	£42,793	£42,746	£43,724	£43,570	£43,140	£43,195
11	150	£42,636	£42,594	£43,386	£43,484	£42,982	£43,016
13	150	£42,403	£42,456	£43,327	£43,511	£43,124	£42,964
16	150	£42,524	£42,699	£43,396	£43,102	£42,978	£42,940
18	150	£42,425	£42,657	£43,331	£43,188	£43,158	£42,952
20	**150**	£42,184	£42,534	£43,372	£43,087	£42,846	**£42,805**
23	150	£42,427	£42,593	£43,427	£43,251	£43,088	£42,957

b=100 combination. On the other hand, for all 5 splits the 13-100 combination provides the lowest MAE values. Overall when taking a look at the averages, the values are very similar for both combinations with a less than £100 difference in RMSE and a slightly larger difference, around £400, in MAE between the combinations. The results are very similar to the results obtained by using the bagging method, because m_{try} is relatively large, suggesting that the same covariates are used as the node splitting variable the majority of the time because they have the most influence on property prices. Therefore, the individual trees in the random forest will be very structurally similar to the bagged trees.

Table 4.4: Table showing the median absolute error (MAE) of property price for each combination of tuning parameters when building a random forest applied to the 5 data splits.

m_{try}	b	Split 1	Split 2	Split 3	Split 4	Split 5	Average
2	50	£18,920	£18,617	£18,445	£19,065	£18,521	£18,713
4	50	£18,160	£17,853	£17,741	£18,099	£18,063	£17,983
6	50	£17,750	£17,596	£17,497	£18,302	£17,625	£17,754
9	50	£17,393	£17,810	£17,445	£18,111	£17,910	£17,734
11	50	£17,222	£17,680	£17,558	£17,713	£17,734	£17,581
13	50	£17,739	£17,788	£17,192	£17,464	£17,441	£17,525
16	50	£17,908	£17,669	£17,520	£17,732	£17,735	£17,713
18	50	£17,500	£17,302	£17,419	£17,593	£17,953	£17,553
20	50	£17,689	£17,510	£17,581	£17,645	£17,548	£17,595
23	50	£17,832	£18,059	£17,465	£18,044	£17,521	£17,784
2	100	£18,683	£18,051	£18,452	£18,839	£18,369	£18,479
4	100	£17,918	£17,842	£17,638	£17,993	£17,887	£17,856
6	100	£17,463	£17,498	£17,501	£18,088	£17,564	£17,623
9	100	£17,485	£17,728	£17,394	£17,720	£17,638	£17,593
11	100	£17,190	£17,691	£17,473	£17,687	£17,504	£17,509
13	**100**	£17,174	£17,582	£17,191	£17,594	£17,358	**£17,380**
16	100	£17,626	£17,520	£17,422	£17,500	£17,519	£17,518
18	100	£17,446	£17,577	£17,366	£17,635	£17,772	£17,559
20	100	£17,808	£17,464	£17,554	£17,656	£17,337	£17,564
23	100	£17,733	£17,642	£17,347	£17,916	£17,751	£17,678
2	150	£18,610	£18,008	£18,376	£18,760	£18,200	£18,391
4	150	£17,635	£17,834	£17,426	£18,128	£17,863	£17,777
6	150	£17,547	£17,423	£17,424	£17,783	£17,605	£17,556
9	150	£17,310	£17,400	£17,319	£17,538	£17,694	£17,452
11	150	£17,116	£17,381	£17,498	£17,589	£17,467	£17,410
13	150	£17,219	£17,609	£17,095	£17,648	£17,374	£17,389
16	150	£17,545	£17,625	£17,255	£17,542	£17,626	£17,519
18	150	£17,343	£17,464	£17,297	£17,560	£17,679	£17,469
20	150	£17,665	£17,400	£17,337	£17,564	£17,444	£17,482
23	150	£17,520	£17,456	£17,260	£17,899	£17,606	£17,548

Table 4.6 shows how the property price predictions of each of the 5 different models investigated so far in this thesis compare with one another. Across all 5 data splits, the spatial model, bagging and random forest are very similar to each other in terms of RMSE and MAE respectively, suggesting that these methods are much better at predicting property price in Scotland than the decision tree and the linear model are. At the bottom of Table 4.6, the averages of each of the models are displayed and this allows them to be easily comparable to each other. From these results, it is clear that the spa-

Table 4.5: Table of the root mean square errors (RMSE) and median absolute errors (MAE) of property prices of each of the 5 data splits using the random forest algorithm for 2 different combinations of tuning parameters.

Split	m_{try}	b	RMSE	MAE
1	20	150	£44,676	£18,036
	13	100	£44,391	£17,249
2	20	150	£44,430	£17,815
	13	100	£44,758	£17,304
3	20	150	£41,408	£18,890
	13	100	£41,635	£18,710
4	20	150	£40,173	£16,946
	13	100	£40,221	£16,866
5	20	150	£42,163	£17,817
	13	100	£42,258	£17,230
Average	20	150	£42,570	£17,901
	13	100	£42,653	£17,472

tial model has the lowest RMSE and MAE hence indicating that it has better predictive performance than the other models. Despite there being very little difference in MAE, £12, between the spatial model and the model created using the bagging method, the difference in RMSE between these models is much more apparent at around £1,000.

It is advantageous to investigate the feature importance plots for the machine learning methods to see which of the covariates influence the predictions. Figures 4.4 and 4.5 show the feature importance plots of the bagging model and the random forest model respectively for 1 of the 5 data splits. The random forest model is selected to be impurity based so it can be compared to the bagging model and maintains consistency. The ordering of the covariates in the feature importance plot is determined by the sum of squared errors (SSE) value of each covariate, defined in Section 4.2.1. As both the bagging and random forest methods have a large number of trees, an average SSE for each of the 5 splits is calculated for every covariate. Then, these average SSE values are added together to give a total SSE value across all 5 splits and the covariates can be ordered by importance from lowest SSE (most important) to highest SSE (least important). For further details, please see Chapter 9 of Hands-on Machine Learning with R book (Boehmke and Greenwell, 2019). Both models share the same covariates at the top of their respective plots, indicating the similarity between the models which could explain why the RMSEs and MAEs are so similar in both cases. In the random forest model, council tax clearly has a much more significant effect than any of the other covariates. Whereas, in the bagging model, council tax is not the only covariate of substantial importance as mean number of

Table 4.6: Table showing the root mean square errors (RMSE) and median absolute errors (MAE) of each of the 5 data splits using a decision tree, a linear model, a spatial model where $k = 7$, bagging and a random forest where m_{try}=20 and b=150.

Split	Model	RMSE	MAE
1	Tree	£59,625	£25,504
	Linear	£47,995	£20,038
	Spatial($k = 7$)	£44,068	£18,386
	Bagging	£44,745	£18,037
	Random Forest	£44,676	£18,036
2	Tree	£58,692	£24,581
	Linear	£48,973	£19,902
	Spatial($k = 7$)	£43,775	£18,111
	Bagging	£44,805	£17,487
	Random Forest	£44,430	£17,815
3	Tree	£55,160	£25,259
	Linear	£45,631	£20,910
	Spatial($k = 7$)	£40,122	£17,431
	Bagging	£41,175	£18,793
	Random Forest	£41,408	£18,890
4	Tree	£54,119	£25,195
	Linear	£43,524	£18,704
	Spatial($k = 7$)	£39,439	£17,624
	Bagging	£39,795	£17,049
	Random Forest	£40,173	£16,946
5	Tree	£56,049	£23,683
	Linear	£45,875	£18,984
	Spatial($k = 7$)	£40,861	£17,244
	Bagging	£42,131	£17,490
	Random Forest	£42,163	£17,817
Average	Tree	£56,729	£24,845
	Linear	£46,400	£19,708
	Spatial($k = 7$)	£41,653	£17,759
	Bagging	£42,530	£17,771
	Random Forest	£42,570	£17,901

rooms, percentage of flats and percentage of semi or detached properties all have similar importance. Overall, across both models, both property and Data Zone characteristics have large effects on the model. However, despite having little impact on the bagging model, easting and northing, both of which are physical geographical covariates, are two of the top 10 most influential covariates on the random forest, although the absolute size of their importance is still low. When referring back to Table 4.6, bagging is only slightly better at predicting property prices than the random forest, which would suggest that although they place higher in the random forest feature importance plot than they do

in bagging, easting and northing have very little impact on the model as a whole. The spatial model, which has the best RMSE and MAE so far, uses the full set of covariates and also takes into account spatial geometry via a neighbourhood matrix between the Data Zones in order to make predictions, whereas the machine learning methods directly utilise the central points of each Data Zone to represent space. These differences in the geometry used could have an effect on the model and may be a reason why the spatial model has the lowest RMSE and MAE so far.

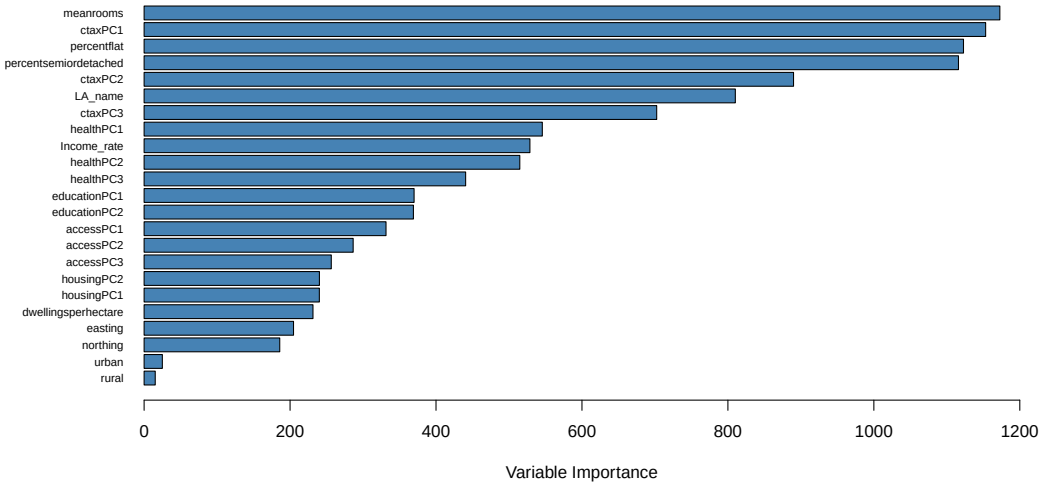

Figure 4.4: Feature Importance Plot of the bagging model.

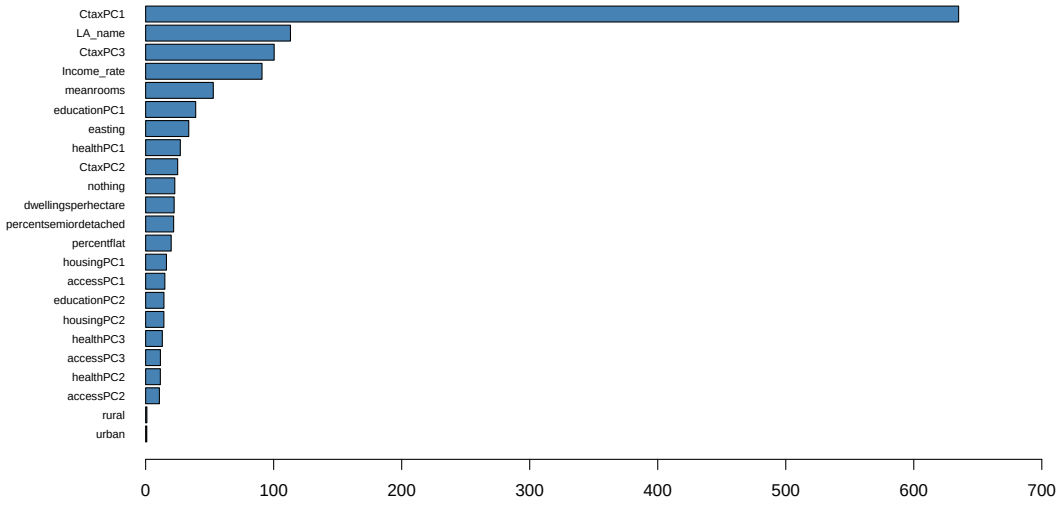

Figure 4.5: Feature Importance Plot of the impurity based random forest model.

Overall, this shows that the spatial model has outperformed the machine learning

methods of prediction, and is the method which has the best predictive performance so far with regards to the property price data set. However, it is worthwhile to explore whether gradient boosting, a machine learning method which is a further development of bagging and random forests, can outperform the spatial model and produce a more accurate model for predicting property prices in Scotland.

4.5 Gradient Boosting

The final machine learning algorithm investigated in this chapter is gradient boosting. Gradient boosting, unlike random forest and bagging where an ensemble of large deep trees are constructed, builds an ensemble of shallow trees in which each tree is an improvement on the one that precedes it (Boehmke and Greenwell, 2019). Although shallow trees are usually poor indicators of prediction on their own, if appropriate tuning parameters are considered and selected, the results of an ensemble of shallow trees using the gradient boosting method can be very successful.

4.5.1 Structure

As seen previously, the other machine learning methods investigated, bagging and random forest, create large complex trees with low bias and high variance (Breiman, 2001). By taking an average across the ensemble of trees in both of these methods it reduces the variance and hence the predictive performance improves as the error in the residuals is minimized (Breiman, 1996). Despite also being a machine learning method of prediction, gradient boosting does not follow a similar structure and instead consists of simple shallow decision trees of high bias and low variance where averaging is not required.

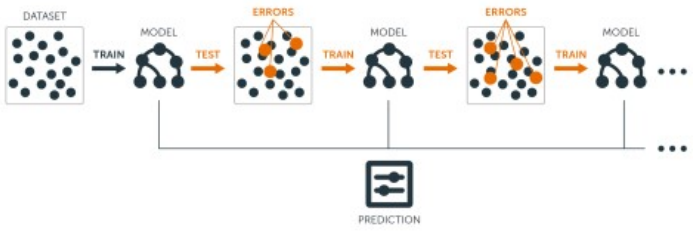

Figure 4.6: Diagram of the sequential improvement of a gradient boosting machine taken from Boehmke and Greenwell [2019].

To begin with, a weak model which is a single shallow decision tree typically with the tree depth tuning parameter of around 1 to 6 is built, and its performance is assessed through the size of the prediction errors. This performance is sequentially improved by constructing new trees where each tree will attempt to fix where the previous tree went wrong using the prediction errors. It does this by focusing in on a specific line in the training data where the largest prediction errors occur, and adapts to reduce or simply eliminate the error (Boehmke and Greenwell, 2019). This repeated process, which can be seen in Figure 4.6 taken from the Gradient Boosting chapter of Hands-on Machine Learning with R (Boehmke and Greenwell, 2019), will continue with each tree building on its predecessor until a specific stopping criteria is reached – usually the set number of trees is achieved. Initially, the errors in the residuals will be extensive and there will be lots of room for improvement, but as the process goes on this error will be sequentially reduced as the trees are continually improving through each new tree that is constructed (Boehmke and Greenwell, 2019). The process can be summarised using the following algorithm, which again is taken from Hands-on Machine Learning with R (Boehmke and Greenwell, 2019):

Algorithm 1 Algorithm for Gradient Boosting

1: **while** Stopping criteria is not reached **do**
2: Fit a decision tree to the data : $F_1(x) = y$
3: Fit the next decision tree to the residuals of the previous: $h_1(x) = y - F_1(x)$
4: Add this new tree to our algorithm: $F_2(x) = F_1(x) + h_1(x)$
5: Fit the next decision tree to the residuals of $F_2(x) : h_2(x) = y - F_2(x)$
6: Add this new tree to our algorithm : $F_3(x) = F_2(x) + h_2(x)$
7: **end while**

4.5.2 Tuning Parameters

In a basic gradient boosting model, two different types of tuning parameters are considered, boosting (the number of trees and the learning rate) and tree specific (tree depth and the minimum number of observations in terminal nodes) (Boehmke and Greenwell, 2019). The number of trees is a common tuning parameter in tree-based machine learning methods, and often plays a very important part in determining predictive performance (Breiman, 1996). The aim of gradient boosting machines is to create an ensemble of trees in which each tree is an improvement on the tree created before it. This process can go on for as long as allowed, in some cases creating thousands upon thousands of trees, which can often lead to over-fitting because an average is not taken. Therefore, it is important to find an optimal number of trees which minimizes the RMSE and MAE values in an ap-

propriate amount of time as sometimes these processes can be very time consuming. After evaluating this tuning parameter in the bagging and random forest methods with 50, 100, and 150 trees, it was concluded that the optimal number of trees was 150 because, in general, this number produced the lowest RMSE and MAE values. Hence, in order to be consistent and since gradient boosting machines often require a large number of trees, the number of trees parameter is set so that 150 trees are constructed using gradient boosting.

Another boosting specific tuning parameter is the learning rate (l_r), otherwise known as shrinkage, which is the value of the contribution of each tree to the final outcome (Boehmke and Greenwell, 2019). This is a value between 0 and 1, however typically it lies between 0.001 and 0.3. In order to find the optimal value of the learning rate, larger values such as 0.3 should be considered first, and then the value should be reduced progressively in increments until the optimum is achieved (Boehmke and Greenwell, 2019). In general, as the learning rate decreases, the accuracy of the model will increase but more trees will be required. For the property price data set, the learning rate will be evaluated at 0.3, 0.1, 0.05, 0.01, 0.005, and 0.001 with 150 trees as recommended in Chapter 12.3 of Hands-on Machine Learning with R (Boehmke and Greenwell, 2019).

In gradient boosting, the depth of each of the trees (t_d) must be managed as it can have an impact on the final outcome. It is possible for trees to be very large and have only one observation in each terminal node, but this is often an inaccurate depiction of the data as it is an exact copy of the training set so tree depth is usually set between 1 and 8 with the majority being between 3 and 8 (Boehmke and Greenwell, 2019). Shallow trees of depth 1 or 2, essentially stumps, will be successful if a large number of them are constructed, but deeper trees are generally preferred for large data sets and allow for special relationships to be captured by the gradient boosting machines (Boehmke and Greenwell, 2019). They do however, increase the risk of over-fitting if they are too large. Therefore, to find the optimal value using gradient boosting machines in predicting property prices, tree depth of sizes 1, 3, 5, and 8 will be assessed.

The final tuning parameter to be set in the gradient boosting machines is the minimum number of observations in each terminal node. This tuning parameter is the least influential of the four as it does not have a significant impact on predictive performance (Boehmke and Greenwell, 2019). Since both the bagging and random forest methods considered set this parameter at a value of 2, to stay consistent in the gradient boosting machine it will also be set at 2.

4.5.3 Choosing the tuning parameter combination

Like the random forest and CAR models, before any predictions are made on the test set, the optimal combination of values of the tuning parameters, specifically l_r and t_d, must be chosen by evaluating the training-validation set. It should be noted that the two other tuning parameters used in gradient boosting, minimum number of observations in terminal nodes and number of trees, are assumed to be fixed values here.

The data are split at random into 5 training-test data splits, and the tuning parameters are estimated in the same way as the previous models, which is omitted here for brevity. This produces 24 combinations of results for each of the 5 data splits seen in Tables 4.7 and 4.8 which show the root mean square error and median absolute error obtained by each combination respectively.

Table 4.7: Table showing the root mean square error (RMSE) of property price for each combination of tuning parameters using the gradient boosting method applied to the 5 data splits.

t_d	l_r	SPLIT 1	SPLIT 2	SPLIT 3	SPLIT 4	SPLIT 5	Average
1	0.3	£45,694	£45,302	£46,282	£46,995	£45,844	£46,023
1	0.1	£46,460	£46,324	£47,187	£47,647	£46,712	£46,866
1	0.05	£49,427	£49,415	£50,189	£50,746	£49,997	£49,955
1	0.01	£69,315	£69,093	£70,460	£70,559	£70,201	£69,926
1	0.005	£76,698	£76,489	£78,034	£78,124	£77,749	£77,419
1	0.001	£86,684	£86,686	£87,989	£88,396	£87,870	£87,525
3	0.3	£44,096	£43,250	£44,309	£44,784	£44,535	£44,195
3	0.1	£43,788	£43,538	£44,241	£44,691	£43,846	£44,021
3	0.05	£45,081	£45,071	£45,728	£45,946	£45,523	£45,470
3	0.01	£60,571	£60,486	£61,657	£61,842	£61,358	£61,183
3	0.005	£71,228	£71,145	£72,445	£72,688	£72,188	£71,939
3	0.001	£85,475	£85,505	£86,764	£87,207	£86,655	£86,321
5	0.3	£44,343	£44,176	£45,273	£44,655	£44,008	£44,491
5	0.1	£42,881	£42,175	£43,435	£43,484	£43,092	£43,013
5	0.05	£43,718	£43,505	£44,372	£44,510	£44,197	£44,060
5	0.01	£57,588	£57,625	£58,607	£58,987	£58,512	£58,264
5	0.005	£68,983	£68,958	£70,180	£70,537	£70,058	£69,743
5	0.001	£84,966	£84,990	£86,263	£86,719	£86,182	£85,824
8	0.3	£44,991	£44,984	£46,412	£46,129	£45,609	£45,625
8	**0.1**	£42,040	£41,903	£43,284	£43,093	£42,562	**£42,576**
8	0.05	£42,855	£42,380	£43,516	£43,388	£43,308	£43,089
8	0.01	£55,490	£55,489	£56,362	£56,682	£56,332	£56,071
8	0.005	£67,417	£67,403	£68,492	£68,853	£68,440	£68,121
8	0.001	£84,571	£84,592	£85,859	£86,308	£85,764	£85,419

Table 4.8: Table showing the median absolute error (MAE) of property price for each combination of tuning parameters using the gradient boosting method applied to the 5 data splits.

t_d	l_r	SPLIT 1	SPLIT 2	SPLIT 3	SPLIT 4	SPLIT 5	Average
1	0.3	£19,318	£19,353	£19,333	£20,317	£19,698	£19,604
1	0.1	£19,829	£19,984	£19,886	£20,615	£20,153	£20,093
1	0.05	£20,453	£20,558	£20,139	£20,600	£20,707	£20,491
1	0.01	£28,253	£27,810	£27,995	£28,270	£27,966	£28,059
1	0.005	£33,459	£32,438	£32,910	£33,089	£32,990	£32,977
1	0.001	£43,402	£42,640	£43,033	£43,849	£43,459	£43,276
3	0.3	£18,812	£18,782	£18,633	£19,027	£18,546	£18,760
3	0.1	£18,193	£18,322	£18,077	£18,703	£18,187	£18,296
3	0.05	£18,895	£18,952	£18,899	£19,187	£18,772	£18,941
3	0.01	£24,272	£24,130	£24,264	£24,527	£24,334	£24,306
3	0.005	£30,434	£29,653	£30,205	£30,252	£29,932	£30,095
3	0.001	£43,415	£42,549	£43,034	£43,757	£42,808	£43,113
5	0.3	£19,276	£18,877	£18,962	£19,173	£18,934	£19,044
5	0.1	£17,783	£17,691	£17,971	£18,160	£17,866	£17,894
5	0.05	£17,845	£18,249	£18,231	£18,612	£17,929	£18,173
5	0.01	£22,957	£22,498	£22,522	£22,965	£22,772	£22,743
5	0.005	£29,035	£28,733	£28,924	£29,095	£28,743	£28,906
5	0.001	£43,035	£42,480	£42,760	£43,271	£42,640	£42,837
8	0.3	£19,282	£19,321	£19,217	£19,993	£19,297	£19,422
8	**0.1**	£17,835	£17,490	£17,602	£17,677	£17,228	**£17,566**
8	0.05	£17,422	£17,828	£17,526	£17,848	£17,448	£17,614
8	0.01	£21,918	£21,538	£21,630	£22,029	£21,626	£21,748
8	0.005	£28,344	£27,783	£27,973	£28,362	£27,919	£28,076
8	0.001	£42,858	£42,204	£42,388	£42,813	£42,349	£42,522

Table 4.7 shows that the lowest value of RMSE across the 5 splits is £42,576 which is obtained when the tree depth is 8 and the learning rate is 0.1. On the other hand, the highest RMSE is £87,525, when the tree depth is 1 and learning rate is 0.001. As there is a £44,949 difference in RMSE, this shows that in the case of the property price data set, shallow trees are insufficient and the learning rate, when it is too small, leads to poor predictive performance. Interestingly, when examining the splits individually, all 5 splits produce the lowest RMSE using the (tree depth= 8, learning rate= 0.1) combination and the highest RMSE when the (tree depth=1, learning rate=0.001) combination is used. This suggests that there is a clear difference between the combinations and highlights how sensitive the results are to small changes in the tuning parameters. When inspecting Table 4.8 it can be seen again that the combination which generates the optimal MAE of £17,566 is when the tree depth is 8 and the learning rate is 0.1. Likewise, the least optimal combination is when the tree depth is 1 and the learning rate is 0.001 which

yields a MAE of £43,276, £25,710 more than the best value. Overall these results conclude that the best combination of tuning parameters is when the tree depth is set at 8 and the learning rate is 0.1, therefore predictions on the test set will be made using this combination.

4.5.4 Test Set Predictions

The gradient boosting model is refit to the entire training set with $l_r = 0.1$ and $t_d = 8$ and property price predictions can be made on the test set using the fitted model for each of the 5 training/test splits. The resulting RMSE and MAE values can be seen in Table 4.9 alongside the results from all of the other methods of prediction evaluated previously, with their respective averages displayed at the bottom of the table. In general, for gradient boosting the RMSE values are very similar to each other across the 5 splits and follow a similar pattern to the other tree-based machine learning methods with Split 4 having the lowest RMSE of the 5. Likewise, the MAE values across all 5 splits are also very similar to each other.

4.6 Discussion

In conclusion, when looking at the tree-based machine learning methods, all 3 of the ensemble tree-based methods have very similar predictive performance as their RMSE and MAE values are all around the same value. The single decision tree is omitted from the following discussion because it is the basis for the 3 ensemble methods and is expected to perform poorly compared to the other methods due to the lack of tuning parameters. As there is a substantial difference in the RMSE and MAE values of the single tree compared to the ensemble methods, this highlights that ensembles of trees are much more successful in prediction and that tuning parameters have a positive impact on predictive ability.

The average RMSE and MAE values for each of the methods across all 5 data splits are displayed at the bottom of Table 4.9. Overall, with regards to the machine learning methods, bagging has the lowest RMSE value of £42,530 which is £40 less than random forest and £296 less than gradient boosting. The prediction metric RMSE penalises more heavily for outliers suggesting that bagging is therefore the best method of the three as it deals best with outliers in this particular data set. However, the differences in RMSE between the methods are all very small in comparison to the large values of the property price data, showing that there is actually very little between the 3 machine learning

Table 4.9: Table showing the root mean square errors (RMSE) and the median absolute errors (MAE) of each of the 5 data splits using a decision tree, a linear model, a spatial model where $k = 7$, bagging, a random forest where m_{try}=20 and b=150 and gradient boosting.

Split	Model	RMSE	MAE
1	Linear	£47,995	£20,038
	Spatial($k = 7$)	£44,068	£18,386
	Tree	£59,625	£25,504
	Bagging	£44,745	£18,037
	Random Forest	£44,676	£18,036
	Gradient Boosting	£44,381	£17,815
2	Linear	£48,973	£19,902
	Spatial($k = 7$)	£43,775	£18,111
	Tree	£58,692	£24,581
	Bagging	£44,805	£17,487
	Random Forest	£44,430	£17,815
	Gradient Boosting	£45,686	£17,389
3	Linear	£45,631	£20,910
	Spatial($k = 7$)	£40,122	£17,431
	Tree	£55,160	£25,259
	Bagging	£41,175	£18,793
	Random Forest	£41,408	£18,890
	Gradient Boosting	£41,225	£17,888
4	Linear	£43,524	£18,704
	Spatial($k = 7$)	£39,439	£17,624
	Tree	£54,119	£25,195
	Bagging	£39,795	£17,049
	Random Forest	£40,173	£16,946
	Gradient Boosting	£39,688	£16,730
5	Linear	£45,875	£18,984
	Spatial($k = 7$)	£40,861	£17,244
	Tree	£56,049	£23,683
	Bagging	£42,131	£17,490
	Random Forest	£42,163	£17,817
	Gradient Boosting	£43,150	£18,726
Average	Linear	£46,400	£19,708
	Spatial($k = 7$)	£41,653	£17,759
	Tree	£56,729	£24,845
	Bagging	£42,530	£17,771
	Random Forest	£42,570	£17,901
	Gradient Boosting	£42,826	£17,710

methods and they all have similar predictive performance with bagging just slightly outperforming the rest. Again, the MAE values highlight this similar predictive performance

as the values are not very different to each other, this time with gradient boosting having the lowest MAE by only £61.

In terms of comparing the tree-based machine learning methods to the methods previously investigated, the RMSE results suggest that the spatial CAR model when KNN=7 performs best in prediction for this particular property price data set because it has the lowest average RMSE, beating bagging, the best machine learning method, by £877. However, the gradient boosting method has the lowest MAE, £49 less than the spatial model. Therefore, it can be said that in the spatial data context classical a-spatial machine learning methods are broadly comparable to standard spatial models. If a best model had to be chosen, one would potentially choose the spatial CAR model when KNN=7 because it has the lowest RMSE meaning that it is best at dealing with outliers in this data set. However, in another context the machine learning methods may be more suitable as it is dependent on the makeup of the data set and how large an impact the outliers make. As seen from Figure 4.7, both machine learning methods and spatial methods seem to do worst at high price extremes as property price predictions are generally lower than the actual property prices, indicating that these methods of prediction are much more accurate at lower property price values than higher ones.

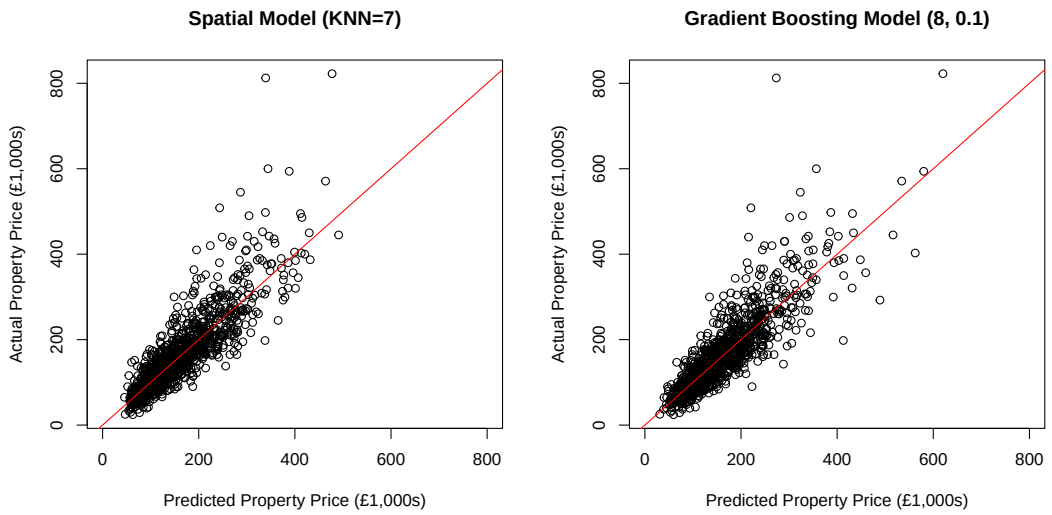

Figure 4.7: Scatterplots of property price predictions on the test set against actual test set property prices using spatial modelling with KNN= 7 (L) and gradient boosting using t_d= 8 and l_r= 0.1 (R).

Overall, both the spatial CAR model and the tree-based machine learning methods have their advantages, the spatial CAR model is good because it picks up on spatial

autocorrelation present in the structure of the data, while the machine learning methods account for non-linear covariate effects. So, logically since both methods produce fairly similar predictions, it may be of interest to attempt to further improve on these predictions by incorporating the strengths of both machine learning methods and spatial methods together. Thus the next chapter will investigate if combining these methods I can achieve a more accurate model for predicting property prices in Scotland.

Chapter 5

Property price prediction by combining spatial and machine learning methods

5.1 Introduction

In Chapters 3 and 4 the predictive abilities of spatial CAR models and machine learning methods were investigated in a spatial areal unit data context and compared with one another through prediction metrics such as root mean square error (RMSE) and median absolute error (MAE). The results of this investigation showed that there is not a substantial difference in the predictive performance of the spatial CAR model and the tree-based machine learning methods compared, namely bagging, random forests and gradient boosting machines. Therefore, in this chapter I will investigate whether combining the strengths of spatial methods and machine learning methods together using geographically weighted random forests will improve the accuracy of property price predictions in Scotland.

5.2 Geographically Weighted Random Forests

An established prediction method which combines spatial information and random forests is the geographically weighted random forest (GWRF) algorithm (Georganos et al., 2021). This algorithm, although very similar to the traditional random forest, adds a degree of complexity by accounting for space. It does this by fitting a separate "local" random forest for each observation in the training set using only it's neighbouring observations, so that the prediction of each observation in the test set depends on a local random forest

"close" to it in space.

5.2.1 Structure

In the property price context a random forest is fitted for each individual Data Zone in the training set using only the training points "close" to the Data Zone in question, making each local random forest unique. A detailed explanation of how a random forest is constructed can be seen in Section 4.4.1. In order to make predictions on a Data Zone in the test set, the closest training Data Zone is chosen and predictions are made using a weighted average of it's associated local random forest and a global random forest constructed from the entire training set. This is contrary to the random forests discussed in Section 4.4 where only a single global random forest was constructed based on all Data Zones in the training set.

5.2.2 Implementation

Like the other machine learning methods assessed in Chapter 4, the local random forest also has various tuning parameters including the number of variables to consider at each split (m_{try}) and the number of trees in the forest (b), both of which are used in the traditional random forest. For this property price data set, I have fixed the values of these specific tuning parameters at $m_{try} = 20$ and $b = 150$ due to the computational complexity of this algorithm (a separate random forest is fitted for each training set data point) and because this specific combination of ($m_{try} = 20$, $b = 150$) was the random forest combination which produced the lowest RMSE in the previous chapter.

Two new tuning parameters are introduced in the local random forest, the first of which is the bandwidth (bw) parameter. The bandwidth controls the number of training points (Data Zones) close to the Data Zone in question which the local random forest is constructed with. This must be a value less than the total number of data points (Data Zones) in the training set, as otherwise if equal, it will produce a global random forest such as that constructed in Section 4.4. So, the bandwidths of 100, 300 and 500 neighbours will be investigated on the property price data set. These values are chosen because even though there are 6,881 Data Zones in total, as the bandwidth increases, the local random forest becomes more computationally demanding and a previous study of this method has shown that in general, using lower values of bw results in a more accurate prediction (Georganos et al., 2021). Furthermore, due to the shortness of time allocated to this research thesis, the lower values of bw, 100, 300 and 500, are a better choice as

they are able to be computed within a shorter time period.

The other tuning parameter introduced in the geographically weighted random forest is the weight parameter (a). The weight parameter controls how much the predictions are determined by the local random forest model and how much by the global random forest model. To see the impact of how predictive performance varies depending on a, this parameter will be evaluated at a= 0.25, 0.5, 0.75 and 1. If a= 0.25 this means that the predictions from the local model hold 25% of the overall weight while the predictions from the global model hold 75%, so this prediction is determined more on the global model than the local model. On the other hand if a=1 then the prediction will be solely determined on the local random forest model with no weighting on the global model whatsoever (Georganos et al., 2021). Note, setting a=0 results in a global random forest which was fitted in Section 4.4.

5.2.3 Choosing the optimal tuning parameter combination

Like the other machine learning methods, before any predictions are made on the test set, the optimal combination of values of the tuning parameters, bw and a, must be chosen by evaluating the same 10-fold cross validation approach used previously. Specifically, the data are randomly split into 5 training-test data splits using the splitting process shown in Figure 2.3.4. Note, the m_{try} and b parameters are fixed at 20 and 150 respectively. Then, the test sets for each of the 5 splits are removed and the local random forest algorithm is applied to the training sets. In order to choose the optimal tuning parameter combination, (bw, a), each of the training splits are further partitioned into 10 smaller sub-splits of roughly equal size. The local random forest is then fit to 9/10 of the training set for each combination of tuning parameters and is then used to predict the 10th, the validation set. This process is then repeated 10 times for each split, with a different validation set used each time, using 10-fold cross validation. This produces 12 different combinations of results for each of the 5 splits with bw=(100, 300, 500) and a= (0.25, 0.5, 0.75, 1), and Tables 5.1 and 5.2 show the values of RMSE and MAE obtained by each combination.

It can be seen from Table 5.1 that the lowest average value of RMSE is £42,717 which is when bw=300 and a=0.5. On the contrary, the highest average value of RMSE is £45,746 when bw=100 and a=1, creating an overall range of £3,029. There is not substantial difference in the RMSE values between the combinations, however when a= 1 the RMSE values are slightly higher. This would suggest that the optimal geographically weighted random forests are those which incorporate both the local and global models as the RMSE values lower when this occurs. Moreover, Table 5.2 shows that the combination

Table 5.1: Table showing the root mean square error (RMSE) of property price for each combination of tuning parameters when building a geographically weighted random forest applied to the 5 data splits.

bw	*a*	Split 1	Split 2	Split 3	Split 4	Split 5	Average
100	0.25	£42,548	£42,539	£43,486	£43,239	£42,911	£42,944
300	0.25	£42,802	£42,628	£43,444	£43,274	£43,035	£43,036
500	0.25	£42,921	£43,045	£43,571	£43,563	£43,330	£43,286
100	0.5	£42,425	£42,207	£43,794	£43,277	£42,674	£42,875
300	0.5	£42,457	£42,300	£43,255	£43,047	£42,528	**£42,717**
500	0.5	£42,696	£42,755	£43,526	£43,400	£43,022	£43,080
100	0.75	£43,260	£42,874	£45,139	£44,335	£43,452	£43,812
300	0.75	£42,737	£42,576	£43,697	£43,484	£42,686	£43,036
500	0.75	£42,966	£42,950	£43,965	£43,736	£43,197	£43,363
100	1	£45,054	£44,551	£47,506	£46,391	£45,225	£45,746
300	1	£43,673	£43,478	£44,792	£44,612	£43,525	£44,016
500	1	£43,750	£43,648	£44,898	£44,585	£43,864	£44,149

Table 5.2: Table showing the median absolute error (MAE) of property price for each combination of tuning parameters when building a geographically weighted random forest applied to the 5 data splits.

bw	*a*	Split 1	Split 2	Split 3	Split 4	Split 5	Average
100	0.25	£17,239	£17,081	£17,030	£17,417	£17,233	£17,200
300	0.25	£17,644	£17,220	£17,373	£17,507	£17,367	£17,422
500	0.25	£17,638	£17,561	£17,300	£17,796	£17,586	£17,576
100	0.5	£17,051	£16,856	£16,985	£16,924	£16,947	**£16,953**
300	0.5	£17,356	£17,065	£17,340	£17,277	£17,145	£17,237
500	0.5	£17,577	£17,287	£17,084	£17,608	£17,142	£17,340
100	0.75	£17,798	£17,483	£17,445	£17,376	£17,449	£17,510
300	0.75	£17,456	£17,276	£17,352	£17,338	£17,339	£17,352
500	0.75	£17,489	£17,383	£17,300	£17,619	£17,337	£17,426
100	1	£18,573	£18,450	£18,455	£18,523	£18,497	£18,499
300	1	£18,021	£17,829	£17,756	£17,846	£17,752	£17,841
500	1	£17,864	£17,617	£17,885	£18,202	£17,478	£17,809

which produces the lowest MAE of £16,953 is when bw=100 and a=0.5. Similar to the RMSE, the least optimal combination is when bw=100 and a=1 as the MAE is the highest at £18,499 when this occurs creating an overall MAE range of £1,546. Again, the highest average MAE values also seem to be when a=1, which would suggest that in order to achieve the most accurate model for prediction, the geographically weighted random forest prediction should not be determined on the local model alone. Therefore, since the best combination of tuning parameters are not the same for RMSE and MAE, we will continue to make predictions on the test set in the following section using both

the bw=300 and a=0.5 combination and the bw=100 and a=0.5 combination.

5.2.4 Test Set Predictions

The geographically weighted random forest is refit to the entire training set, firstly with bw=100 and a=0.5 and then with bw=300 and a=0.5. Predictions can then be made on the test set using the fitted model for each of the 5 training-test data splits. The resulting RMSE and MAE values are evaluated and can be seen in Table 5.3 where the two tuning parameter combinations can be compared with one another in order to find out which is best. To get a general overview of each combination, the average values across the 5 splits are calculated and are shown at the bottom of Table 5.3 .

Table 5.3: Table of root mean square errors (RMSE) and median absolute errors (MAE) of property prices of each of the 5 data splits using the geographically weighted random forest algorithm for 2 different combinations of tuning parameters.

Split	bw	a	RMSE	MAE
1	100	0.5	£44,507	£17,721
	300	0.5	£45,012	£17,517
2	100	0.5	£44,380	£16,372
	300	0.5	£44,621	£17,356
3	100	0.5	£41,257	£17,550
	300	0.5	£40,900	£17,466
4	100	0.5	£39,995	£16,497
	300	0.5	£40,397	£16,503
5	100	0.5	£43,114	£17,048
	300	0.5	£42,319	£17,734
Average	100	0.5	£42,651	£17,038
	300	0.5	£42,650	£17,315

Firstly, it should be noted that of the tuning parameter combinations being evaluated, both have the same value of a, 0.5. This suggests that in terms of making predictions on the property price data set with a geographically weighted random forest, it is best to have equal proportions of the local model and the global model. In general, when looking at the RMSE and MAE values across the 5 splits, there is not a substantial difference between both tuning parameter combinations as respective values are all very similar to each other. The average values displayed at the bottom of Table 5.3 show the RMSE values being almost exactly the same, a difference of only £1. This highlights that in terms of RMSE, there is no bandwidth in particular that is better than the other of the two presented. On the other hand, there is a £277 difference in the average MAE values

with bw=100 having the lower value, although given the scale of the data this difference is small.

5.3 Discussion

In conclusion, by using state-of-the-art geographically weighted random forests to combine the spatial information in the data and random forest models, I found that the GWRF produced very similar results to the other models in terms of both RMSE and MAE. Therefore, since the GWRF did not improve on the predictions made by the spatial CAR models and the a-spatial tree-based machine learning methods on their own and took substantially longer to implement, it suggests these methods work better alone rather than combined in the form of a GWRF.

In general, for all 12 tuning parameter combinations, the geographically weighted random forest had very similar RMSE and MAE values. By comparing the results of these combinations to one another, I found that there was not one single combination that produced the best RMSE and the best MAE. Although, when looking at the tuning parameters individually, the best values of RMSE and MAE were both obtained when a had a value of 0.5, which is an equal mix of the global model and the local model. This suggests that by predicting the property price of a Data Zone by using only the property prices of the Data Zones nearby is not the best approach. Furthermore, the optimal RMSE value was obtained when the bandwidth parameter, bw, was equal to 300, whilst the best MAE value was when bw=100. This shows that lower bandwidths are generally slightly better for predicting property prices than higher bandwidths because bw=500 was not chosen by either metric. So, even though the GWRF performed best when a=0.5 and bw=100 and 300, there was very little difference between all the combinations suggesting that the results are relatively robust to the tuning parameters.

As the geographically weighted random forest is essentially a combination of the spatial information in the data and the traditional random forest model, it is of interest to first compare the GWRF results to the other models compared in this thesis. From Table 5.4, the overall average RMSE of the GWRF is £42,651 while the spatial CAR model where knn = 7 is £1,000 lower at £41,653. On the other hand, the GWRF has a lower MAE by around £700 which shows that one method is not definitively better than the other. Similarly, when comparing the GWRF to the random forest, the average RMSE for the random forest is £42,570, approximately £80 less than the GWRF. Although this is a very small difference in the grand scheme of things, it shows that in the context of

property prices in Scotland, using the global random forest model alone produces slightly better results with regards to RMSE than using an equal combination of the global and local models. On the contrary, the MAE of the GWRF is £863 less than that of the random forest, which highlights that the GWRF performs better than both spatial and machine learning methods in terms of MAE and slightly worse in terms of RMSE. Since the RMSE is a prediction metric which penalizes more heavily for extreme values, this leads to the conclusion that despite all 3 of these models being broadly comparable to one another, the GWRF is better at prediction when there are very few extreme values while the random forest and spatial CAR model are better at predicting extreme property price values than the GWRF.

Overall, when evaluating Table 5.4, it can be seen that excluding the linear model and the decision tree, the tree-based machine learning methods, spatial CAR model and the geographically weighted random forest all have very similar predictive performance across all 5 splits as the RMSE and MAE values are all within around £2,000 and £1,000 of one another respectively. On average, the best value of RMSE, £41,653, is obtained using the spatial CAR method, which suggests that when presented with areal unit data like this particular data set with extreme price values, that accounting for space using spatial random effects is more important for making accurate predictions than flexible covariate response relationships delivered by the random forests. However, the spatial CAR model was outperformed in terms of MAE by the geographically weighted random forest model which has the lowest overall average MAE of £17,038. Therefore, this indicates that there is not a clear answer as to which method is best for prediction as a single model does not outperform the rest in terms of both RMSE and MAE. Due to there being very little difference in the results between the methods across the 5 splits, it indicates that these methods excluding the linear model and the decision tree, are all broadly comparable to one another. Thus each of the methods have different strengths, and the most suitable method for prediction is likely to vary depending on the data set in question. A data set with lots of outliers will benefit more by using the spatial CAR method for prediction, while on the other hand a data set with fewer outliers is likely to have more accurate predictions using the geographically weighted random forest.

In the future, if I have the opportunity to investigate combining spatial and machine learning methods without being limited by time, it would be interesting to find out if there are other ways which will improve the modelling of property price predictions in Scotland. One method of improvement is to look at the tuning parameters in more detail. When I was choosing the best tuning parameter combination for the geographically

weighted random forest, I had set the values of a, the proportion determined by the local model, to increase in increments of 0.25 from 0.25 to 1. The value which produced the optimal RMSE and MAE values were when a=0.5, an even combination of the global model and the local model. However, when comparing to the traditional random forest in Chapter 4.4 when a=0, the GWRF had a poorer RMSE by only £80. So, since this difference is very small, if a was increased in smaller increments such as by 0.01, there may be a value between 0.25 and 0.75 where the GWRF has a lower RMSE than the random forest and thus is a more accurate model overall.

Furthermore, by investigating whether there are ways of combining spatial and machine learning methods other than the geographically weighted random forest could be beneficial for building a more accurate model for prediction. The GWRF was chosen as the method which combines space and machine learning because it is a popular algorithm that has software to allow implementation. But, of all the well performing tree-based machine learning methods, the random forest performed the worst in terms of both RMSE and MAE. There is a chance these values could be improved upon if bagging or gradient boosting was combined with spatial methods instead of the random forest, so if there was more time allocated to the thesis these methods could be researched and evaluated.

Table 5.4: Table showing the RMSE and the MAE of each of the 5 data splits using a decision tree, linear model, spatial model where $k = 7$, bagging, random forest where m_{try}=20 and b=150 and gradient boosting.

Split	Model	RMSE	MAE
1	Linear	£47,995	£20,038
	Spatial($k = 7$)	£44,068	£18,386
	Bagging	£44,745	£18,037
	Random Forest	£44,676	£18,036
	Gradient Boosting	£44,381	£17,815
	GWRF(bw=100, $a = 0.5$)	£44,507	£17,721
	GWRF(bw=300, $a = 0.5$)	£45,012	£17,517
2	Linear	£48,973	£19,902
	Spatial($k = 7$)	£43,775	£18,111
	Bagging	£44,805	£17,487
	Random Forest	£44,430	£17,815
	Gradient Boosting	£45,686	£17,389
	GWRF(bw=100, $a = 0.5$)	£44,380	£16,372
	GWRF(bw=300, $a = 0.5$)	£44,621	£17,356
3	Linear	£45,631	£20,910
	Spatial($k = 7$)	£40,122	£17,431
	Bagging	£41,175	£18,793
	Random Forest	£41,408	£18,890
	Gradient Boosting	£41,225	£17,888
	GWRF(bw=100, $a = 0.5$)	£41,257	£17,550
	GWRF(bw=300, $a = 0.5$)	£40,900	£17,466
4	Linear	£43,524	£18,704
	Spatial($k = 7$)	£39,439	£17,624
	Bagging	£39,795	£17,049
	Random Forest	£40,173	£16,946
	Gradient Boosting	£39,688	£16,730
	GWRF(bw=100, $a = 0.5$)	£39,995	£16,497
	GWRF(bw=300, $a = 0.5$)	£40,397	£16,503
5	Linear	£45,875	£18,984
	Spatial($k = 7$)	£40,861	£17,244
	Bagging	£42,131	£17,490
	Random Forest	£42,163	£17,817
	Gradient Boosting	£43,150	£18,726
	GWRF(bw=100, $a = 0.5$)	£43,114	£17,048
	GWRF(bw=300, $a = 0.5$)	£42,319	£17,734
Average	Linear	£46,400	£19,708
	Spatial($k = 7$)	£41,653	£17,759
	Bagging	£42,530	£17,771
	Random Forest	£42,570	£17,901
	Gradient Boosting	£42,826	£17,710
	GWRF(bw=100, $a = 0.5$)	£42,651	£17,038
	GWRF(bw=300, $a = 0.5$)	£42,650	£17,315

Chapter 6

Conclusion

6.1 Discussion

The main aim of this thesis, as set out in Section 1.1, was to investigate different methods for predicting property prices in Scotland and compare how accurate those predictions were. This was done by evaluating various different methods such as the linear model (Section 2.4), the spatial conditional autoregressive (CAR) model (Chapter 3), tree-based machine learning methods (Chapter 4) and finally a spatially adjusted machine learning algorithm, namely the geographically weighted random forest (GWRF) (Chapter 5), using a range of prediction metrics such as root mean square error (RMSE) and median absolute error (MAE). In order to prevent the results being biased and achieve accuracy in predictions, each of the methods were assessed via 5 training and test data splits, which were then further split using the 10-fold cross validation technique to tune the parameters and find the optimal tuning parameter combinations.

Firstly, it was important to gain an insight into the characteristics of the data set and identify any key patterns which emerged. These findings were discussed in Chapter 2. The average median property prices in Scotland's Data Zones in 2018 ranged from as low as £19,500 to as high as £878,000, with the majority priced between £50,000 and £250,000. When looking at specific covariates which may have an effect on property price, mean number of rooms and council tax are two covariates which tend to influence property price the most. However, since there was not one single covariate that can be used to determine the price of a property, this led to the conclusion that there are multiple factors that contribute to the price of a property. The linear model was used as a simple prediction model in order to see the impact of covariates and was also set as the original baseline model for the more complex models to beat.

The predominant method for modelling areal unit data is spatial CAR models, hence this was the natural first choice model to use in Chapter 3 to attempt to improve on the predictions made by the linear model. Spatial CAR models use the same linear covariate model as the linear model, but by accounting for the spatial correlation between observations after covariate adjustment, the accuracy of the predictions is improved. This essentially means that one would assume that the Data Zones with missing values will be similar to those in neighbouring Data Zones with available values, once the covariate effects have been accounted for. The strength of the spatial correlation present in the residuals of the linear model was measured using Moran's I test (Moran, 1950), and significant spatial correlation was found. Then the spatial model was fitted to the data with 10 different values of k, the number of nearest neighbours for constructing the spatial correlation structure via the neighbourhood matrix, through 10-fold cross validation on the training validation set. It was found in Tables 3.2 and 3.3 that the best values of k for RMSE and MAE were k=7 and k=3 respectively. Then when assessing these values of k on the entire training set to predict the test set in Table 3.4, k=7 produced the optimal RMSE and MAE values. Overall, this method offered a dramatic improvement over the linear model, improving RMSE by around £5,000 (approximately 10%) and MAE by around £2,000 (approximately 10%).

After looking at the classical spatial statistical methods, the next step in this thesis in Chapter 4 was to study different machine learning methods, as they are currently popular for prediction problems and generally perform well. I focused specifically on tree-based methods because they are the one of the most common classes of machine learning models used today. So, to begin with a simple decision tree model was constructed due to its simplicity and the fact that it is the basis for all of the other tree-based machine learning methods. As expected, Table 4.1 showed that this model performed poorly because it is known to be a weak leaner (Boehmke and Greenwell, 2019). After studying the single decision tree, I then moved on to investigate more complex ensemble machine learning methods such as Bagging (Section 4.3), Random Forest (Section 4.4) and Gradient Boosting (Section 4.5). These methods all had very similar results as they follow the same general principles, with bagging being a special case of the random forest model with m_{try} being equal to the full set of covariates. By studying the different tuning parameter combinations for each method using 10-fold cross validation on the training-validation set, the optimal tuning parameter combinations were 150 trees for Bagging (Table 4.2), 150 trees and m_{try}=20 and 100 trees and m_{try}=13 for Random Forest (Tables 4.3 and 4.4), and 150 trees, learning rate=0.1 and tree depth=8 for Gradient Boosting (Tables 4.7 and

4.8). In general, the tuning parameters did not have a very large impact on the results suggesting that most methods are fairly robust to tuning parameter selection. When comparing the three tree-based methods with one another, despite being very similar, Bagging proved to have the lowest RMSE value whilst Gradient Boosting had the lowest value of MAE. Overall, Table 4.9 shows that there was not a lot of difference between the tree-based machine learning methods and the spatial CAR models, suggesting that in general neither is clearly preferable. If one had to be selected, spatial CAR models are slightly better in RMSE by approximately £877, whilst the tree-based machine learning method of Gradient Boosting is slightly better in MAE by approximately £49.

Since spatial methods and machine learning methods produced very similar results, it was of interest in Chapter 5 to study a recently proposed method, the fusion of these two approaches, the GWRF (Georganos et al., 2021). The GWRF fits a separate local random forest for each Data Zone using only it's spatially neighbouring Data Zones. Then predictions of missing values depend on a proportion of the local random forest closest to it in space and the global random forest, the latter accounting for all Data Zones. By carrying out the same 10 fold-cross validation technique on the training-validation data splits as used for the spatial models and the machine learning models, the optimal tuning parameter combinations were obtained from the results of Tables 5.1 and 5.2. The optimal GWRF models had a=0.5, an equal proportion of the global and local models, and lower bandwidths of 100 and 300 nearest Data Zones. Generally speaking, when compared to spatial and machine learning methods in Table 5.4, the GWRF was better at achieving a lower MAE, an improvement of nearly £700 on Gradient Boosting when bw=100, however it is not as good regarding RMSE compared to the spatial CAR model and machine learning methods apart from Gradient Boosting.

In conclusion, all of the complex (not linear model or single decision tree) methods of prediction studied are broadly comparable to each other as they produced very similar results. As seen in Table 5.4, these methods all had values of RMSE around £42,500 and MAE values of around £17,600, a clear improvement on the linear model. In this data set spatial autocorrelation is present and is important, as shown by the spatial CAR model substantially outperforming the linear model. Moreover, linear and non-linear effects are also important as the tree-based machine learning models are much better at accurately predicting the property prices than the linear model. The combining of spatial methods and machine learning methods in the form of the geographically weighted random forest does not perform better than the spatial CAR model or the random forest overall, however like the other methods it is much better at prediction than the linear model. Therefore,

the conclusion can be drawn that there is not one method that is the "best" at predicting the areal unit property price data analysed in this thesis, as it depends on the prediction metric (RMSE or MAE) used.

6.2 Future Work

This research provided an interesting insight into different methods of prediction for areal unit data in the context of Scottish property prices. It concluded that, of the methods studied, there is no particular "best" method of prediction which substantially outperformed the others. However, if there was more time allocated to this research, there are a few things that could have been done to achieve a more accurate set of results.

Firstly, it is important to keep in mind that the conclusion that I reached is only based on the results of one property price data set. If a similar study was carried out on another property price data set of perhaps another year or another study region, a different conclusion may occur. Therefore, in the future, if I was allocated more time to complete this research project, I would explore other property price data sets and simulate predictions using the same methods as in this thesis. By doing this, I would be able to compare the results across different data sets and reach a more comprehensive conclusion over whether there is a method which is best at predicting property prices in general.

In addition, another improvement which could be made is the prediction methods chosen to be studied. When investigating property price predictions by combining spatial methods and machine learning methods, I only considered one method, the geographically weighted random forest (Georganos et al., 2021). As mentioned previously, this method was selected because it is an established method and had software in R to allow ease of implementation. However, as seen from Section 4.4, of the tree-based machine learning models, the random forest did not have the optimal average RMSE or MAE value. Therefore, this would encourage me to explore the respective combinations of bagging and gradient boosting with the spatial information in the data in the future, as this could improve the accuracy of predictions and outperform the current methods I have investigated in this thesis.

Finally, as far as prediction goes, there are a plethora of different methods that can be used to make predictions on data. In Chapter 4, where property price predictions were made using classical machine learning methods, only tree-based methods were considered and studied. There are other non-tree-based methods which could be investigated on this

property price data set if there was no time constraint, an example being neural networks (Boehmke and Greenwell, 2019). There may be a chance that there are machine learning methods other than those studied in this thesis that have lower RMSE and MAE values and are able to more accurately predict property prices in Scotland.

References

Bensic, M., Sarlija, N., and Zekic-Susac, M. (2005). Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. *Intelligent Systems in Accounting, Finance & Management: International Journal*, 13(3):133–150. 38

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43:1–20. 30

Bivand, R. S., Pebesma, E. J., Gomez-Rubio, V., and Pebesma, E. J. (2008). *Applied spatial data analysis with R*, volume 747248717. Springer. 27

Boehmke, B. and Greenwell, B. M. (2019). *Hands-on machine learning with R*. CRC press. viii, 2, 18, 19, 39, 40, 47, 48, 52, 55, 56, 57, 74, 77

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24:123–140. 55, 56

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32. 46, 55

Breiman, L. (2017). *Classification and regression trees*. Routledge. 40

Brown, H. (2022). Scotland's 'poshest' and 'most desirable' villages, according to recent report. https://www.scotsman.com/lifestyle/homes-and-gardens/scotlands-most-desirable-villages-according-to-recent-report-3530080. Accessed: 2022-10-21. 6

Carmona, P., Climent, F., and Momparler, A. (2019). Predicting failure in the us banking sector: An extreme gradient boosting approach. *International Review of Economics & Finance*, 61:304–323. 39

Cruz, J. A. and Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2:117693510600200030. 38

Dormann, C. F., McPherson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., Davies, R. G., Hirzel, A., Jetz, W., Kissling, D. W., et al. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30(5):609–628. 2

Gaikwad, D. and Thool, R. C. (2015). Intrusion detection system using bagging ensemble method of machine learning. In *2015 international conference on computing communication control and automation*, pages 291–295. IEEE. 38

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC. 32

Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., and Kalogirou, S. (2021). Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International*, 36(2):121–136. 2, 4, 38, 64, 65, 66, 75, 76

Gómez-Rubio, V. (2021a). Bayesian inference with INLA: Chapter 1 introduction to bayesian inference. https://becarioprecario.bitbucket.io/inla-gitbook/ch-mixed.html. Accessed: 2023-01-13. 33

Gómez-Rubio, V. (2021b). Bayesian inference with INLA: Chapter 2 the integrated nested laplace approximation. https://becarioprecario.bitbucket.io/inla-gitbook/ch-mixed.html. Accessed: 2023-01-13. 33

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer. 48

Knoll, L., Breuer, L., and Bach, M. (2019). Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. *Science of the total environment*, 668:1317–1327. 39

Lee, D. (2013). CARBayes: an R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, 55(13):1–24. 29

Lee, D. (2023). Modelling areal data. Spatial Statistics 4H and 5M. Lecture Slides, University of Glasgow, Glasgow, UK. 30

Lee, D. and Mitchell, R. (2013). Locally adaptive spatial smoothing using conditional auto-regressive models. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, pages 593–608. 2, 30

Leroux, B., Lei, X., and Brewslow, N. (1999). Estimation of disease rates in small areas: a new mixed model for spatial dependence. *Statistical models in epidemiology, the environmental and clinical trials*, pages 135–178. 31

Marshall, C. (2013). Scotland's best performing schools revealed. https://www.scotsman.com/education/scotlands-best-performing-schools-revealed-1549864. Accessed: 2022-10-21. 6

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23. 26, 28, 74

Public Health England (2018). England's poorest areas are fast food hotspots. https://www.gov.uk/government/news/englands-poorest-areas-are-fast-food-hotspots. Accessed: 2022-10-15. 17

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 5

R INLA Project (2020). What is INLA? https://www.r-inla.org/what-is-inla. Accessed: 2023-01-11. 2, 33

Redhead, H. (2015). Living near a waitrose 'increases the value of your property by 12 percent'. https://metro.co.uk/2015/04/03/living-near-a-waitrose-increases-the-value-of-your-property-by-12-percent-5134605/. Accessed: 2022-10-15. 17

Rightmove (2023). Find estate agents and letting agents in the uk. https://www.rightmove.co.uk/estate-agents.html. Accessed: 2023-09-17. 1

Robertson, C. and Gray, A. (2021). Data splitting. MM404: Statistical Modelling and Analysis. Lecture Slides, University of Strathclyde, Glasgow, UK. 18

Rue, H. and Held, L. (2005). Gaussian Markov Random Fields: Theory and Applications (1st ed.). chapman and hall/crc. https://doi.org/10.1201/9780203492024. Accessed: 2023-02-23. 30

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392. 2, 32, 33

Scottish Government (2004). Scottish neighbourhood statistics data zones background information. https://www.gov.scot/publications/scottish-neighbourhood-statistics-data-zones-background-information/pages/. Accessed: 2022-10-18. 5

Scottish Government (2012). Scottish local government financial statistics 2010-11. https://www.gov.scot/publications/scottish-local-government-financial-statistics-2010-11/. Accessed: 2022-10-11. 6, 11

Scottish Government (2017). Local authorities: factsheet. https://www.gov.scot/publications/local-authorities-factsheet/. Accessed: 2022-10-19. 5

Scottish Government (2020a). Scottish index of multiple deprivation 2020. https://www.gov.scot/collections/scottish-index-of-multiple-deprivation-2020/?utm_source=redirect&utm_medium=shorturl&utm_campaign=simd. Accessed: 2022-10-21. 11, 15

Scottish Government (2020b). Simd 2020 technical notes. https://www.gov.scot/binaries/content/documents/govscot/publications/statistics/2020/09/simd-2020-technical-notes. Accessed: 2022-10-28. 16, 17

Scottish Government (2021). House prices. https://statistics.gov.scot/. Accessed: 2022-09-26. 2, 6

Shaw, V. (2017). Waitrose effect can 'boost house prices by thousands of pounds'. https://www.independent.co.uk/property/house-prices-latest-waitrose-effect-sainsburys-marks-and-spencer-uk-property-a7760926.html. Accessed: 2022-10-15. 17

Shekhar, S. and Xiong, H. (2007). *Encyclopedia of GIS: Correlation and Autocorrelation*. Springer Science & Business Media. 28

Su, Y., Tian, X., Gao, R., Guo, W., Chen, C., Chen, C., Jia, D., Li, H., and Lv, X. (2022). Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis. *Computers in Biology and Medicine*, 145:105409. 38

Tehrany, M. S., Jones, S., and Shabani, F. (2019). Identifying the essential flood conditioning factors for flood prone area mapping using machine learning techniques. *Catena*, 175:174–192. 39