Alícia Chimeno Sarabia
Silvia Ferrer Vázquez

ARIMAX Model for Real Case Series

Advanced Statistical Modeling

Abstract

This project uses advanced statistical modeling to predict London housing prices through time series analysis, specifically with an Seasonal ARIMA model applied to the UK House Price Index (UK HPI) from January 1995 to November 2024. The goal is to estimate future price trends and help determine savings needed for property purchase in the next year. The analysis addresses stationarity and variance issues, followed by model assessments using autocorrelation, residual analysis, and forecast accuracy.

Project Description

Data

The UK House Price Index (UK HPI) captures changes in the value of residential properties. The UK HPI uses sales data collected on residential housing transactions, whether for cash or with a mortgage. Properties have been included: from England, Scotland, and Northern Ireland. Data is available at a national and regional level, as well as counties, local authorities and London boroughs.

For more information on the dataset, see [2], and for the data source, see [1]. In our use case, we used only information about London.

1 Time Series Analysis

1.1 Identification

The London housing price series is from January 1995 until November 2024 with monthly frequency.

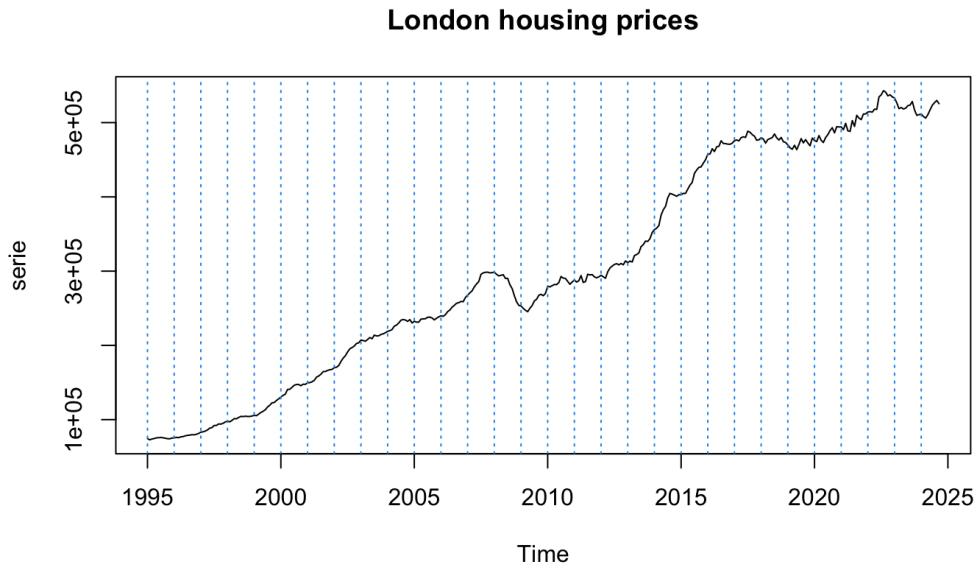
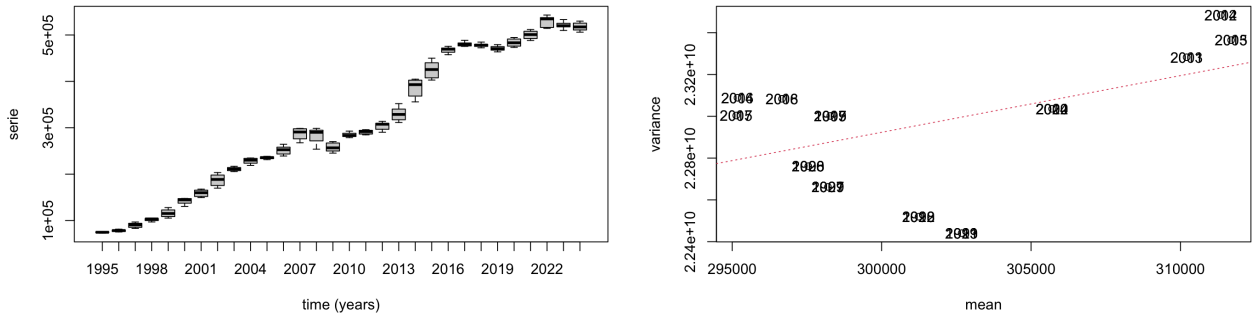


Figure 1: Original Time Series (`serie`)

1.1.a) Stationarity

We shall diagnose if the series is stationary. If not, try to transform the series into a stationary one.

Is the variance constant? To diagnose the non-constant variance, we will check 2 plots: Box plot and mean/var plot



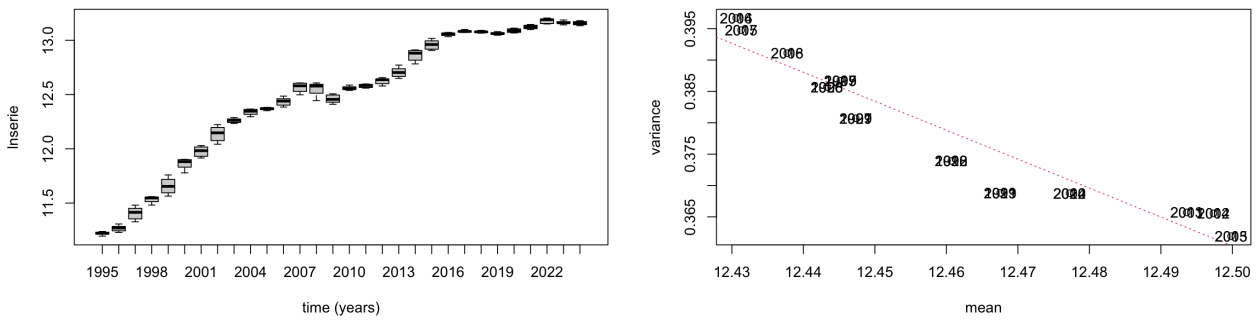
(a) Boxplot (b) Mean-var plot

Figure 2

In plot (a), we observe that the variance is not consistent across all years, as indicated by the varying box sizes. Larger boxes correspond to higher variance, while smaller boxes suggest lower variance. Plot (b) reveals a positive relationship between the mean and variance: higher mean values correspond to higher variance. This behavior suggests the need for a scale transformation to stabilize the variance. To investigate further, the Box-Cox method is applied to determine whether a logarithmic transformation would be appropriate.

```
> BoxCox.lambda(serie+1)
[1] 0.5087956
```

The Box-Cox lambda estimate is approximately 0.51, which is close to zero. This supports the use of a logarithmic transformation for the data. After applying the transformation, we reevaluate the variance using the same plots.



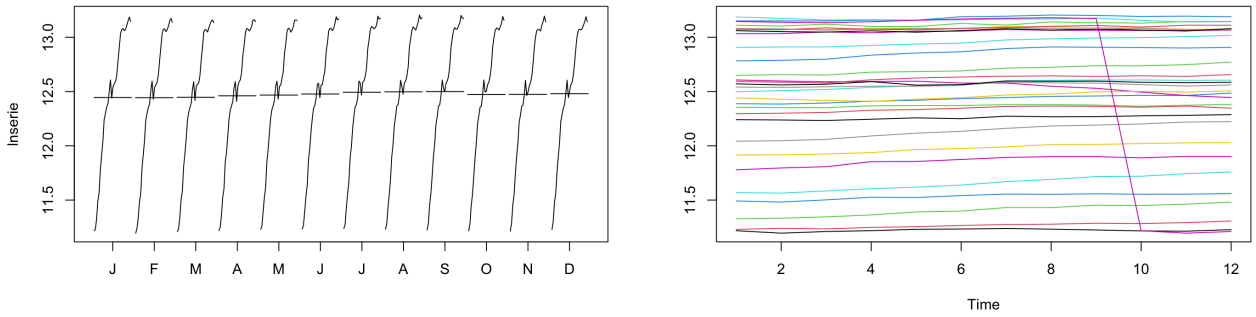
(a) Boxplot (b) Mean-var plot

Figure 3

The post-transformation plot (a) shows smaller, more consistent box sizes, indicating improved

variance stability. In plot (b), higher mean values now correspond to smaller variances, suggesting that the transformation reduced the mean-variance relationship, achieving more constant variance. This validates the effectiveness of logarithmic transformation, a common approach to price data, and supports proceeding with the transformed series, `lnserie`, for further analysis. See the time series plot in 17.

Is there a Seasonal Pattern?



(a) Monthplot (b) ts.plot by month

Figure 4

The monthly plots indicate consistent behavior throughout the months, suggesting that there is no significant monthly seasonal pattern in the data. However, the additive decomposition of the time series reveals a seasonal pattern. Refer to the additive decomposition plot in Figure 18 for a visual representation. To further investigate this, we examine the ACF plot to identify the correlation and the specific nature of the seasonal pattern.

Is the mean constant? The original time series, `serie` 1 , and the transformed one, `lnserie` 17, seem to follow a linear or quadratic trend. The mean is not constant due to the visible upward trend. We will apply one regular difference trying to make the mean constant, `d1lnserie`.

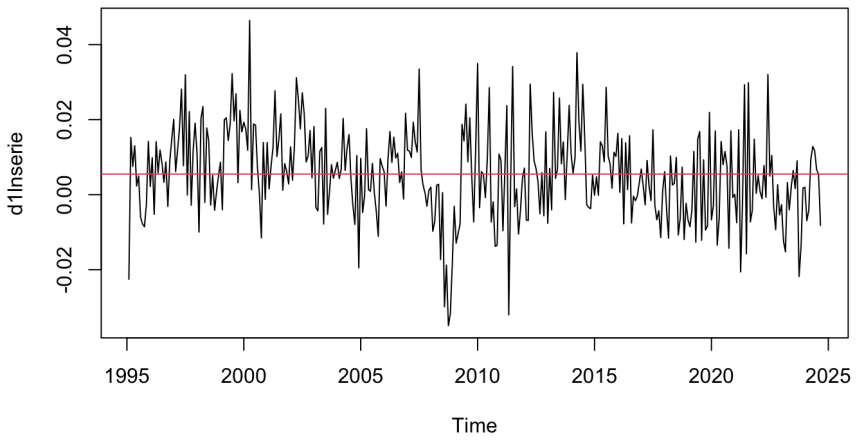
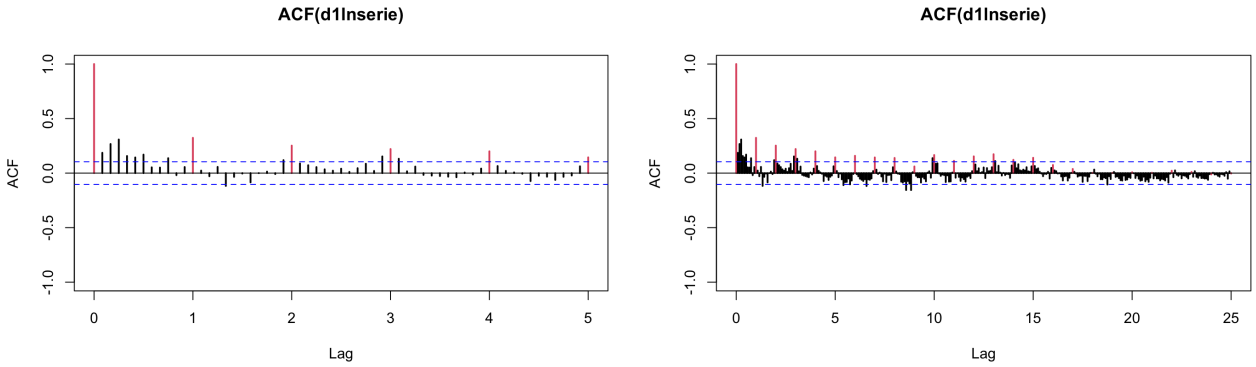


Figure 5: Differenciated Time Series (`d1lnserie`)

The mean is now close to 0 and follows a more stable and constant trend. To check for over-differentiation, we perform an additional differencing and compare the variances. Our results indicate that no further differencing is necessary, as the variance remains stable after the first difference.

1.1.b) Auto-correlation analysis

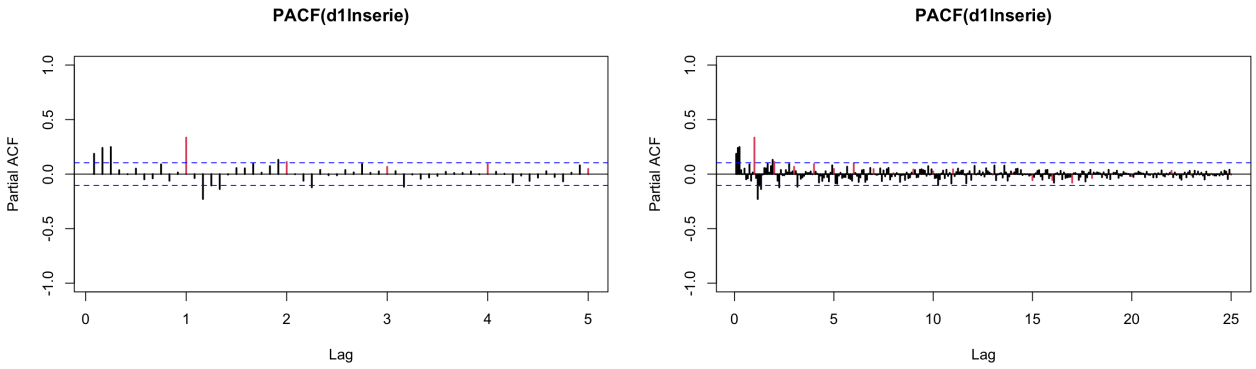
In order to build an arima mode, we need to plot the ACF and the PACF. Then we have to choose the parameters p and q. We suspected that our data had a yearly seasonal patter from our plot in the additive serie decomposition, therefore we colored each year lag in red.



(a) ACF Plot Showing 5 Lags (b) ACF Plot Showing 25 Lags

Figure 6

The ACF plot exhibits a distinct spike at lag 12, and upon closer inspection, a recurring cycle every 12 months suggests the presence of yearly seasonality in the data.



(a) PACF Plot Showing 5 Lags (b) PACF Plot Showing 25 Lags

Figure 7

For the seasonal component, we consider multiple lags corresponding to the seasonality period. In this case, we use D = 0, P = 1, and Q = 1, as both the autocorrelation and partial autocorrelation functions suggest that the seasonal effects eventually fade to zero. There is no clear recurring pattern that justifies the inclusion of higher-order AR or MA terms, indicating that AR(1) and MA(1) are appropriate for modeling the seasonal structure.

To identify the non-seasonal components, we examine the ACF and PACF plots for a few lags. The ACF shows a slowly decaying pattern. The PACF reveals that the last lag is significantly non-zero, which suggests an autoregressive process of order 1 (AR(1)). Based on these observations, we will experiment with an AR(1) model and also try an ARMA(1, 1) model to capture both the autoregressive and moving average components in the regular (non-seasonal) part. As seen in the previous section, we will use one differentiation (d=1) although alse seems worth to try (d=2) because the data seems to be similar to a quadratic.

- `ARMA(p=1,d=1,q=0)(P=1,D=0,Q=1)[12]`
- `ARMA(p=1,d=1,q=1)(P=1,D=0,Q=1)[12]`
- `ARMA(p=2,d=2,q=1)(P=1,D=0,Q=1)[12]`

1.2 Estimation

1.2.a) Estimation of the models

MODEL 1: ARIMA(1,1,0)(1,0,1)[12]

```
arima(x = lnserie, order = c(1, 1, 0), seasonal = list(order = c(1, 0, 1), period = 12))

Coefficients:
         ar1    sar1     sma1
      0.1715  0.9571  -0.7920
s.e.  0.0535  0.0299   0.0754

sigma^2 estimated as 0.0001214:  log likelihood = 1096,  aic = -2184
```

MODEL 2: ARIMA(1,1,1)(1,0,1)[12]

```
arima(x = lnserie, order = c(1, 1, 1), seasonal = list(order = c(1, 0, 1), period = 12))

Coefficients:
         ar1      ma1    sar1     sma1
      0.9119  -0.7297  0.9714  -0.8490
s.e.  0.0317   0.0465  0.0237   0.0659

sigma^2 estimated as 0.0001043:  log likelihood = 1122.49,  aic = -2234.99
```

MODEL 3: ARIMA(2,2,1)(1,0,1)[12]

```
arima(x = lnserie, order = c(2, 2, 1), seasonal = list(order = c(1, 0, 1), period = 12))

Coefficients:
          ar1      ar2      ma1    sar1     sma1
      -0.5990  -0.2677  -0.4011  0.9880  -0.8849
s.e.   0.0972   0.0796   0.0952  0.0105   0.0480
```

```
sigma^2 estimated as 9.665e-05:  log likelihood = 1129.93,  aic = -2247.86
```

1.3 Validation

1.3.a) Residual Analysis

In the model, we assumed that the random noise follows a white noise process, that is, $Z_t \sim \mathcal{N}(0, \sigma^2)$. To validate this assumption, we should verify whether the 3 conditions regarding the distribution are satisfied through residual analysis.

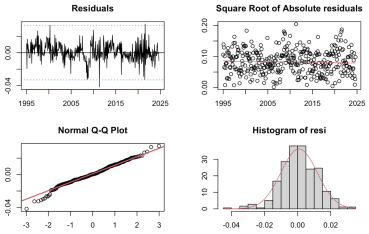


Figure 8: Caption for Plot 1

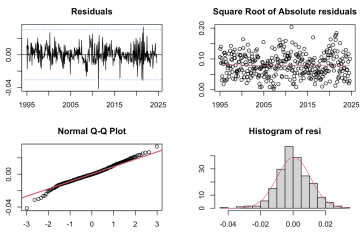


Figure 9: Caption for Plot 2

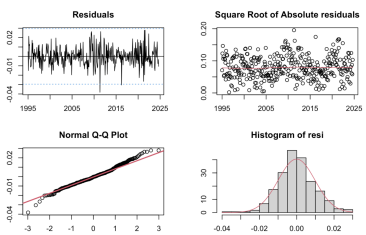


Figure 10: Caption for Plot 3

Figure 11: A collection of three plots.

Model 1:

- σ_t^2 constant:
  - The Breusch-Pagan test indicated no evidence of heteroscedasticity (p-value > 0.05). Therefore, the assumption of constant variance is satisfied.
  - Residual plots: seem to have behave constant except several outliers.
- σ_t^2 normal:
  - QQ-plot: Data seems to be normal distributed because almost all the data follow the line except there are a few elements in the tail that should be treated.
  - Normality tests (Shapiro-Wilk, Anderson-Darling, and Jarque-Bera) indicate significant deviations from normality ($p < 0.05$ for all tests). This suggests that the residuals are not normally distributed, which might indicate a need for model refinement or transformation of data.
- σ_t^2 independence:
  - The Durbin-Watson test show that there is no significant evidence of autocorrelation in the residuals.
  - Ljung-Box test: While the p-value at lag 1 is not significant, the p-values for higher lags (from 2 onward) are extremely small, indicating strong evidence of significant autocorrelation in the residuals. This suggests that while global tests (e.g., Durbin-Watson) indicate no overall autocorrelation, specific patterns or cycles of autocorrelation are present at certain lags.

Model 2: has the same conclusions. Confirms constant variance, fails to proof normality and no overall autocorrelation but significant autocorrelation in several lags.

Model 3:

- σ_t^2 constant:
  - The Breusch-Pagan test indicates that the residuals in the model likely have constant variance. Same as in model 1 and 2.
  - Residual plots: seem to have behave the same as the previous models.
- σ_t^2 normal:
  - QQ-plot: Data seems to be less normal distributed than in model 1 (and 2) in the upper tail but better in the lower tail.
  - Normality tests indicate that data is not normally distributed, same as in the 2 previous models.
- σ_t^2 independence:
  - The Durbin-Watson test show that there is no significant evidence of autocorrelation in the residuals. The DW statistic of 1.995 is close to 2, suggesting that the residuals do not exhibit significant autocorrelation. Similar as the other 2 models.
  - Ljung-Box test: The main difference between the model 3 and model 1 and 2 is this test. For all lags (1 to 48), the p-values are high, suggesting there is no significant evidence of autocorrelation at these lags. In the opposite as model 1 and 2 that detected autocorrelation between lags. See results in

1.3.b) Causality and invertibility

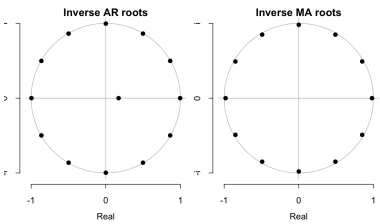


Figure 12: Caption for Plot 1

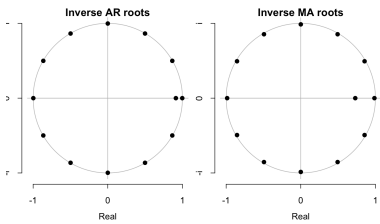


Figure 13: Caption for Plot 2

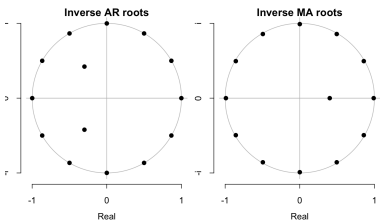


Figure 14: Caption for Plot 3

Figure 15: A collection of three plots.

Model 1: The model is both causal and invertible. This can be confirmed analytically by observing that the modulus of the AR characteristic polynomial roots and the modulus of the MA characteristic polynomial roots are greater than 1. Additionally, the model is graphically confirmed as causal and invertible, since the inverse roots of both the AR and MA polynomials lie inside the unit circle.

Model 2: The model is both causal and invertible. Same reason as for model 1.

Model 3: The model is both causal and invertible. Same reason as for model 1.

Expressions analysis: Parameters are in the Annex 2.4

Model 1: Model 1 clearly identifies a strong seasonal effect with periodicity at lags of 12 and 24 because the other parameters are close to 0. The rapid decay of Psi-weights after initial spikes suggests that shocks have short-lived impacts beyond the seasonal pattern.

Model 2: Similar to Model 1, Model 2 highlights seasonal periodicity at lags of 12 and 24. The presence of oscillations in Pi-weights suggests interactions between AR components, leading to alternating responses over time.

Model 3: Model 3 appears less stable than Models 1 and 2, as indicated by initial negative weights and persistent oscillations. Periodicity at lags of 12 and 24 is weaker and more irregular, suggesting a different dynamic or noise in the model.

1.3.c) Evaluate predictions capability

We will compare the 3 models with the model metrics (AIC, loglike and BIC) and prediction metrics (ME, RMSE and MAE) in the training set and the test set.

	AIC	Loglike	BIC
Model 1	-2184	1096	-2168.5
Model 2	-2234.99	1122.49	-2215.61
Model 3	-2247.86	1129.93	-2224.63

Model 1:

```
                      ME       RMSE         MAE
Training set  0.001059155 0.01110578 0.008561819
Test set     -0.028730149 0.02996202 0.028730149
```

Model 2:

```
                       ME       RMSE         MAE
Training set  0.0004948748 0.01029705 0.007865779
Test set     -0.0259639947 0.02722785 0.025963995
```

Model 3:

```
                       ME        RMSE         MAE
Training set -3.448897e-05 0.009891862 0.007631939
Test set     -2.882083e-02 0.029704729 0.028820833
```

The best model in model metrics is clearly model 3, ARIMA(2,2,1)(1,0,1)[12], and the validation of the model is more accurate in model 3.

1.4 Predictions

In this section we will forecast the future 12 months from the best model selected and provide also confidence intervals. We had to transform our predictions from the logarithmic scale into the original one for a better interpretation.

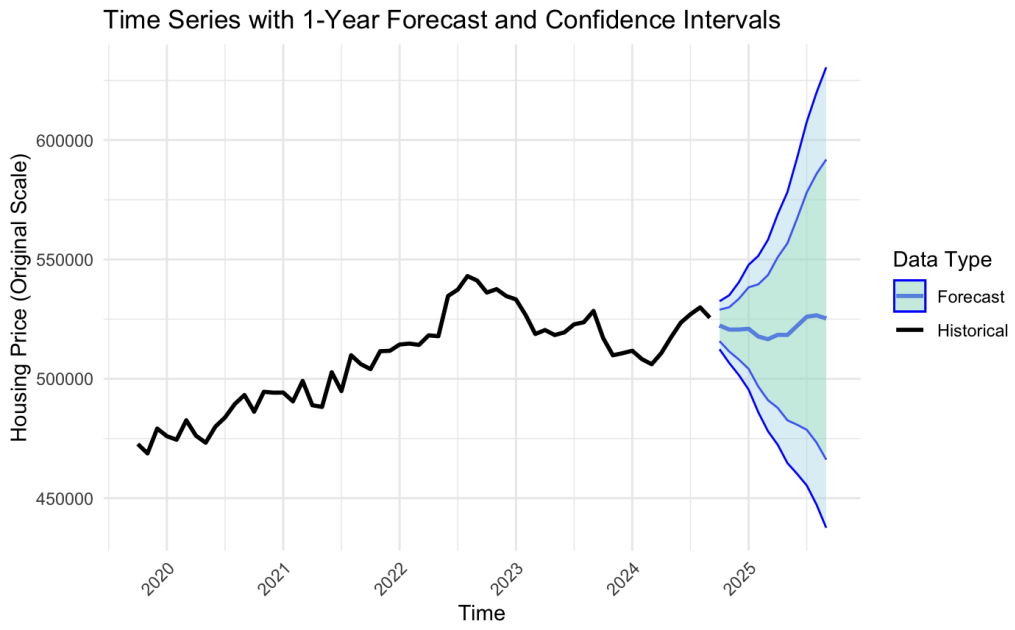


Figure 16

1.5 Conclusions

- The forecast values suggest a slightly fluctuating but overall increasing trend in house prices, with values predicted in the next year generally around the range of £520,000-£526,000.
- The widening confidence intervals indicate greater uncertainty in the forecasts as time progresses. Suggests caution when interpreting the predictions far into the future.
- The upper and lower bounds highlight potential risks and opportunities: At the lower end of the 95% intervals, the price could drop to as low as £455,300 in mid-2025, while at the upper end, it could rise to £607,686. If planning investments, consider these bounds to estimate best and worst-case scenarios.
- For the majority of 2025, the forecast median values hover around £520,000-£522,000. This likely reflects a stabilization in the market with mild growth
- Beyond the time series, additional factors like economic indicators, new housing development, political changes, and infrastructure projects,... should be considered to refine forecasts.

1.6 Motivation

The primary motivation behind analyzing the housing market in London is to make informed, data-driven decisions before purchasing a property. One of us is planning to move to London, and such a decision is both exciting and significant. Understanding housing price trends and the factors that influence them is crucial to gaining insights into the affordability and potential long-term value of properties. Additionally, this analysis highlights the importance of securing a well-paying job to support such a life-changing move. By leveraging time series data and incorporating relevant factors, this study aims to provide prospective buyers with a deeper understanding of the market, enabling them to make well-informed choices in a competitive housing environment.

2 Annex

2.1 Transformed Time series

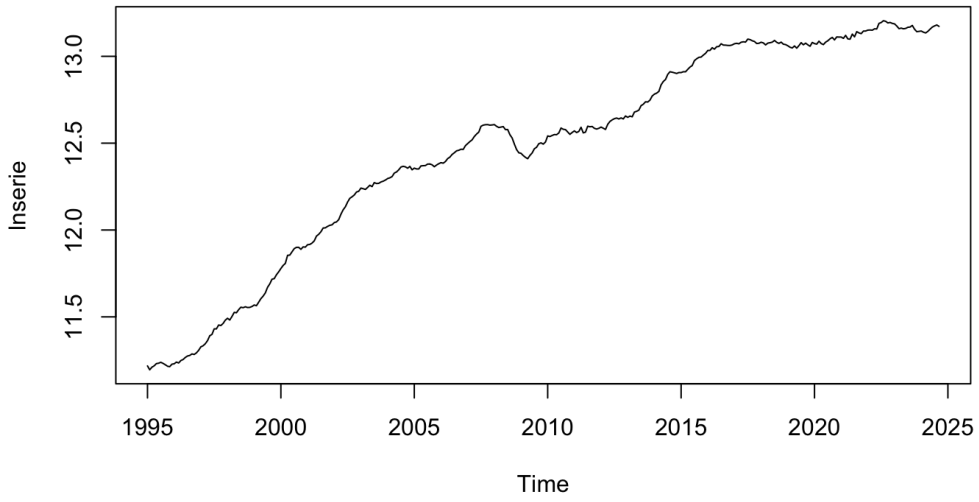


Figure 17: Forecast

2.2 Additive Decomposition of `lnserie`

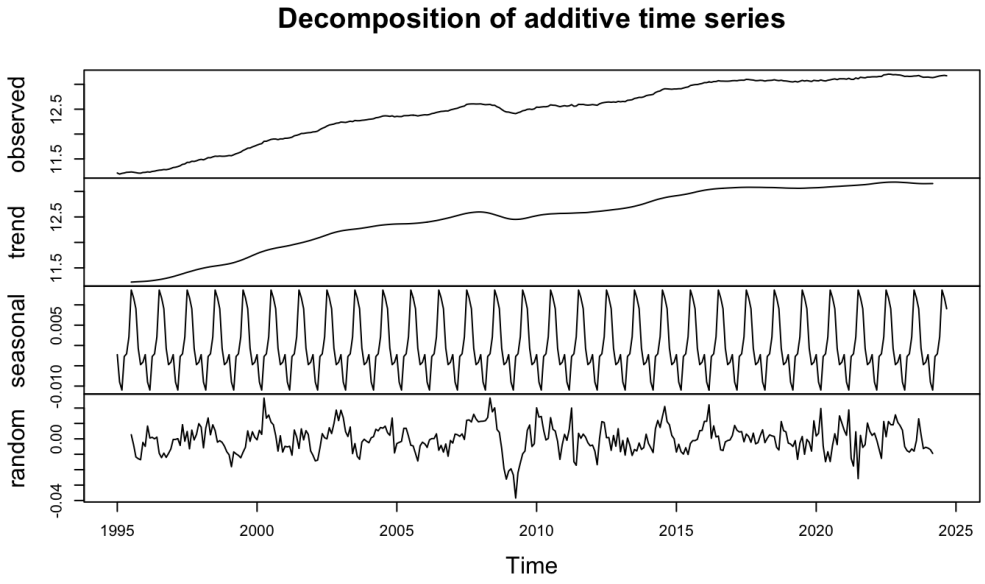


Figure 18: Additive Decomposition of `lnserie`

2.3 Ljung-Box test model 3

```
Ljung-Box test
     lag.df    statistic    p.value
```

```
[1,]      1 1.022366e-04 0.99193256
[2,]      2 9.446917e-03 0.99528768
[3,]      3 3.271435e-02 0.99844164
[4,]      4 4.883515e-02 0.99970670
[5,]     12 1.753563e+01 0.13053495
[6,]     24 3.390122e+01 0.08643994
[7,]     36 4.046930e+01 0.27952235
[8,]     48 4.553930e+01 0.57423503
```

2.4 Expression Parameters

2.4.a) Model 1

```
Psi-weights (MA(inf))

--------------------
       psi 1        psi 2        psi 3        psi 4        psi 5        psi 6        psi 7
1.717681e-01 2.950428e-02 5.067894e-03 8.705026e-04 1.495246e-04 2.568355e-05 4.411615e-06 7.5
       psi 9       psi 10       psi 11       psi 12       psi 13       psi 14       psi 15
1.301615e-07 2.235760e-08 3.840323e-09 1.790264e-01 3.075103e-02 5.282047e-03 9.072871e-04 1.5
      psi 17       psi 18       psi 19       psi 20       psi 21       psi 22       psi 23
2.676885e-05 4.598035e-06 7.897958e-07 1.356617e-07 2.330236e-08 4.002602e-09 6.875193e-10 1.6

Pi-weights (AR(inf))

--------------------
       pi 1        pi 2        pi 3        pi 4        pi 5        pi 6        pi 7        pi
 0.17176810  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.0000000
      pi 10       pi 11       pi 12       pi 13       pi 14       pi 15       pi 16       pi 1
 0.00000000  0.00000000  0.17902644 -0.03075103  0.00000000  0.00000000  0.00000000  0.0000000
      pi 19       pi 20       pi 21       pi 22       pi 23       pi 24
 0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.13766340
```

2.4.b) Model 2

```
Psi-weights (MA(inf))

--------------------
     psi 1      psi 2      psi 3      psi 4      psi 5      psi 6      psi 7      psi 8      p
0.18225199 0.16620332 0.15156786 0.13822116 0.12604974 0.11495010 0.10482787 0.09559698 0.0871
    psi 11     psi 12     psi 13     psi 14     psi 15     psi 16     psi 17     psi 18     ps
0.07250141 0.18855315 0.08260921 0.07533484 0.06870104 0.06265139 0.05713446 0.05210334 0.0475
    psi 21     psi 22     psi 23     psi 24
0.03951553 0.03603588 0.03286265 0.14890470

Pi-weights (AR(inf))
```

```
--------------------
         pi 1          pi 2          pi 3          pi 4          pi 5          pi 6          p
 0.1822519882  0.1329875311  0.0970397284  0.0708089609  0.0516686209  0.0377020981  0.0275108
         pi 8          pi 9         pi 10         pi 11         pi 12         pi 13         pi
 0.0200744112  0.0146481057  0.0106885825  0.0077993563  0.1281271535 -0.0181614587 -0.0132522
        pi 15         pi 16         pi 17         pi 18         pi 19         pi 20         pi
-0.0096700345 -0.0070561316 -0.0051487917 -0.0037570240 -0.0027414645 -0.0020004204 -0.0014596
        pi 22         pi 23         pi 24
-0.0010651201 -0.0007772079  0.1033781501
```

2.4.c) Model 3

```
Psi-weights (MA(inf))

--------------------
        psi 1         psi 2         psi 3         psi 4         psi 5         psi 6         ps
-1.0037564142  0.3524153187  0.0579184702 -0.1344799496  0.0683841776 -0.0055422448 -0.0155502
        psi 8         psi 9        psi 10        psi 11        psi 12        psi 13        psi
 0.0113139733 -0.0027814552 -0.0014011336  0.0016545841  0.1277972043 -0.1289816693  0.0454803
       psi 15        psi 16        psi 17        psi 18        psi 19        psi 20        psi
 0.0073197301 -0.0172577739  0.0088071534 -0.0007309854 -0.0019918978  0.0014551144 -0.0003599
       psi 22        psi 23        psi 24
-0.0001787941  0.0002125384  0.1257336180

Pi-weights (AR(inf))

--------------------
         pi 1          pi 2          pi 3          pi 4          pi 5          pi 6          p
-1.003756e+00 -6.551116e-01 -2.459149e-01 -9.231119e-02 -3.465164e-02 -1.300749e-02 -4.882732e
         pi 8          pi 9         pi 10         pi 11         pi 12         pi 13         pi
-1.832873e-03 -6.880213e-04 -2.582685e-04 -9.694846e-05  1.284105e-01  1.289158e-01  8.414196e
        pi 15         pi 16         pi 17         pi 18         pi 19         pi 20         pi
 3.158509e-02  1.185637e-02  4.450626e-03  1.670670e-03  6.271337e-04  2.354126e-04  8.836885e
        pi 22         pi 23         pi 24
 3.317177e-05  1.245197e-05  1.093235e-01
```

References

[1] London Datastore. *UK House Price Index*. Accessed: 2023-12-21. 2023. URL: https://data.london.gov.uk/dataset/uk-house-price-index.

[2] UK Government. *About the UK House Price Index*. Accessed: 2023-12-21. 2023. URL: https://www.gov.uk/government/publications/about-the-uk-house-price-index/about-the-uk-house-price-index#data-tables?utm_medium=GOV.UK&utm_source=datadownload&utm_campaign=data_tables&utm_term=9.30_21_03_17.