A Statistical Analysis of Housing Prices In Bristol

Group 10: Baggator

Andre Dale, Benjy Bream, Kate Holloway, Ted Mellow, Kit Crowley

July 16, 2025

Mathematical and Data Modelling 3

School of Engineering Mathematics and Technology

University of Bristol

Declaration of AI use

AI (including OpenAI's ChatGPT) has been used for assisting in research for different methods of data analytics that are industry / academic standard, such as suggestions for the ARIMA/ARIMAX model. AI was also used for help in resolving code issues where necessary, as well as for suggesting packages in python that were purpose built for specific statistical investigation use cases, such as for the help of data normalisation with Python's Sklearn. AI has also helped with LaTeX code for formatting components such as tables and figures.

1 Introduction

In the UK, Bristol's house prices are rising faster than the national average, leading to Bristol being the second most expensive area in the UK to buy a house in [1]. This does not however translate into a uniformly prosperous city, with 46% of children in Lawrence Hill living in poverty [2]. Tackling information asymmetry between buyers and sellers in the housing market can help buyers who may lack the time or resources to do their own research to make more informed decisions when it comes to purchasing housing.

This paper's aims are two-fold: To analyse the underlying dynamics of Bristol's housing market, and feed these insights into a model capable of predicting future trends in the housing market.

A spatial and temporal analysis of Bristol's individual ward prices is undertaken with the aim of understanding if there are dominant areas of Bristol that lead the trends and cause a spill-over effect to neighbouring areas. Wards that appear to have notable influence on the market are further investigated, as it may be possible to only analyse a few key areas in order to determine how the entire market will change over time [3].

There are two main types of applicable market forecasting models: traditional, including macroeconomic models, and advanced models which include Auto-Regression Integrated Moving Average (ARIMA) models [4]. Previous research on macroeconomic models for predicting house prices has explored the relationship between income and house prices [5], as well as influencing the Office for Budget Responsibility's (OBR) integration of various macroeconomic factors into its estimation model [6].

The OBR's model benefits from a plethora of available data, and is based on a host of macroeconomic factors that vary from disposable income to the number of households. The model presented is effective at forecasting long-term, but struggled to accurately predict the short term, lagging about a year behind the actual house prices. Long term predictions may be useful for a government or real estate company, but for an individual agent aiming to understand a local market in the short term, this model may be unhelpful. Error correction models look at short-term forecasting based on the income-to-house price ratio [5]. Though this model is effective at predicting the housing market in the short term, the paper states it is hindered by the amount of macroeconomic data available. Due to the shortfalls of both aforementioned models, more advanced models are considered that may be more data-efficient and accurate in the short term.

Studies comparing machine learning methods, including ARIMAX models against neural networks have also shown ARIMAX's ability to accurately predict future trends [7]. Despite being tested on wheat prices, which are more volatile than the housing market, the ARIMA models resulted in the highest accuracy, despite their known limitations to predict sudden market shifts [8]. It has been proposed that a hybrid ARIMAX-neural network model would be even more accurate [9].

ARIMA models have previously been applied to the Manchester housing market [10], with higher success for detached houses but much lower for terraced houses. The model is held back by a lack of data preventing the model from being trained exclusively on post 2008 financial crash data. Further results have been obtained in the Beijing property market [11], reaching success in some districts and less so in others. In that case the model is held back due to the uni-variate nature of ARIMA, and could have been improved with a multivariate model including factors such as location.

Therefore, this paper conducts a detailed analysis of multivariate ARIMA models applied to the housing market, incorporating factors such as different location price data and utilising comprehensive data sources. Furthermore, given the limitations of ARIMA models in forecasting market crashes, a more detailed analysis of the model's performance is conducted in the context of the 2008 financial crisis [12].

2 Exploratory Data Analysis

Here 'house' will refer to any dwelling unit, including all house types and flats. The data used in this report comes from the Office for National Statistics (ONS) [1] and HM Land Registry, a non-ministerial UK government organisation [13]. Using government officiated data ensured the data obtained is reliable, complete, and well formatted.

City wide data for Bristol is extracted from the UK House Price Index (HPI) [13]. The dataset includes median house prices broken down by house type, as well as whether the property is a new or old build. It also includes monthly and yearly percent changes, and sale counts broken down by the aforementioned categories. Data is obtained on a monthly basis from 1995 to 2023.

Median house prices for Bristol wards are extracted from the HPSSA 37 dataset [14]. This data is not broken down by house type and comes in 3-month intervals over the same time period as the HPI.

For any spatial analysis, the central points of the wards and data on their geographies were provided by the ONS ordnance survey [15].

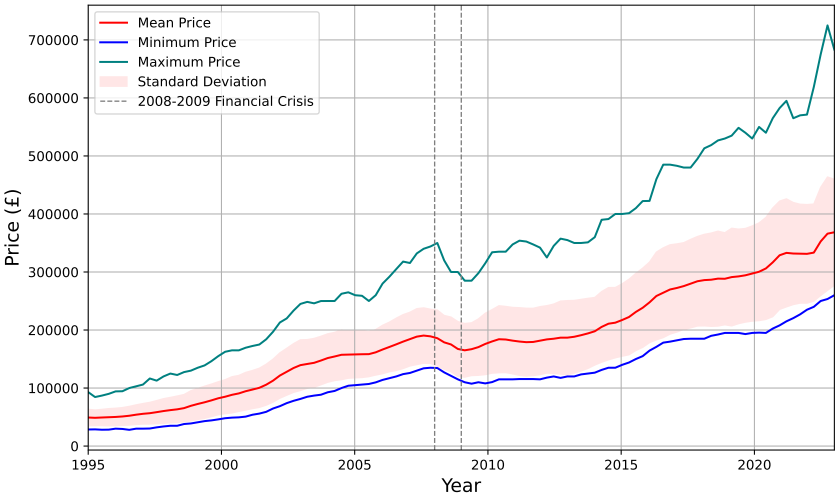

Figure 1: Price of housing in Bristol wards.

Initial analysis of the data reveals that since the 2008 financial crisis, prices have grown by £14,000 a year on average in Bristol, with the gap between the most expensive and cheapest wards widening from £65,000 in 1995 to £420,000 in 2023, highlighted in Figure 1. The highest priced wards have been much more volatile, increasing by 35% in 2022 before dropping by 5.7% in the first 3 months of 2023, while the median price increased by only 9.5% over the same period and remained consistent at the start of 2023. Despite this the lowest-priced wards have appeared immune to the fluctuations, with prices increasing shallow in comparison at a rate of 26% a year since 2021.

2.1 Spatial Investigation

An investigation into how the local geography affects house prices is performed by looking for trends between ward prices [1] and proximity to a particular ward. The distances from wards and the relationship that had on the median housing prices is investigated. Ward centroids were obtained from the ONS ordinance survey [15] and the Euclidean distance between wards is then calculated, normalised between 0 and 1, and inverted so higher values indicate closer wards. As different trends were identified in pre and post 2009 results, only data post 2009 is included in the report to ensure the results could be applied to current and future investigations.

The Pearson Correlation coefficient, R [16], is used as a measure of this price-distance relationship. For a target ward and their ward-pair combinations, all R values were calculated, and stored in a vector, which itself is correlated for R^2 with a vector of the target ward's distances from the other wards. The indexing of wards is consistent for both vectors, resulting in an output value that is defined as the target ward's Distance Correlation coefficient. This is performed for all wards, with their Distance Correlation coefficient's plotted in Figure 2.

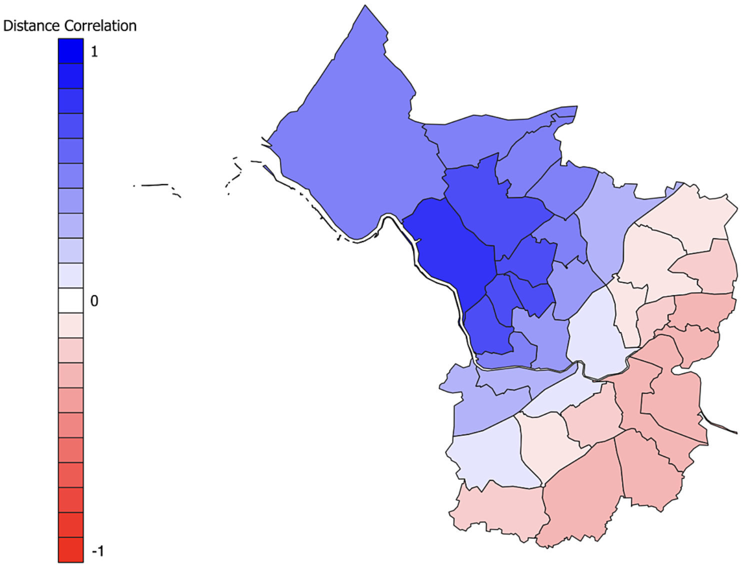

Figure 2: Map of the Distance Correlation coefficients of Bristol wards. Positive values for a ward imply a house price increases the closer to the ward.

Stoke Bishop stands out in 2, represented by the darkest blue region in the figure, exhibiting the highest Distance Correlation coefficient, implying that increasing distance from this ward corresponds to a decrease in average house price. This can also be seen from the wave like decrease in Distance Correlation coefficients further away from Stoke Bishop. Because of this influence, Stoke Bishop will be used as the primary investigatory ward.

2.2 Rolling Window Correlation

An investigation into how closely each ward's house prices correlate can give insights into how the market fluctuates with time [16]. By converting the HPSSA 37 data into a percentage change between the 3 month intervals, a normalised set of data is obtained.

A rolling correlation method is implemented to see how the correlation between wards changes over time. This method takes a window which is a subset of the time-series data from two wards and a pre-determined window-size, which is the number of consecutive points from each time series used for comparison, e.g. a year is 4 data points. The Pearson correlation coefficient, R is calculated using the data from that window upon both wards [16]. After this calculation, the window slides down the dataset by one row, and finds the correlation coefficient in the next subset of data for both wards. This process continues until the rolling window reaches the end of the dataset. Rolling correlation is performed using window sizes of 1, 2, 3, 5, and 10 years. This gives a range of window sizes that is sensitive enough to capture both short and long-term trends between wards.

Values for Figure 3 were obtained by calculating R for every ward pair permutation with that window size, and finding the mean of all of these ward-pairs to represent typical values for the entire City of Bristol when measured by that rolling window size. This is then repeated for every window size. Figure 3 illustrates that a 1-year rolling window produces the highest correlation coefficient, suggesting that correlations can be best understood between wards when measured or forecasted for one year.

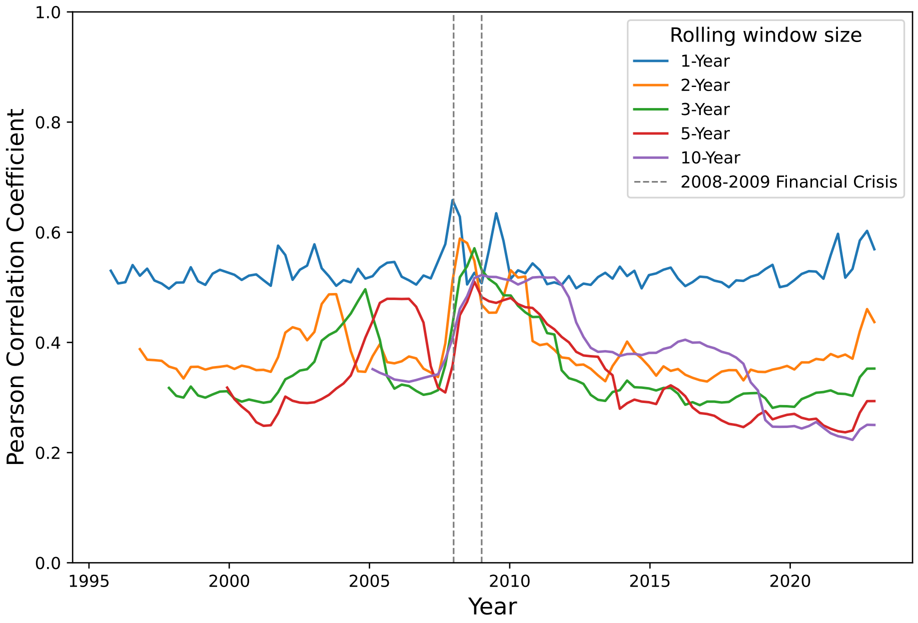

Figure 3: Mean Pearson correlation coefficient of Bristol wards for various rolling window widths.

All ward correlation comparisons for Stoke Bishop are listed in order from highest R^2 to lowest in Figure 4. The figure shows that there are many wards that have a high correlation coefficient, and thus a predictor model should benefit from including some of these, but after a certain amount the marginal gains will drop off and may even negatively impact the forecast.

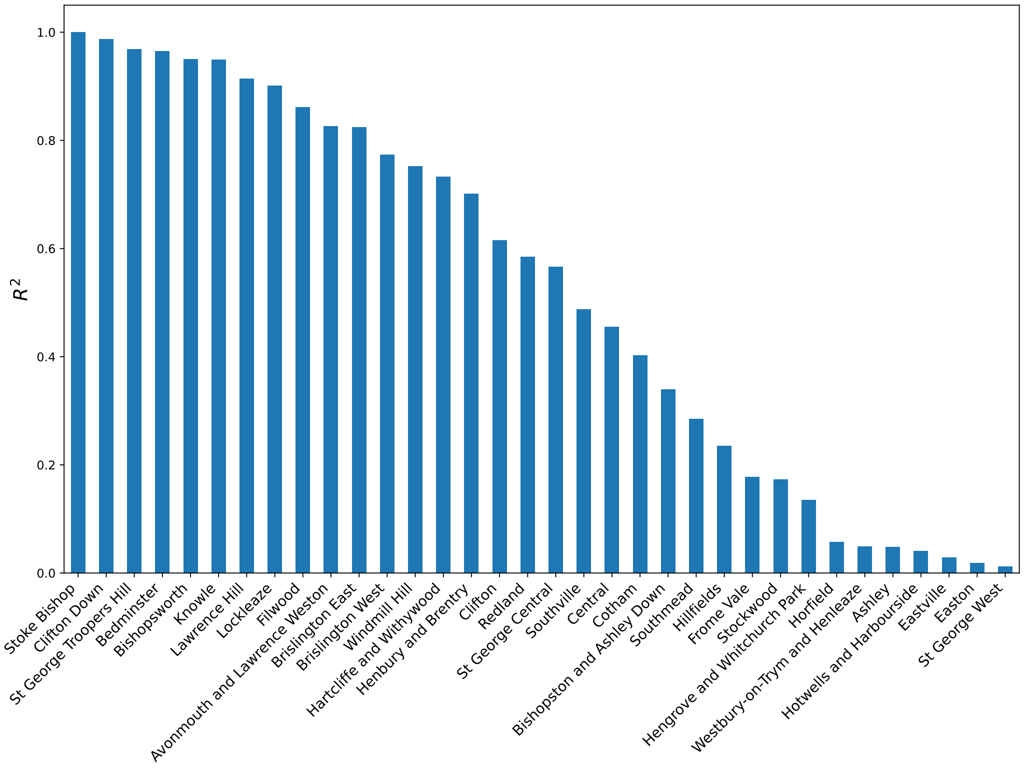

Figure 4: Stoke Bishop R^2 with Bristol wards ordered by highest median price to lowest median price.

For a given ward, the wards exhibiting the highest R^2 correlation with it are denoted its Feature Wards. These Feature Wards, ranked in order of correlation strength serve as feature inputs to the ARIMAX model in Section 3.

2.3 Principle Component Analysis

Principal component analysis (PCA) is performed on the Bristol data extracted from the HPI dataset [13]. The PCA is performed to determine the most significant features that dictate the trends in the Bristol housing market, and to inform what features should be included in the ARIMAX prediction models [17]. The data is NaN imputed where necessary [18], and then normalised between values from 0 to 1 before performing the PCA[1].

Table 1: Top 5 significant loading contribution features for the top 5 Principal Components

PC1 - 46.63%	**PC2 - 22.39%**	**PC3 - 9.65%**	**PC4 - 8.69%**	**PC5 - 4.17%**
SemiDetachedPrice	1m%Change	Cash1m%Change	Cash12m%Change	NewSalesVolume
DetachedPrice	12m%Change	FTB1m%Change	FTB12m%Change	SalesVolume
TerracedPrice	SemiDetached1m%Change	Mortgage1m%Change	Mortgage12m%Change	OldSalesVolume
FlatPrice	Old1m%Change	FOO1m%Change	FOO12m%Change	MortgageSalesVolume
OldPrice	SemiDetached12m%Change	New12m%Change	CashSalesVolume	CashSalesVolume

The PCA revealed that the top 5 principle components explain 91.54% of the variance in the data, shown in Figure 5, indicating there are a few key features driving the overall house price trends. The highest loading contributors to these principal components can be found in table 1, which are the features that had the strongest linear correlation with the principal component.

[1] the methods of which are detailed in Appendix section A.1

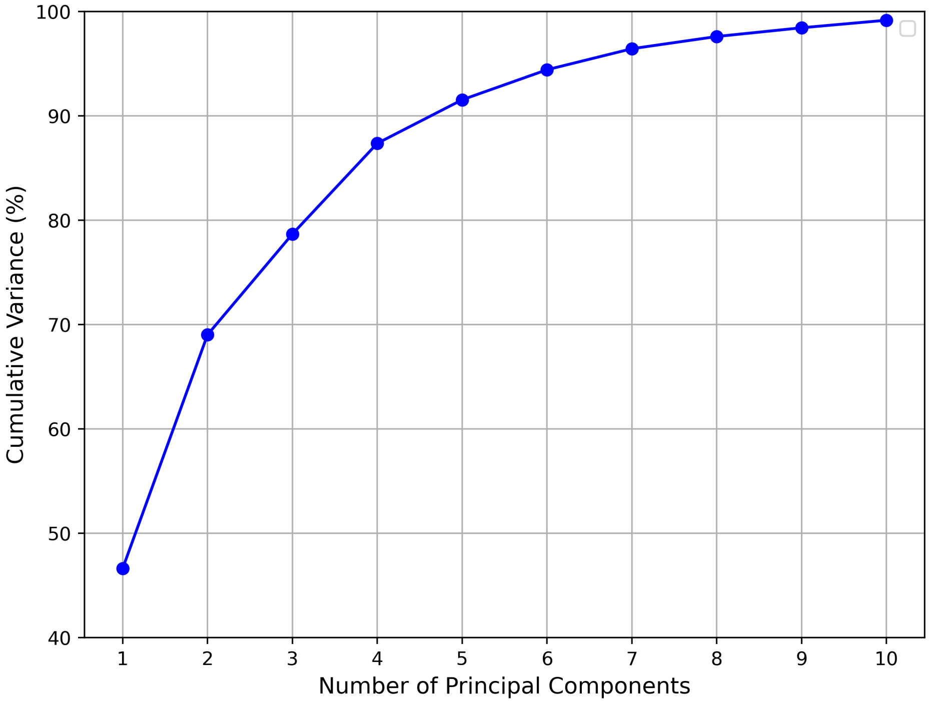

Figure 5: Variance (cumulative) explained by the top 10 principal components

Principal component 1 (PC1) accounted for 46.63% of the overall variance. The loading contributions were all within ±0.0002 of one another, which suggests that price changes across different property types all follow a common market trend, due to their equal strength correlation. This implies no particular type of housing disproportionally affects the overall pricing trends.

The top contributors to PC2, which accounted for 22.39% of the variance, are all short term and long term price changes for the different property types, suggesting PC2 capturing the rate of housing market changes throughout Bristol. Similar to PC1, their loading contributions are all relatively equal, which supports the claim that no particular type of housing has a significant sway on the market.

PC3, PC4 and PC5 each explain a small amount of the variance, so it is unreliable to make conclusive statements about them, however for forecasting purposes with predictive models they may prove beneficial for improving accuracy and reliability.

It is often claimed, incorrectly, that volume of houses being sold has a large impact on the price of housing [19]. PC5 supports this contradiction, with the top 5 loading contribution features being sales volume but only explaining 4.17% of the variance in prices.

3 Predictive Model Methods

This section of the paper will forecast the time series data of both the HPSSA 37 [14] and HPI [1] datasets. HPSSA 37 is used to predict all wards house prices, though the results of Stoke

Bishop will be explored in further detail due to the results in section 2.1 suggesting it is a 'leader ward'.

Autoregressive Integrated Moving Average (ARIMA) models are the principal forecasting models used in this paper. Suitable for non-stationary time-series data whilst accurately capturing historical trends [20], ARIMA models and variants have been used for housing market investigations with proven success[10]. Although ARIMA is designed for univariate data, not considering relationships between multiple variables, it can easily be adapted for multi-variate data (ARIMAX) that can investigate multiple variables and longer term behaviours. This section of the paper will explore how incorporating multi-variate data into ARIMA affects its forecasting capabilities, as well as methods of improving these forecasts.

The mathematical formulation of an ARIMAX model is described below,

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \sum_{k=1}^{m} \beta_k X_{t,k} + \epsilon_t \tag{1}$$

Where Y_t is the predicted time series value at time t, c is a constant term, ϕ_i is a measure of how strong past values effects are on the next prediction, Y_{t-i} are the values of the time series at time $t - i$. θ_j is a measure of how strong the past errors effects are on the next prediction, ϵ_{t-j} are the magnitude of the errors at time $t - j$. β_k are the weighting coefficients of the different exogenous variables which is left constant, and m is the number of exogenous features, with $X_{t,k}$ representing the value of the kth-exogenous variable at time t. ϵ_t is the error term at time t, with p and q being tunable hyperparameters.

4 Predictive Model Results

4.1 Ward Forecasting

Both the ARIMA and ARIMAX models will be evaluated on a 70/30 train-test split on the entire HPSSA 37 dataset to attempt to accurately predict the housing market change for specific wards. Each model will then be retrained on data pre 2008 to forecast the 2008 crash in order to evaluate their forecasting robustness in the face of sudden change in the housing market.

The Feature Wards are used for the ARIMAX multi-variate exogenous variables. The model is trained and tested with the next strongest Feature Ward added each time, recording their respective R^2 value for that models prediction against the test data. The number of exogenous features, m (1), included in the final ARIMAX predictive model will be the combination of Feature Wards that outputs the highest R^2 value with the test set data. The success metric R^2 is chosen because of its inherent normalisation for comparison of accuracy between model iterations, and to be able to be compared with the results from similar papers [10].

The ARIMAX performs well when predicting the testing set for the majority of wards, with only 3 wards having an $R^2 < 0.75$. The 19 cheapest wards, bar Central, have an $R^2 \geq 0.9$. Some of the more expensive wards have high R^2 values, though many are much lower with the lowest having an $R^2 < 0.5$. A weak trend emerges from the skew shown in Figure 6, which implies that the model has greater difficulty trying to predict more expensive wards.

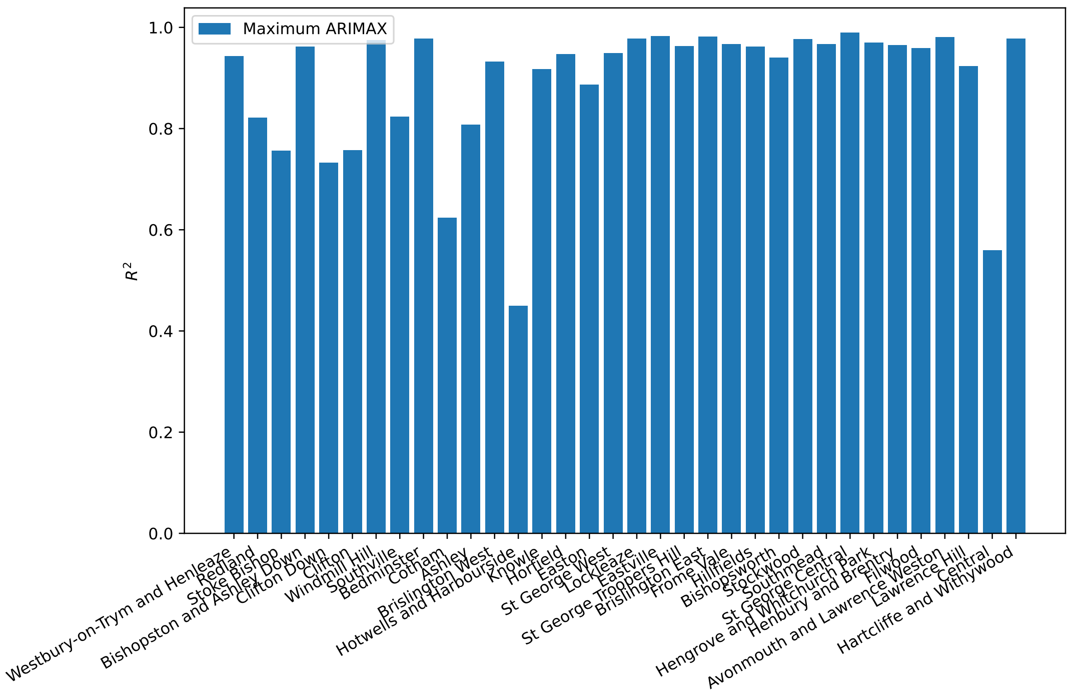

Figure 6: Maximum R^2 values for ARIMAX models trained on HPSSA 37 data with optimal Feature Ward usage, arranged by median house price high to low recorded in 2023.

4.1.1 Forecasting Stoke Bishop

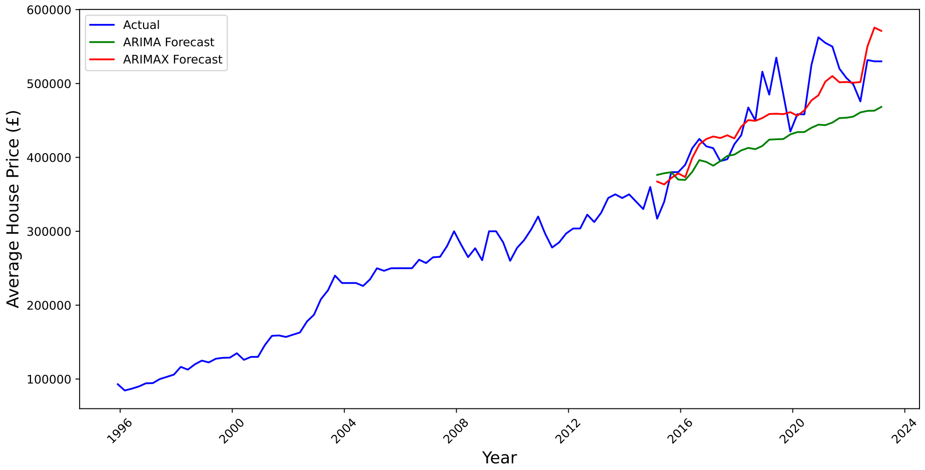

Figure 7: Comparison of ARIMA and ARIMAX testing data against predicted median house prices for Stoke Bishop

Figure 7 shows how the ARIMA and ARIMAX models compare when predicting prices from 2018 to 2024 for Stoke Bishop. The ARIMA model is clearly under fitted to the data producing an R^2 value of 0.1859 and failing to capture any of the peaks or troughs and only capturing the linear increase. However, the ARIMAX model is far more successful, producing an R^2 value of 0.7338. While not peaking as high as it should be, and failing to predict any dips at all, it captures the sharp rises the data exhibits well. It seems that both models have captured the long term linear increase in the data, but the models fail to capture the short duration dips that the market experiences.

4.2 City of Bristol Forecasting

For the City of Bristol forecasting using the larger HPI [13] dataset, more data can be allocated for training whilst maintaining a sufficient amount of data for evaluation. Therefore, each model will be trained and tested via a 80/20 train-test split on the entire data set.

The PCA analysis from section 2.3 allows us to ignore the breakdown of different house types, which is not contained in HPSSA 37 dataset, as the results implied that there is no housing type that disproportionately sways the market, as described in 1.

The ARIMAX model will also take the top 5 principle components of the PCA analysis in section 2.3 as exogenous features. The model will cumulatively add these principle components, with the final model chosen that which outputs the highest R^2 value between the original testing data and ARIMAX predictions - the model deemed most accurate.

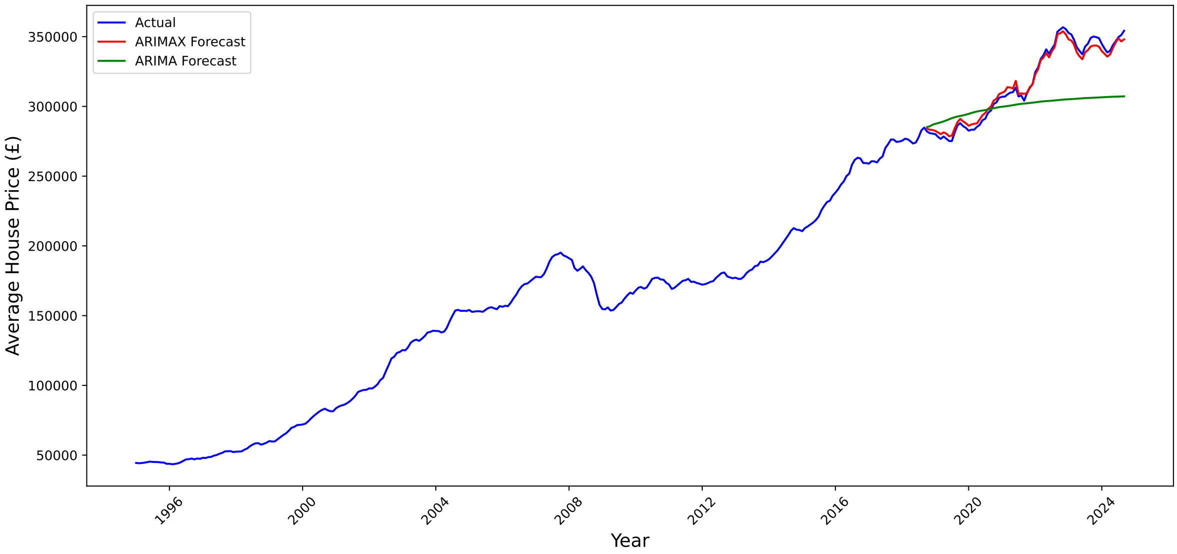

Figure 8: Comparison of ARIMA and ARIMAX testing data against predicted median house prices for the City of Bristol

Figure 8 shows the model's forecast on the test data from 2018-2024 alongside the actual data. Whilst the ARIMA model struggles, it is clear to see that the ARIMAX model performs exceptionally well, with a R^2 value of 0.9847. This makes it the most suitable model to predict future values.

When comparing to the models capabilities on the HPSSA 37 dataset, it is clear to see that a larger dataset that includes external economic indicators vastly improves the forecasting capabilities of ARIMAX, with highest the R^2 values recorded increasing significantly in this ARIMAX model compared to the R^2 of 0.7338.

As a result of this accuracy and implied reliability, the 12 months following September 2024 have been forecast, which are shown in Figure 9. The model predicts the prices will follow the same seasonal pattern, whilst increasing overall. These 12 months have been chosen to be forecasted due to the correlation strength found in this window size, detailed in section 2.2. Any longer and the reliability of predictions could decrease, especially with short term fluctuations.

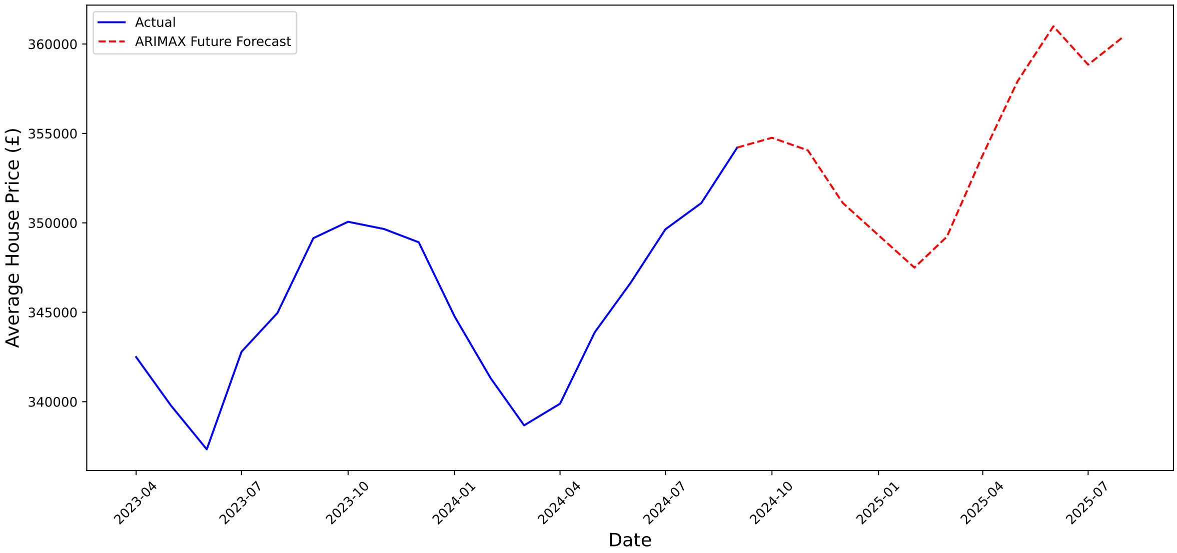

Figure 9: ARIMAX Forecast - Future House Prices: City of Bristol

4.2.1 Financial Crisis Robustness

To test the reliability of the models in the face of a financial crisis, an investigation is performed on how well they dealt with a major change within the data trends. The markets recovery after the Financial Crisis of 2008 [12] is used as the test dataset, with the years prior being used for training.

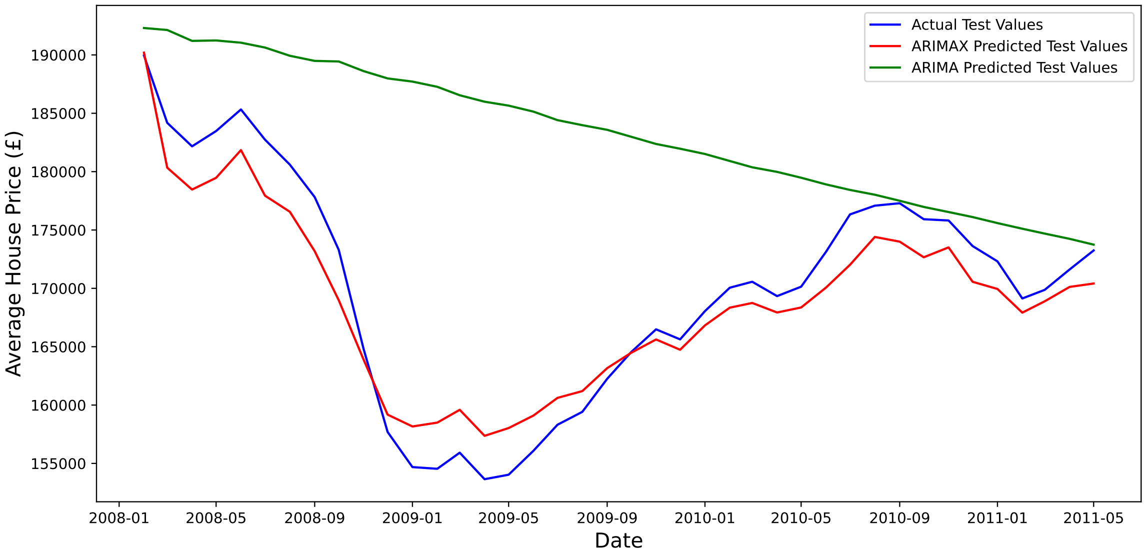

Figure 10: ARIMAX Post 2008 Test Data: City of Bristol

Figure 10 shows how the ARIMA and ARIMAX models perform against the 2008 Financial Crisis, it is clear to see that the ARIMA model's performance is very poor, reinforcing the conclusions of Stevenson [8] with respect to ARIMA models and their weak ability to forecast and consider financial crashes. However, the ARIMAX model closely follows the actual trends, with seemingly no lag.

5 Discussion

The difference between ARIMA and ARIMAX models is clear, with ARIMAX consistently outperforming its predecessor. This is unsurprising, with Gao [11] highlighting the benefits of a multivariate ARIMA in house price prediction. The ability of the more successful ARIMAX model in predicting the individual ward data varied across all wards, with most predictions being very accurate, and the more expensive wards performing worse when forecasted. There may be many underlying reasons for this, for example Figure 1 illustrates greater volatility in the prices of wealthier wards, indicating increased difficulty for a model to make accurate predictions. Furthermore, the figure reveals that the median ward price is skewed towards the minimum, indicating that most wards fall within the lower price range. Consequently, a multivariate model incorporating other wards as exogenous variables is expected to perform more consistently when predicting lower-priced wards, as they exhibit more similar price trends. As such, though Stoke Bishop is highlighted as an important ward due to its geography, its value in predictions is counteracted by these forces resulting in poor predictions for this ward, and being of low value in predicting other wards.

The final ARIMAX model forecasting city-wide house prices resulted in a very accurate prediction of the test data, qualitatively mimicking the micro-trends successfully. Furthermore, the model does extremely well over the 2008 financial crisis period, predicting a downfall with negligible lag. While more testing needs to be done on these models, it may be that recording price data on a more granular level is enough to drastically improve performance.

An analysis involving non-linear correlations is not performed, though could give additional insights that are not clearly displayed by the correlations measured. More complex predictors, such as an LSTM neural network could potentially make more reliable conclusions about the future housing market due to their complexity and more developed weights and insights, and as such could be developed in future works [9].

Not assuming the correlations are linear may also allow for a better choice of exogenous features. A Monte Carlo method may be able to find more effective wards to include as exogenous features, but is not attempted here due to the computation complexity of these models.

6 Conclusion

The empirical results of spatial, temporal, and PCA have been presented in this report. Stoke Bishop is highlighted as an influential ward, where price trends spill over into neighbouring wards, alongside other novel findings such as there not being one house type that disproportionately affects the housing market. Forecasting is modelled by ARIMA and ARIMAX models, with the highest success being achieved by the multivariate ARIMAX model. The more expensive wards tended to underperform when compared to those with lower house prices. The city-wide ARIMAX model yielded the highest success, closely fitting the test data. Finally the effects of the 2008 financial crash were analysed, as previous research has highlighted the potential weakness of ARIMA models in coping with crash behaviour. Both models when tested on the crash performed better than expected, being able to predict a dip in price but were not able to fully capture the unstable dynamics. The modelling objectives of forecasting Bristol's housing market is met, with models for individual wards as well as Bristol as a whole over the next year being created. This will hopefully encourage prospective homebuyers to make better informed decisions, and housing charities, such as Baggator, can understand Bristol's housing market with more accuracy. Future research could incorporate supply-side data, such as ward-specific construction data, to better capture underlying trends in house price volatility. Furthermore, incorporating information on upcoming housing projects could enhance the predictive accuracy of the model, particularly when combined with economic indicators signalling potential market downturns. This approach would help mitigate the limitation of relying solely on historical data for forecasting.

References

[1] Office for National Statistics. https://www.ons.gov.uk/.

[2] Open Data Bristol. https://opendata.bristol.gov.uk/datasets/eed2a0352d164c36ae810e026a7d9994_3/explore.

[3] M. H. Pesaran and Takashi Yamagata. The spatial and temporal diffusion of house prices in the uk. Technical report, Working Paper, [Institution Name], 2009. Please update the institution and any additional details as available.

[4] E. Pagourtzi, V. Assimakopoulos, T. Hatzichristos, and N. French. Real estate appraisal: A review of valuation methods. *Journal of Property Investment & Finance*, 21(4):383–401, 2003.

[5] Stephen Malpezzi. A simple error correction model of house prices. *Journal of Housing Economics*, 8(1):27–62, Mar 1999. doi: https://doi.org/10.1006/jhec.1999.0240.

[6] Toby Auterson. Working paper no.6 forecasting house prices. *OBR*, 6, 2014. URL https://obr.uk/docs/dlm_uploads/WP06-final-v2.pdf. Working Paper.

[7] Abhishek Yadav. A comparative study of time series, machine learning, and deep learning models for forecasting global price of wheat. *Operations Research Forum*, 5(4), Nov 2024. doi: https://doi.org/10.1007/s43069-024-00395-9.

[8] Simon Stevenson. A comparison of the forecasting ability of ARIMA models. *Journal Name*, Volume(Issue):Page–Page, Year. Incomplete bibliographic information; please update with full details.

[9] Choujun Zhan, Yonglin Liu, Zeqiong Wu, and Mingbo Zhao. A hybrid machine learning framework for forecasting house price. *Expert Systems with Applications*, 233:120981–120981, Dec 2023. doi: https://doi.org/10.1016/j.eswa.2023.120981.

[10] Hamed Ahmed Al-Marwani. Modelling and forecasting property types' price changes and correlations within the city of manchester, uk.

[11] Yue Gao. Predicting housing price in beijing using arima models. *Emory Theses and Dissertations*, 2022. URL https://etd.library.emory.edu/concern/etds/7p88ch78f. ETD ID: 7p88ch78f.

[12] Velma Loh / James Scruton. The 2008 recession 10 years on, Apr 2018. URL https://www.ons.gov.uk/economy/grossdomesticproductgdp/articles/the2008recession10yearson/2018-04-30.

[13] HM Land Registery. Uk house price index. https://www.gov.uk/government/statistics/uk-house-price-index-for-november-2024, 2024. Last accessed Monday 10th Febuary 2024.

[14] Office for National Statistics. Median house prices by ward: Hpssa dataset 37. https://www.ons.gov.uk/peoplepopulationandcommunity/housing/datasets/medianpricepaidbywardhpssadataset37, 2023. Last accessed Monday 10th Febuary 2024.

[15] Office for National Statistics. Ordnance survey uk bfc. https://geoportal.statistics.gov.uk/datasets/bd816ce7fa384745851f942786671f87_0/explore?location=51.469955%2C-2.548967%2C12.66&showTable=true, 2024. Last accessed Monday 10th Febuary 2024.

[16] Russell T. Warne. *Statistics for the Social Sciences*, volume 2. Dec 2020. ISBN 9781108894319. doi: https://doi.org/10.1017/9781108894319.

[17] Yihui Qiu Zhangyu Li. Feature selection based on improved principal component analysis. *ACM Digital Library*, 2023.

[18] Kenji Doya Florian Lalande. Numerical data imputation: Choose knn over deep learning. *Springer Nature*, 2022.

[19] Ian Mulheirn. *Tackling the UK housing crisis: is supply the answer?* 2019. URL https://housingevidence.ac.uk/wp-content/uploads/2024/05/20190820b-CaCHE-Housing-Supply-FINAL.pdf.

[20] Vaia I. Kontopoulou, Athanasios D. Panagopoulos, Ioannis Kakkos, and George K. Matsopoulos. A review of arima vs. machine learning approaches for time series forecasting in data driven networks. *Future Internet*, 15(8), 2023. ISSN 1999-5903. doi: 10.3390/fi15080255. URL https://www.mdpi.com/1999-5903/15/8/255.

A Appendix

A.1 Data handling

Data preprocessing is necessary for the development of forecasting models. Empty columns were removed from the analysis, and in cases where NaN's needed to be imputed, a KNN imputation [18] of the missing data is performed. This particular NaN imputation process involves finding the k-nearest neighbours to the missing data, which are data points most similar to the missing data point using the other available features / data entries, then using a weighted average of these neighbours respective values as an imputed estimate for the NaN value. For this analysis, the 5 nearest neighbours were used for the imputation estimate. This parameter choice provided durability to outliers whilst also not removing any understanding of local patterns due to over-smoothing, where generalisations are made that may lose insights into patterns within the data.

When data normalisation is necessary, the python module Sklearn's StandardScaler is implemented, using a z-score normalisation. This is performed after the handling the NaN imputation. For the PCA implementation, Sklearn's PCA is utilised.