Assignment 3: The Danish housing market

02417 - Time Series Analysis

GROUP

Pavlou, Ioannis - s212858
Blachet, Apolline - s222903
Kapakoglou, Georgios - s223001
Kozaris, Charalampos - s230224

January 20, 2024

Contents

List of Figures

List of Tables

Part 1 - Univariate models

In the first part of the assignment a suitable ARIMA model is found for the average sales price per square meter for houses in Denmark. The dataset that is used comprises of quartely values of sales prices between 1992 - 2022, along with quarterly values of the inflation rate and the interest rate from the Danish National Treasury. The average prices are given for all of Denmark and averaged within the capital region, Sealand, Middle Jutland and Rural areas (defined as southern Denmark and north Jutland). All data is obtained from Danmarks Statistik (`https://www.dst.dk/da/Statistik/emner/oekonomi`). The data are used in order to estimate the components of the model and the contribution of inflation and interest rate as external regressors is investigated. The determined model is then used to make predictions about the house prices for future time steps.

Question 4.1: Presenting the data

Firstly, the average sales prices of houses for each yearly quarter are presented in Figure 1. An increasing trend is clearly evident, so the time series is not considered stationary and the first order moment representation (mean) changes over time. Secondly, the evolution of interest rates over time is presented in Figure 2. A clear decreasing trend can be observed, again leading to the conclusion that the time series is not stationary. Lastly, the inflation rate for the same time period is presented in Figure 3. Even though some stationarity can be observed for a great part of the time series, the ten last time-steps lead to the conclusion that the time series is not stationary.

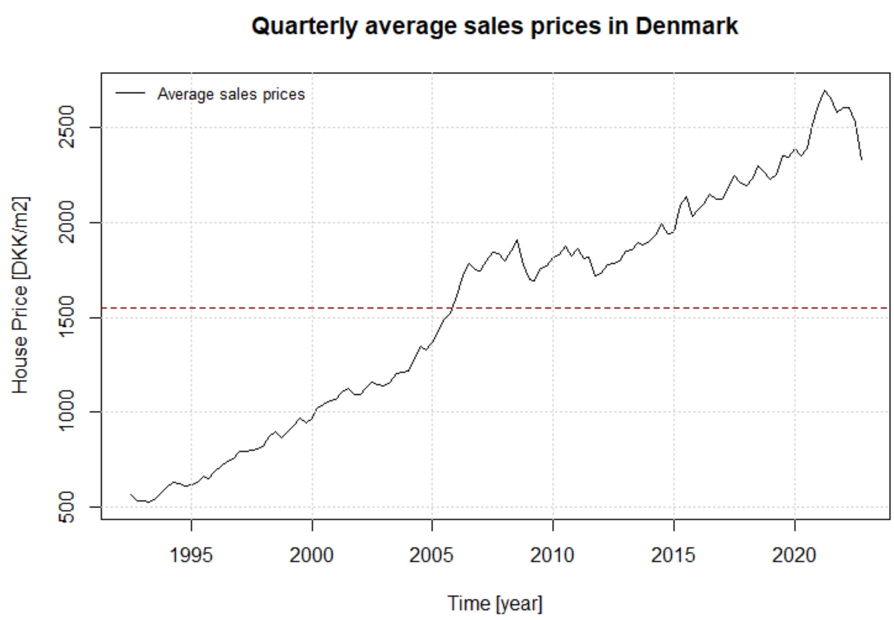

Figure 1: Quarterly average sales prices

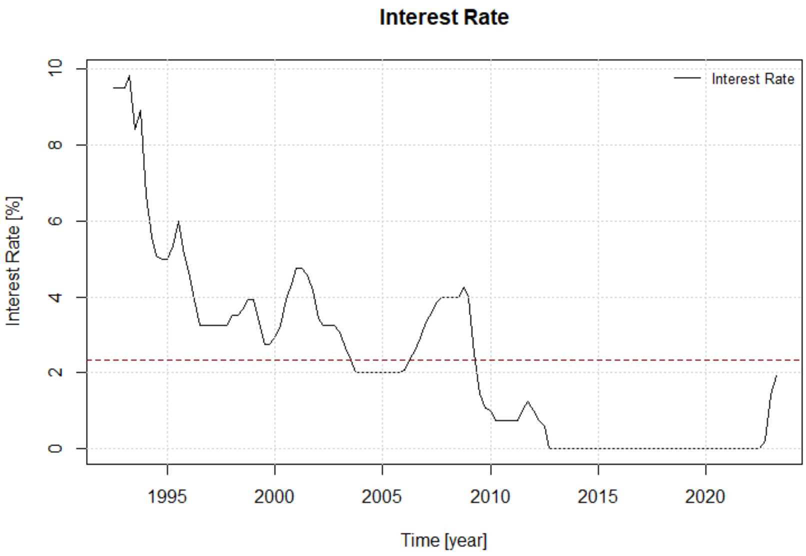

Figure 2: Interest rate

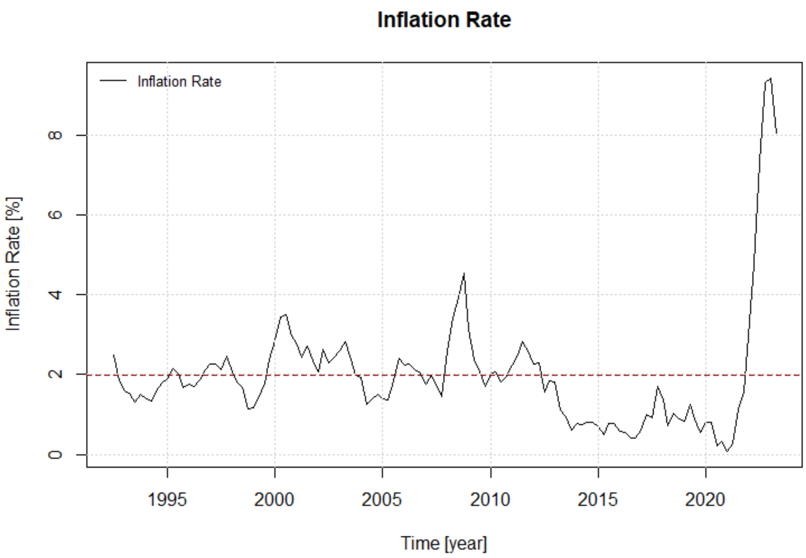

Figure 3: Inflation rate

A non-stationary time series has statistical properties that change over time, such as a trend, and in this case the patterns observed in the past may not continue in the future, making it difficult to build a reliable model. Therefore, it is important to ensure stationarity in the time series. In order to do so a type of transformation, such as differencing, should be considered.

Question 4.2: ACF and PACF

The ACF and PACF plots for the house prices data are presented in Figure 4. The same plots were also produced for the differenced data and are presented in Figure 5. Since the raw time series data are not stationary, they cannot be used to draw reliable conclusions about the model structure. As a result, only the plots of the transformed data are considered to make an assumption about a potential model structure.

In the differenced ACF plot we observe two almost significant quarter lags in the first year and no exponential decay, but an implied oscillation. In the differenced PACF plot we observe two significant quarter lags in the first year. However, in both plots a seasonality is implied in lag equal to one year (four quarters). Thus, a first model structure approach would be and a Seasonal ARIMA $(0, 1, 2) \times (2, 0, 0)_{[4]}$, assuming that the auto-regregressive component derives from the seasonal part of the model and the fact that we difference our data to achieve stationarity.

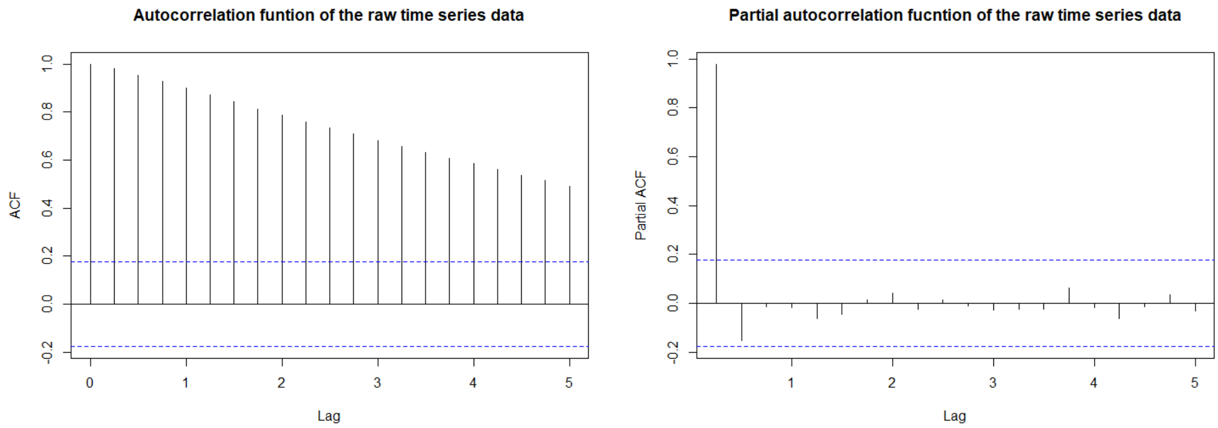

Figure 4: ACF and PACF of the raw time series data respectively

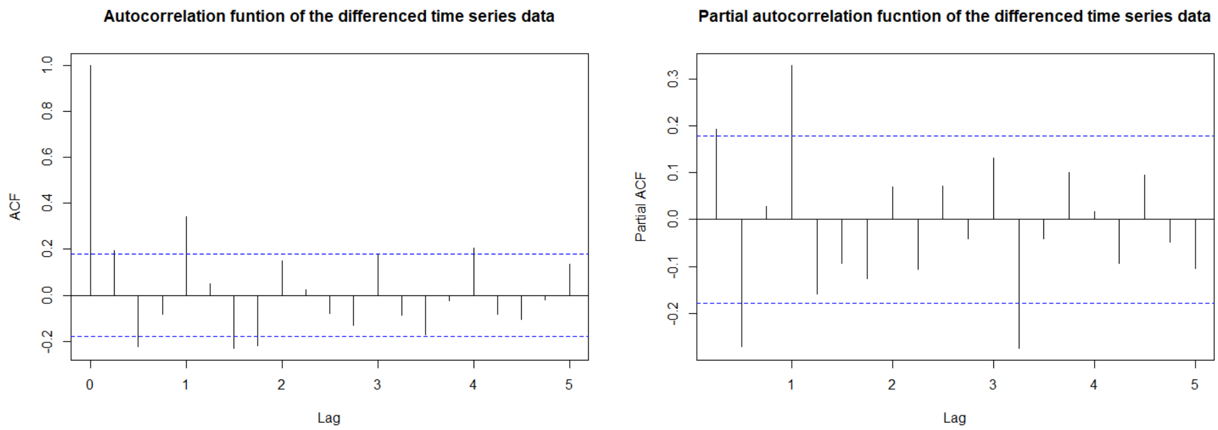

Figure 5: ACF and PACF of the differenced time series data respectively

The plots of the cross-correlation function between the house sales times series of each region are shown in Figure 6. Since the data are not stationary, these plots do not offer information about the cross-correlation between the times series. Therefore, a robust approach would be to apply pre-whitening with our first guess of model structure and examine the true cross-correlation between the different regions of Denmark regarding their house pricing.

The plots derived after applying pre-whitening are illustrated in Figure 7. The goal of pre-whitening is not to eliminate the cross-correlation between the series, but rather to make it easier to interpret and detect the relationship between the series by removing the autocorrelation within each series. In other words, pre-whitening is not meant to completely remove the dependence between the series, but rather to simplify the analysis by removing the effects of autocorrelation within each series.

So, the high correlation at lag 0 is not necessarily a sign that the pre-whitening is not working properly. It's important to look at the correlation at other lags to see if there is a significant cross-correlation between the prewhitened series. Overall, the fact that the lag 0 CCF values are high suggests that there may be a strong linear relationship between the pre-whitened series at lag 0, while the non-significance of other lags does not necessarily indicate that there is no relationship between the series at those lags. However, it shall be noted that "Sealand" has significant cross-correlation for more lags, which potentially indicates that the prices on that region are affected linearly by the other regions as well.

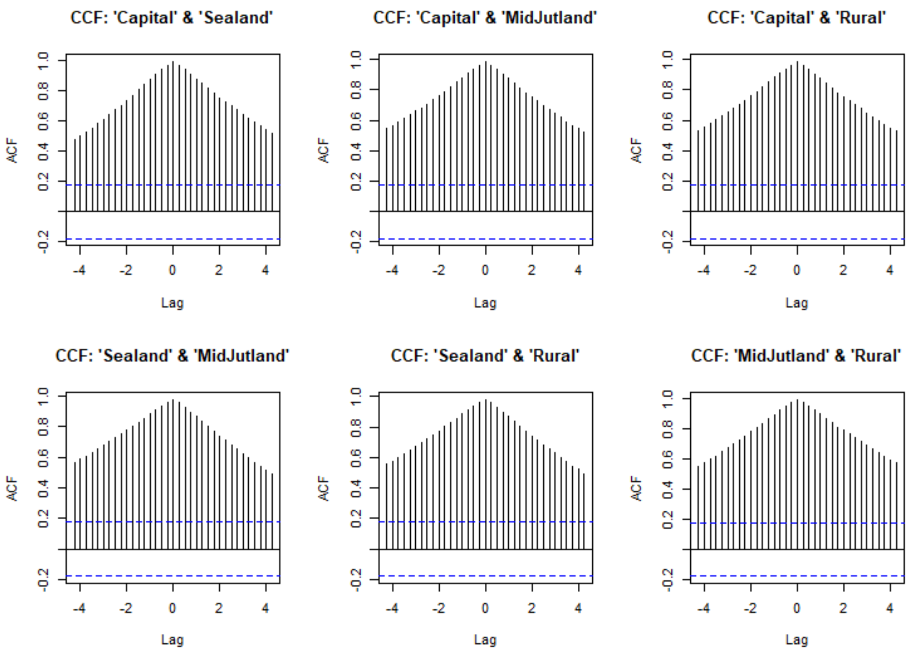

Figure 6: Cross-correlation function between the four raw region time series data

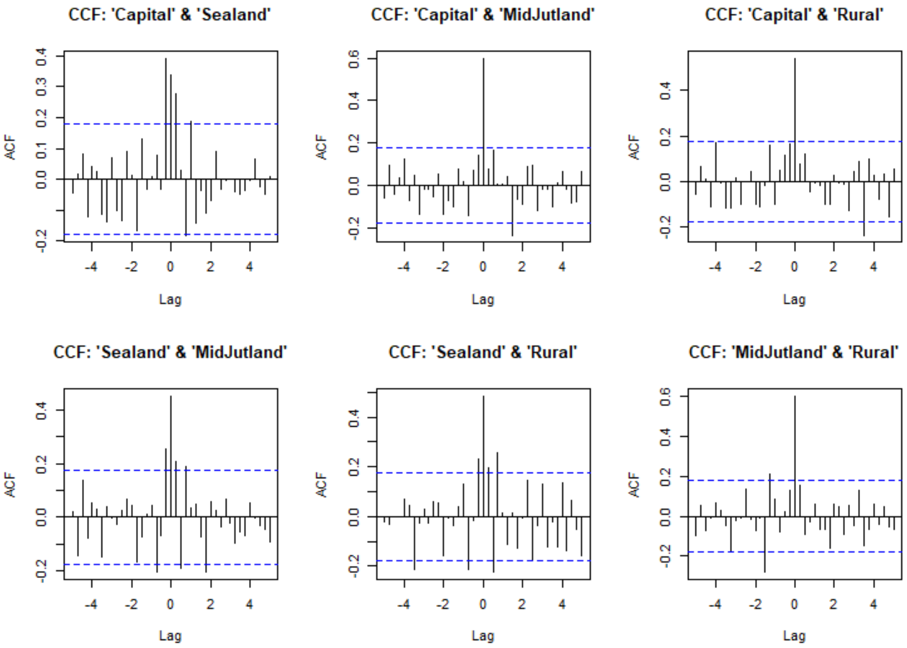

Figure 7: Cross-correlation function between the four pre-whitened region time series data

Question 4.3: Univariate model selection

In order to find a suitable univariate model for house prices in Denmark, the "auto.arima" function in "RStudio" was used:

Listing 1: Univariate model selection

```
model <- auto.arima(dif_den, seasonal=TRUE, trace=TRUE, stationary
    =TRUE) # ARIMA(0,0,1)(2,0,0)[4] with zero mean
m1 <- arima(x = denmark, order = c(0,1,1), seasonal = list(order =
    c(2, 0, 0), period = 4))
```

This function uses the "Akaike Information Criterion" (AIC) to evaluate the performance of considered models and suggests the model with the lowest calculated value. Since we are using quarterly data, the function was set to evaluate seasonal models. Considering that we have differenced the data to cause stationarity, this was also inputted in the function. The result is a suggested Seasonal ARIMA $(0, 0, 1) \times (2, 0, 0)_{[4]}$ model with zero mean.

It is important to note that the "auto.arima" function was used on already differenced data. In order to reflect this at the final model, a Seasonal ARIMA $(0, 1, 1) \times (2, 0, 0)_{[4]}$ model is selected. The variance of the error term, σ^2, is estimated to be 1752. The log-likelihood and AIC are -624.24 and 1256.49, respectively. To sum up, the estimated values of the model coefficients with their respective standard error are presented in Table 1:

Table 1: Chosen ARIMA model coefficients

Coefficients	ma1	sar1	sar2
Estimated value	0.4236	0.3999	0.1557
Standard error	0.0946	0.0931	0.0977

Question 4.4: Residual diagnostics

Checking the residuals derived from ARIMA model predictions for white noise is crucial, as they determine the accuracy and reliability of the model. Non-white noise residuals indicate that the model has not captured all underlying patterns in the data or has violated one or more assumptions as stationarity, independence, or normality. In general, examining residuals for patterns and deviations from expected distribution helps identify potential problems in the model structure. From Figures 8 and 9 it can be observed that there is no significant autocorrelation left in the residuals, thus fulfilling the assumption of residual independence.

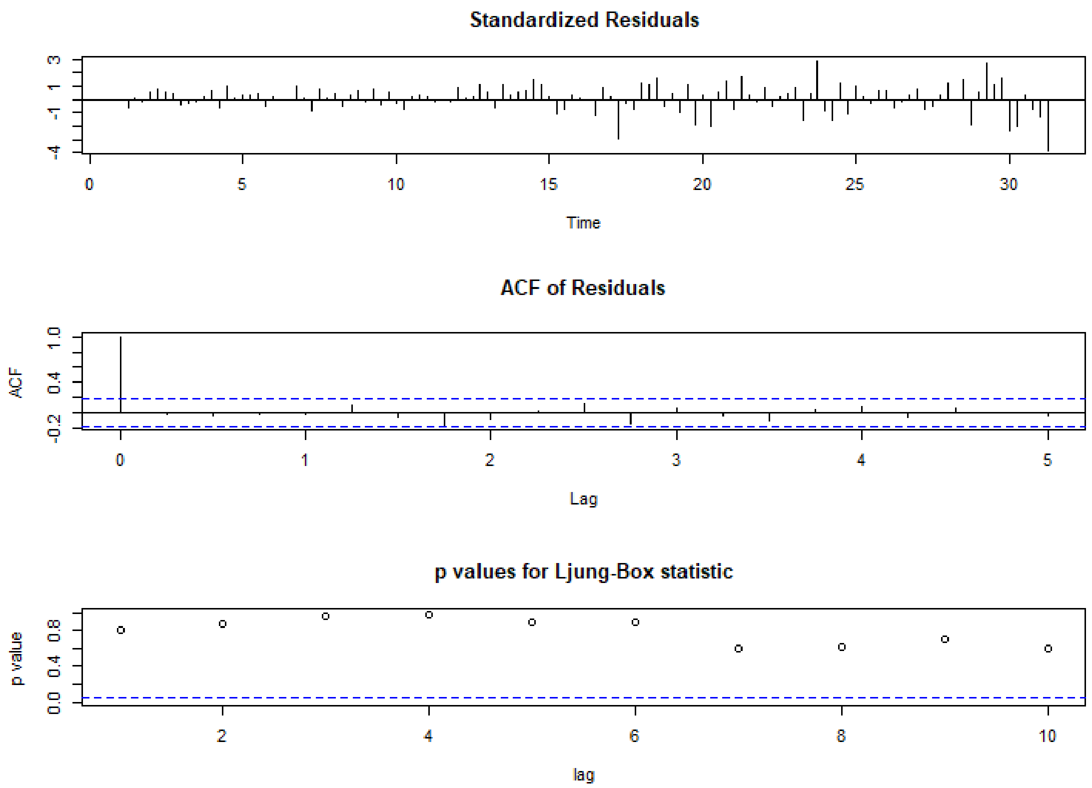

Figure 8: Time series diagnostics for the model predictions

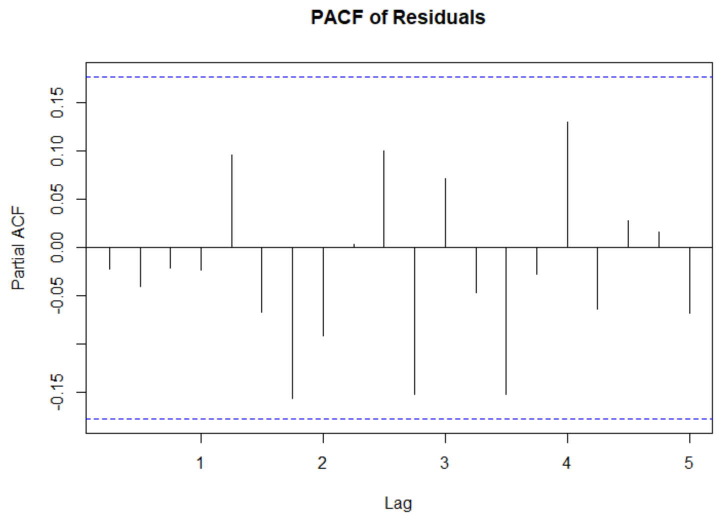

Figure 9: Partial autocorrelation function of the residuals

The purpose of the plot illustrated in Figure 10 is to assess whether the residuals have any significant cyclic patterns or any non-randomness, which could indicate that the model is not adequately capturing all the patterns in the data. Therefore, we can check if the residuals are white noise. If there were any significant peaks or patterns in the plot, that would suggest that the residuals have some structure left, and the model is not a good fit for the data. However, in our case everything seems fine.

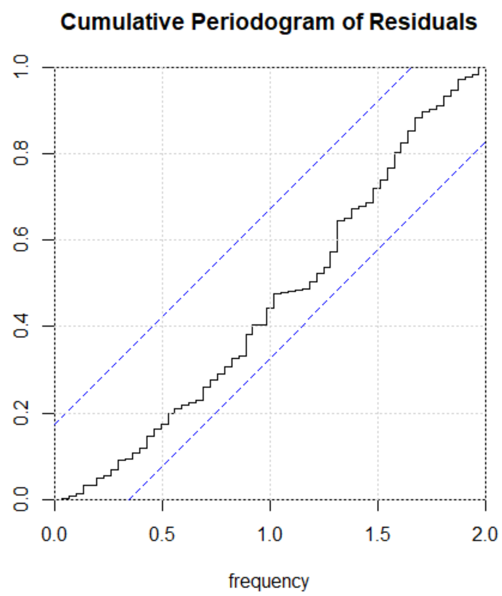

Figure 10: Cumulative periodogram of the residuals

The Q-Q plot presented in Figure 11 is an effective diagnostic tool for verifying the normality assumption of the residuals. The expected pattern of the points on the plot is a straight line when the residuals are normally distributed. However, significant deviations from a straight line may indicate a violation of the normality assumption, which requires further investigation or model modification. It can be concluded that the normality assumption of the residuals is satisfied to a satisfactory extent, with the presence of some potential outliers, such as observations 66 and 122.

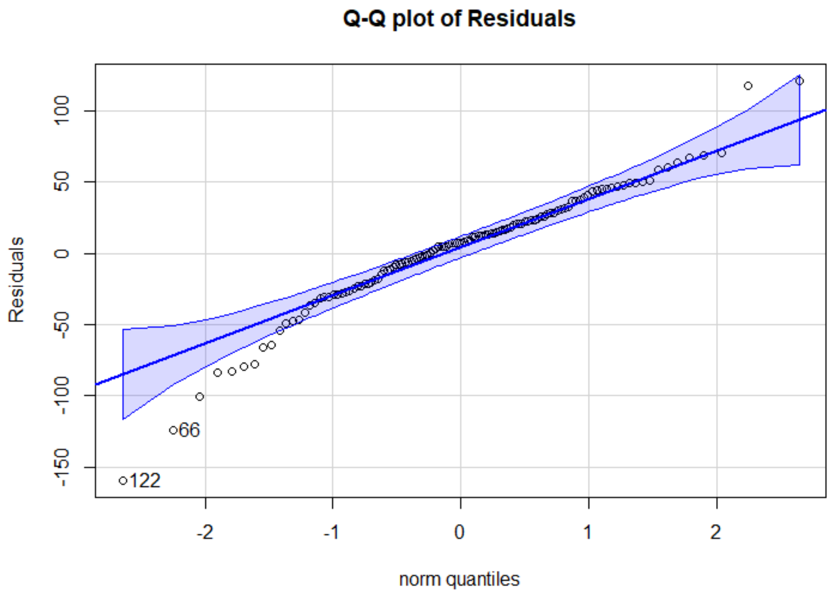

Figure 11: Q-Q plot of the residuals

Question 4.5: Forecasting the future house prices - I

Based on the selected ARIMA model, predictions for six time steps beyond the training data are generated. More specifically, in Table 2 are included the predicted average house prices with their respective 95% prediction intervals. Additionally, they are illustrated in Figure 12. It shall be stated that these predictions do not take any external inputs into consideration. Overall, it can be observed that the predicted values are lower than the previous three years of observed house prices.

Table 2: Predictions for six time steps beyond the training data, until 2024 Q1

Time step (quarter)	**2022 Q4**	**2023 Q1**	**2023 Q2**	**2023 Q3**	**2023 Q4**	**2024 Q1**
Lower 95% p.i.	2250.176	2231.910	2172.900	2065.483	2030.057	2011.598
House price (DKK/m2)	2292.038	2304.740	2267.006	2176.873	2165.096	2170.331
Upper 95% p.i.	2333.900	2377.570	2361.112	2288.263	2300.135	2329.064

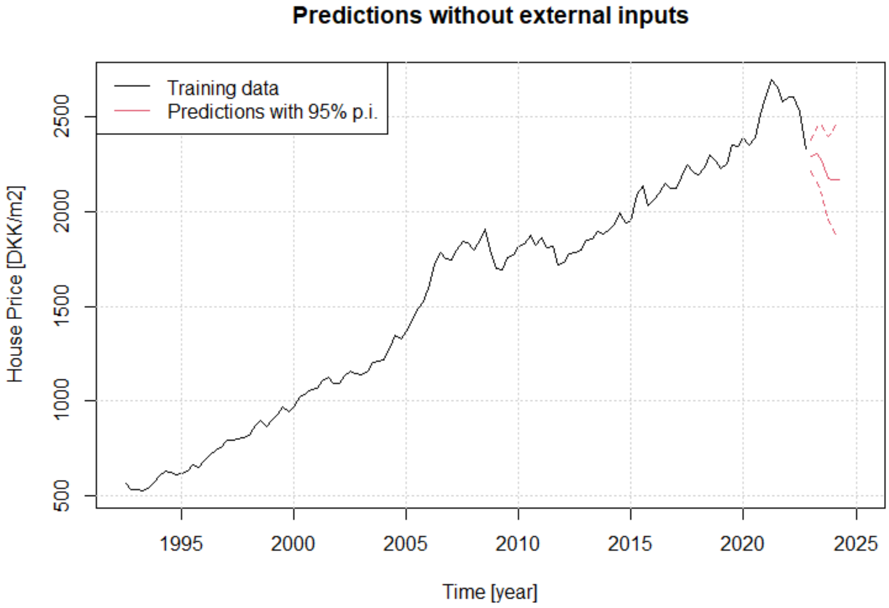

Figure 12: Forecasting of the future house prices without external inputs

Question 4.6: External inputs

At this section it is examined whether the information we have about external inputs can improve the chosen model's prediction performance or not. In order to extend the ARIMA model to an ARIMAX both interest and inflation rate are used as external regressors. The following lines of code are applied in "RStudio":

Listing 2: External inputs

```
externalreg <- cbind(cumsum(interest[1:122]), cumsum(inflation[1:122])) # cumsum cf week 7 (diff(cumsum()) = identity)
m2 <- arima(x = denmark, order = c(0,1,1), seasonal = list(order = c(2, 0, 0), period = 4), xreg = externalreg)
(m2)
```

The estimated coefficients with their respective standard errors are presented in Table 3. The variance of the error term, σ^2, is estimated to be 1540. The log-likelihood and Akaike's Information Criterion (AIC) are -617.07 and 1246.13, respectively. A lower value of AIC for the ARIMAX model compared to the value of 1256.49 for the ARIMA model indicates that the ARIMAX model is a better fit to the data. Hence, it can be concluded that the increase in model complexity by including the external regressors is justified by the improvement in the quality of future house price predictions achieved by the ARIMAX model.

Table 3: ARIMAX model coefficients

Coefficients	ma1	sar1	sar2	externalreg1	externalreg2
Estimated value	0.3904	0.4143	0.3069	0.7034	-15.040
Standard error	0.1010	0.0893	0.0988	3.7106	3.9041

Question 4.7: Forecasting the future house prices - II

The updated ARIMAX model is used to make new predictions for six time steps beyond the training data. The predicted average house prices along with their 95% predictions intervals are presented in Table 4 and plotted in Figure 13. The forecasted prices are notably lower than those predicted by the model without the external regressors, with the decreasing trend being even steeper than the one the ARIMA model predicted.

Table 4: Predictions for six time steps beyond the training data, until 2024 Q1

Time step (quarter)	2022 Q4	2023 Q1	2023 Q2	2023 Q3	2023 Q4	2024 Q1
Lower 95% p.i.	2184.581	2092.504	1958.984	1783.996	1677.703	1581.383
House price (DKK/m2)	2223.818	2159.705	2045.543	1886.313	1802.134	1727.795
Upper 95% p.i.	2263.055	2226.906	2132.102	1988.630	1926.565	1874.207

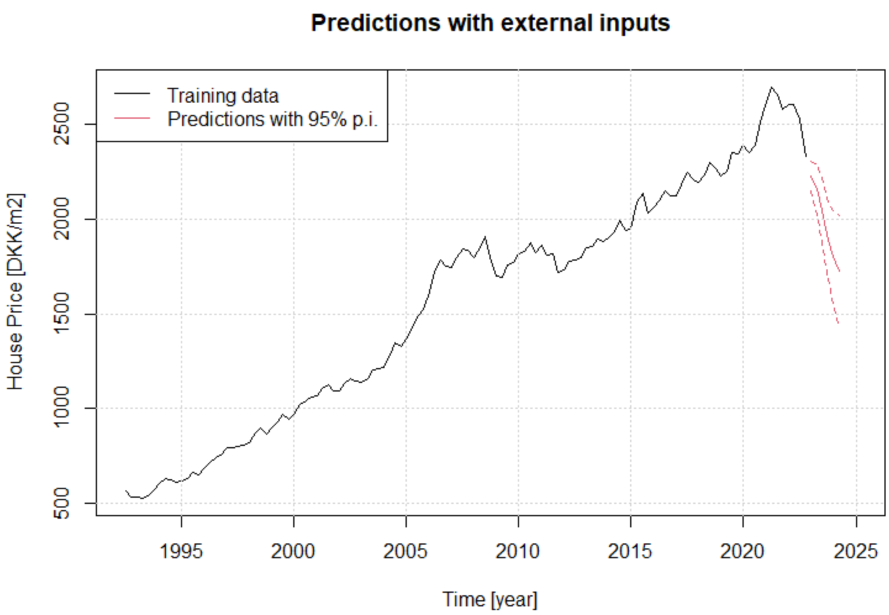

Figure 13: Forecasting of the future house prices with external inputs

Question 4.8: Conclusions - I

The ARIMAX model developed above can be considered well-informed regarding the general trend in house prices in Denmark, taking into account historical data in each season of the past thirty years along with inflation and interest rates. Overall, both the ARIMA and ARIMAX models predict a decrease in the coming yearly quarters, meaning that a drop in house prices should be expected. This leads to the conclusion that this is not an economically wise moment to buy a house in Denmark, since it is possible that in the not-too-distant future the same house could be bought at a lower price. Especially, ARIMAX predicts a monotonic decrease on the average house sales prices the in the near future quarters.

However, both models have a common limitation/weakness, which is that they are trained on average observations about Denmark in general. Consequently, characteristics like the hedonistic benefits of houses closer to the sea or transportation benefits from houses in the capital region etc. that could lead to different resilience in house prices among different regions of Denmark are neglected. Thus, the spatial variability of the way prices are affected in periods of economic growth or imbalance is flattened out and homogenized by training the model only in one time series of data. The way to deal with this would be to train a different ARIMAX model for each different region time series individually, but that increases automatically both the computational time and the cost exponentially the more the regions and the more detailed the analysis.

Another limitation that has to be taken into consideration are the different synergistic or antagonistic effects that could arise between different regions on a dynamic future environment which again could change the characteristics of some of the time series, but not all of them. Hence, a model that could examine these interactions simultaneously between the different regions that affect the house pricing system in Denmark potentially could lead to more robust predictions about the future prices and generate an economically wiser investment decision, by taking into consideration both the seasonal and spatial variability.

Part 2 - Multivariate models

In the second part of this assignment, a ”Multiple Autoregressive Integrated Moving Average with Exogenous Variables” (MARIMAX) model is developed to predict the value of house prices in the four regions of Denmark simultaneously. The goal is to make future predictions for the entire dataset until the first quarter of 2024. The MARIMAX model allows us to incorporate multiple time series and exogenous variables, such as inflation and interest rates, simultaneously into a single model for more accurate and reliable predictions. Furthermore, the results of the MARIMAX model are evaluated using various diagnostic tests, such as residual analysis and model selection criteria, to ensure the accuracy and reliability of the model. Finally, based on the predicted house prices, arguments are presented regarding the best economic decision for the purchase of a house in Denmark. The analysis takes into account the predicted future trends in house prices and the potential risks and benefits of purchasing a house at different times and regions.

Question 4.9: Re-presenting the data

The quarterly average sales prices for each of the four regions of Denmark are illustrated in Figure 14. Similarly as before, it is evident that there is an upward trend in the data and non-stationary time series have statistical properties that change over time making it difficult to build a reliable model. Therefore, differencing was applied to eliminate the upward trend as illustrated in Figure 15. Another useful preliminary insight would be that capital region house prices fluctuate much more in periods of economic instability like around 2008, while the other regions seem to be more resilient and behave similarly. However, the capital region has the highest value increase over time. Lastly, it is pretty clear that there is periodicity throughout the most years, which suggests that house prices are season dependent.

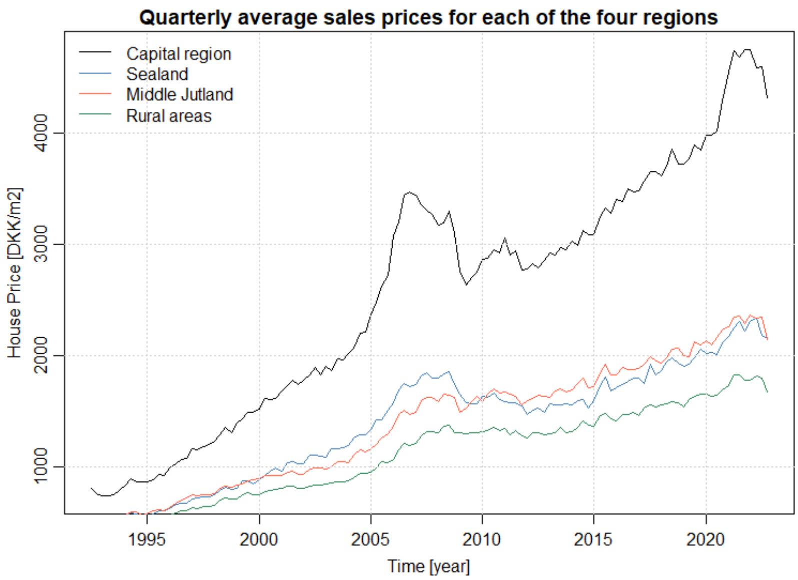

Figure 14: Quarterly average sales prices for each of the four regions in Denmark (raw data)

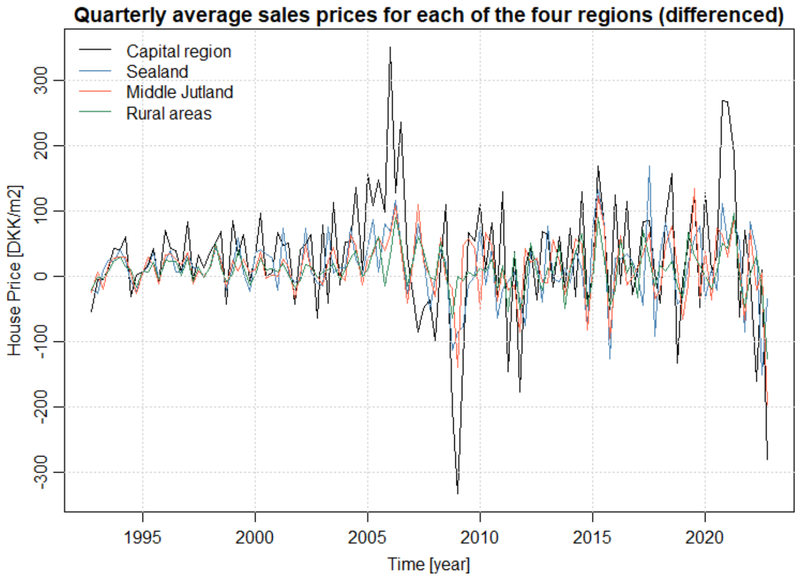

Figure 15: Quarterly average sales prices for each of the four regions in Denmark (diff. data)

Question 4.10: ACF and PACF (multivariate)

The multivariate ACF plots for the prices in the four regions of Denmark are illustrated in Figure 16. Since the four processes are non-stationary, the decay in the ACF plots of the raw data goes very slowly to zero. Thus, it would be meaningful to check the multivariate ACF and PACF plots of the differenced input time series data.

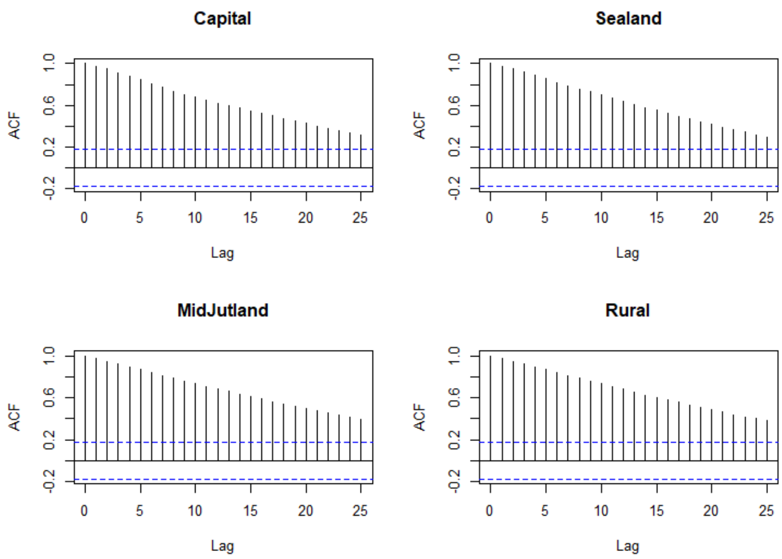

Figure 16: Multivariate ACF for the prices in the four regions (raw data)

After differencing the raw data, the multivariate ACF and PACF plots for the prices in the four regions of Denmark are illustrated in Figures 17 and 18 respectively:

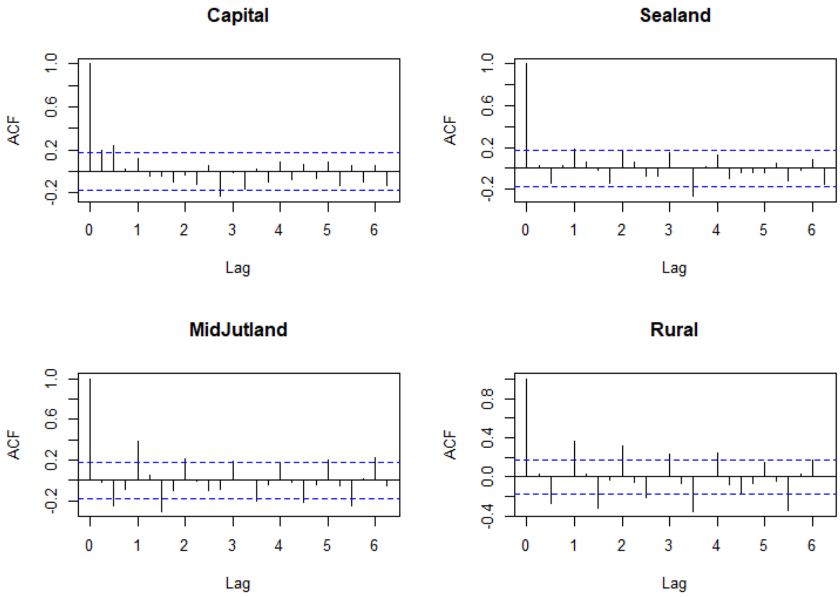

Figure 17: Multivariate ACF for the prices in the four regions (differenced data)

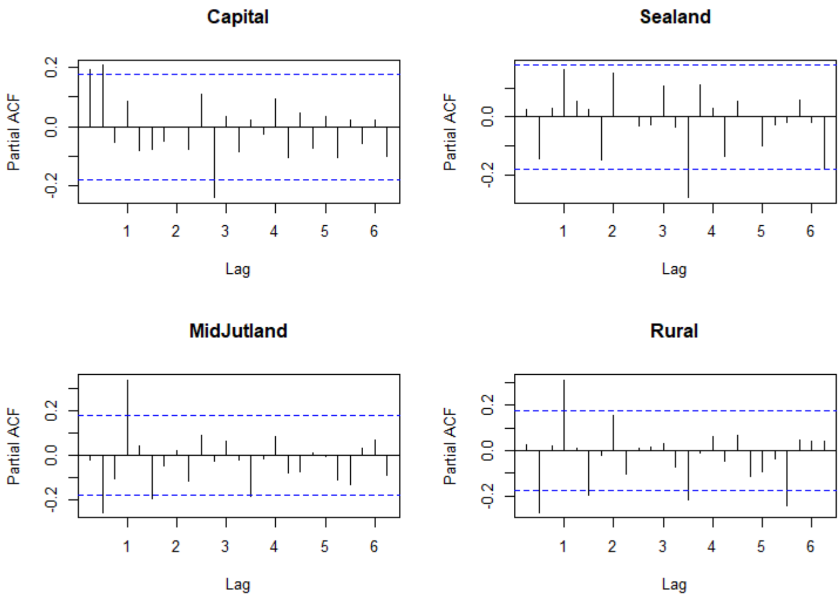

Figure 18: Multivariate PACF for the prices in the four regions (differenced data)

Regarding the ACF plots it can be observed that there is no exponential decay and from ”Mid-Jutland” and ”Rural” plots a seasonality every two quarters is implied. Regarding the PACF plots we could say that there are two significant lags for ”Capital” but only the second lag for the other regions in the first period. However, again a seasonality ever every two quarters in ”Sealand”, ”MidJutland” and ”Rural” is evident. All in all, it is pretty difficult to make an accurate argument for a potential model structure, but a first approach could be a Seasonal MARIMAX model of the structure $(0, 0, 1) \times (1, 0, 0)_{[4]}$ applied on the differenced data.

Question 4.11: Multivariate model selection

In order to find a suitable MARIMAX model for house prices in Denmark, the ”marima” package in ”RStudio” was used. This package takes as input time series of more than one variables, as well as external regressors, which in our case are interest rates and inflation. The external regressors can either be considered as simple regression variables or as independent variables.

Model 1: The first model that was tested was a simple MARIMAX $(0, 0, 1)$ applied on the differenced data, with interest rate and inflation rate used as simple regression variables. The resulting log determinant plot showed that the residual covariance matrix converges smoothly for a small number of iterations but when plotting the ACF and PACF of the residuals a periodicity of 4 lags was observed in the ACF plots, indicating seasonality that was not taken into account in the model.

Model 2: The second model that was tested was a Seasonal MARIMAX $(0, 0, 1) \times (1, 0, 0)_{[4]}$ applied on the differenced data, with interest rate and inflation rate used as simple regression variables. The resulting log determinant plot again showed good convergence for a small number of iterations, and a significant improvement was observed at the ACF plots of the residuals, however there was still some autocorrelation structure left in the residual that was not explained mostly for the first two series.

Model 3: The third model that was tested was a Seasonal MARIMAX $(0, 0, 1) \times (2, 0, 0)_{[4]}$ applied on the differenced data, with interest rate and inflation rate used as simple regression variables. The resulting log determinant plot illustrated in Figure 19 showed good convergence, by using ”penalty = 2” in order to incorporate the AIC in our analysis. The ACF plots of the residuals were similar to those of Model 2. In this case however the one-step predictions seemed to be closer to the training data decreasing model's MSE, but the estimated model coefficients indicated that the house prices of the four regions were not dependent on the external inputs of interest and inflation rate which is not something that we expected and could not be trusted.

Model 4: The fourth model that was investigated had the same structure as Model 3, but with the interest rate and the inflation rate used as independent variables. The resulting log determinant plot converged a bit slower, as illustrated in Figure 20. Afterwards, the ”step.slow” function was used to incorporate the step function of AIC in model selection, but with a smoother way. The resulting model coefficients indicated the expected dependence of some time series with the external inputs, something that was expected. Thus, we believe that interest and inflation rate should be used as independent variables.

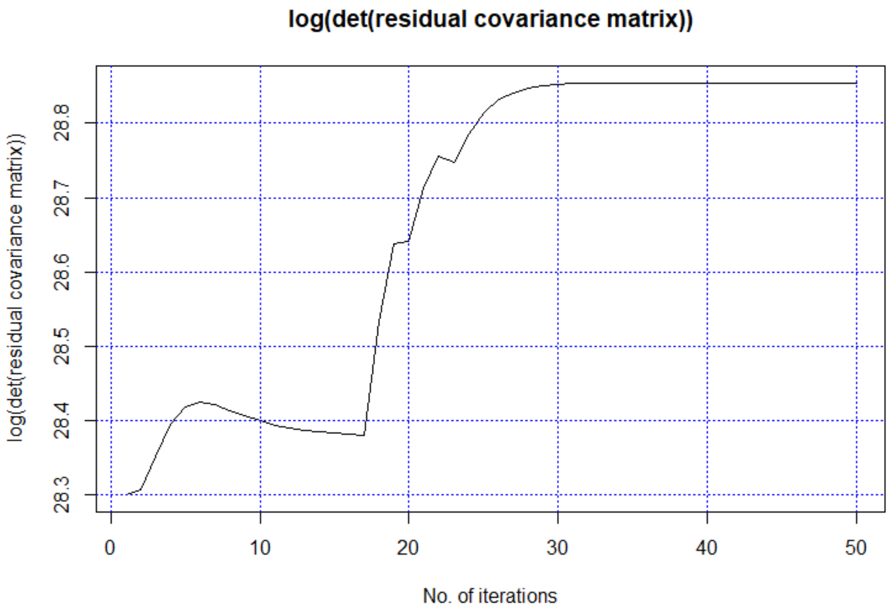

Figure 19: Model 3 - Convergence of the logarithmic determinant of the residual covariance matrix

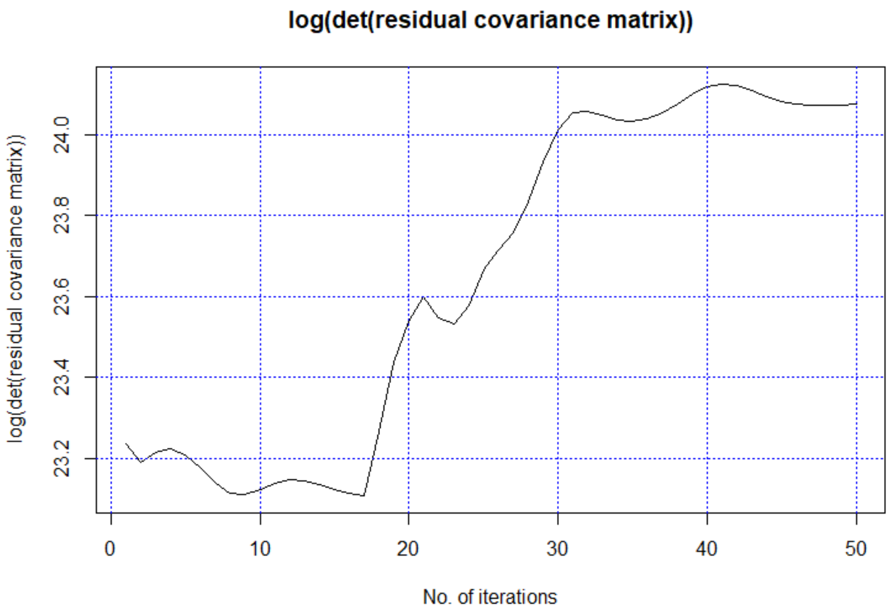

Figure 20: Model 4 - Slower convergence of the logarithmic determinant of the residual covariance matrix

Question 4.12: Multivariate explanation

Model 4 is the one that was finally chosen and used to generate future predictions. As mentioned the structure is a Seasonal MARIMAX $(0, 0, 1) \times (2, 0, 0)_{[4]}$ model applied on the differenced data. The respective univariate structure would be a Seasonal ARIMAX $(0, 1, 1) \times (2, 0, 0)_{[4]}$ model applied on the raw data. The difference is that the multivariate MARIMAX model is trained simultaneously on four different time series data, which depict the evolution of house prices in the four regions and examining at the same time all potential interactions as depicted in Listing 3. Finally, some residual diagnostics plots regarding the multivariate model are included in the Appendix.

Regarding the structure illustrated in the Listing, the AR component is depicted until lag=8 as recognized by the model in order to capture the seasonality of the model structure and MA at lag=1 refers to non-seasonal MA1 model component. Furthermore, the respective f-values assess the significance of the estimated coefficients and values higher than 2 indicate significant coefficients. Additionally, the way to read the Listing is to have in mind that the rows $y1 - y6$ refer to the four regions plus the two rates and the columns the same. Thus, the table elements indicate the dependencies between them.

Briefly, we observe significant dependencies between different regions at lag=4 and lag=8 for the AR component. It is noteworthy that at lag=8 the rural areas showed significant dependencies with all the other regions. Regarding the non seasonal MA component we observe that there are significant dependencies between the capital region with all the other regions and similarly for Sealand. Lastly, it should be noted that there is significant dependency on the MA component between the capital region and the interest rate, as well as between Middle Jutland and the rural areas with the inflation rate.

Listing 3: Estimated parameters of Model 4

```
AR estimates:
, , Lag=1

    x1=y1 x2=y2 x3=y3 x4=y4   x5=y5   x6=y6
y1      0     0     0     0  0.0000  0.0000
y2      0     0     0     0  0.0000  0.0000
y3      0     0     0     0  0.0000  0.0000
y4      0     0     0     0  0.0000  0.0000
y5      0     0     0     0 -1.1644  0.0000
y6      0     0     0     0  0.0000 -0.8509

, , Lag=2

    x1=y1 x2=y2 x3=y3 x4=y4 x5=y5   x6=y6
y1      0     0     0     0     0  0.0000
y2      0     0     0     0     0  0.0000
```

```
y3     0      0      0      0      0   0.0000
y4     0      0      0      0      0   0.0000
y5     0      0      0      0      0   0.0000
y6     0      0      0      0      0  -0.4311

, , Lag=3

   x1=y1 x2=y2 x3=y3 x4=y4  x5=y5 x6=y6
y1     0     0     0     0 0.0000     0
y2     0     0     0     0 0.0000     0
y3     0     0     0     0 0.0000     0
y4     0     0     0     0 0.0000     0
y5     0     0     0     0 0.2604     0
y6     0     0     0     0 0.0000     0

, , Lag=4

   x1=y1   x2=y2   x3=y3   x4=y4 x5=y5  x6=y6
y1     0 -0.5403  0.0000  0.0000     0 0.0000
y2     0  0.0000  0.0000 -0.3882     0 0.0000
y3     0  0.0000 -0.3489  0.0000     0 0.0000
y4     0 -0.1706  0.0000  0.0000     0 0.0000
y5     0  0.0000  0.0000  0.0000     0 0.0000
y6     0  0.0000  0.0000  0.0000     0 0.4564

, , Lag=7

   x1=y1 x2=y2 x3=y3 x4=y4   x5=y5 x6=y6
y1     0     0     0     0  0.0000     0
y2     0     0     0     0  0.0000     0
y3     0     0     0     0  0.0000     0
y4     0     0     0     0  0.0000     0
y5     0     0     0     0 -0.2213     0
y6     0     0     0     0  0.0000     0

, , Lag=8

   x1=y1 x2=y2  x3=y3   x4=y4  x5=y5 x6=y6
y1     0     0 0.4843 -0.5894 0.0000     0
y2     0     0 0.0000 -0.3729 0.0000     0
y3     0     0 0.0000 -0.3481 0.0000     0
y4     0     0 0.0000 -0.4596 0.0000     0
y5     0     0 0.0000  0.0000 0.1575     0
y6     0     0 0.0000  0.0000 0.0000     0
```

```

AR f-values (squared t-values):
, , Lag=1

  x1=y1 x2=y2 x3=y3 x4=y4    x5=y5   x6=y6
y1     0     0     0     0   0.0000  0.0000
y2     0     0     0     0   0.0000  0.0000
y3     0     0     0     0   0.0000  0.0000
y4     0     0     0     0   0.0000  0.0000
y5     0     0     0     0 764.3933  0.0000
y6     0     0     0     0   0.0000 17.1036

, , Lag=2

  x1=y1 x2=y2 x3=y3 x4=y4 x5=y5  x6=y6
y1     0     0     0     0     0 0.0000
y2     0     0     0     0     0 0.0000
y3     0     0     0     0     0 0.0000
y4     0     0     0     0     0 0.0000
y5     0     0     0     0     0 0.0000
y6     0     0     0     0     0 2.5581

, , Lag=3

  x1=y1 x2=y2 x3=y3 x4=y4   x5=y5 x6=y6
y1     0     0     0     0  0.0000     0
y2     0     0     0     0  0.0000     0
y3     0     0     0     0  0.0000     0
y4     0     0     0     0  0.0000     0
y5     0     0     0     0 28.4703     0
y6     0     0     0     0  0.0000     0

, , Lag=4

  x1=y1   x2=y2   x3=y3  x4=y4 x5=y5   x6=y6
y1     0  9.1644  0.0000 0.0000     0  0.0000
y2     0  0.0000  0.0000 6.7445     0  0.0000
y3     0  0.0000 14.7488 0.0000     0  0.0000
y4     0 10.8223  0.0000 0.0000     0  0.0000
y5     0  0.0000  0.0000 0.0000     0  0.0000
y6     0  0.0000  0.0000 0.0000     0 20.8949

, , Lag=7

```

```
      x1=y1  x2=y2  x3=y3  x4=y4     x5=y5  x6=y6
y1        0      0      0      0    0.0000      0
y2        0      0      0      0    0.0000      0
y3        0      0      0      0    0.0000      0
y4        0      0      0      0    0.0000      0
y5        0      0      0      0   13.0641      0
y6        0      0      0      0    0.0000      0

, , Lag=8

      x1=y1  x2=y2   x3=y3     x4=y4   x5=y5  x6=y6
y1        0      0  3.2719    2.1429  0.0000      0
y2        0      0  0.0000    5.5683  0.0000      0
y3        0      0  0.0000    6.3263  0.0000      0
y4        0      0  0.0000   26.8722  0.0000      0
y5        0      0  0.0000    0.0000  8.7939      0
y6        0      0  0.0000    0.0000  0.0000      0

MA estimates:
, , Lag=1

       x1=y1     x2=y2     x3=y3     x4=y4      x5=y5      x6=y6
y1  0.0000    0.6543    0.0000    0.0000   -68.7589     0.0000
y2  0.3321   -0.3061    0.0000    0.0000     0.0000     0.0000
y3  0.1014    0.2941   -0.4484    0.0000     0.0000   -28.2702
y4  0.0685    0.1810    0.2054   -0.6362     0.0000   -19.0049
y5  0.0000    0.0000    0.0000    0.0000     0.6260     0.0000
y6  0.0000    0.0000    0.0000    0.0000     0.0000     0.6756

MA f-values (squared t-values):
, , Lag=1

      x1=y1  x2=y2  x3=y3  x4=y4  x5=y5  x6=y6
y1     0.00  10.33   0.00    0.0   3.00   0.00
y2    37.61   7.55   0.00    0.0   0.00   0.00
y3     3.91   7.57  10.78    0.0   0.00   9.87
y4     3.97   6.03   4.02   20.7   0.00   9.82
y5     0.00   0.00   0.00    0.0  37.17   0.00
y6     0.00   0.00   0.00    0.0   0.00   9.41
```

Question 4.13: Forecasting the future house prices - III

The future house prices until 2024 Q1 are presented in Tables 5, 6, 7 and 8. Additioanlly, the plots for each reaspective region are illustrated in Figures 21, 22, 23 and 24. A general comment would be that a small decrease is expected between the respective quarters, for example 2024 Q1 in comparison to 2023 Q1, in all regions. Another general observation is the further we aim to forecast in the future the broader the range of the prediction intervals, which means the more uncertainty is existent. Finally, in all regions when examining only 2023 quarter 1 seems to be the most expensive period to buy a house except Middle Jutland, for which quarter 2 is the most expensive period.

Table 5: Capital region - Predictions until 2024 Q1

Time step (quarter)	**2022 Q4**	**2023 Q1**	**2023 Q2**	**2023 Q3**	**2023 Q4**	**2024 Q1**
Lower 95% p.i.	4167.018	4127.787	3992.334	3942.829	3851.444	3865.707
House price (DKK/m2)	4340.841	4396.021	4329.505	4337.062	4311.437	4392.120
Upper 95% p.i.	4514.664	4664.254	4666.676	4731.295	4771.429	4918.533

Table 6: Sealand - Predictions until 2024 Q1

Time step (quarter)	**2022 Q4**	**2023 Q1**	**2023 Q2**	**2023 Q3**	**2023 Q4**	**2024 Q1**
Lower 95% p.i.	2039.564	2046.966	2011.276	1922.378	1914.694	1923.737
House price (DKK/m2)	2123.761	2177.717	2175.904	2115.015	2137.520	2175.830
Upper 95% p.i.	2207.959	2308.468	2340.532	2307.652	2360.345	2427.923

Table 7: Middle Jutland - Predictions until 2024 Q1

Time step (quarter)	**2022 Q4**	**2023 Q1**	**2023 Q2**	**2023 Q3**	**2023 Q4**	**2024 Q1**
Lower 95% p.i.	2122.219	2122.456	2105.411	2005.076	1999.132	1999.990
House price (DKK/m2)	2194.960	2225.937	2232.395	2151.847	2174.510	2201.517
Upper 95% p.i.	2267.701	2329.417	2359.379	2298.618	2349.888	2403.044

Table 8: Rural areas - Predictions until 2024 Q1

Time step (quarter)	**2022 Q4**	**2023 Q1**	**2023 Q2**	**2023 Q3**	**2023 Q4**	**2024 Q1**
Lower 95% p.i.	1664.422	1695.091	1654.632	1615.523	1596.691	1605.957
House price (DKK/m2)	1712.812	1765.962	1742.403	1717.429	1716.877	1743.615
Upper 95% p.i.	1761.201	1836.832	1830.173	1819.334	1837.063	1881.273

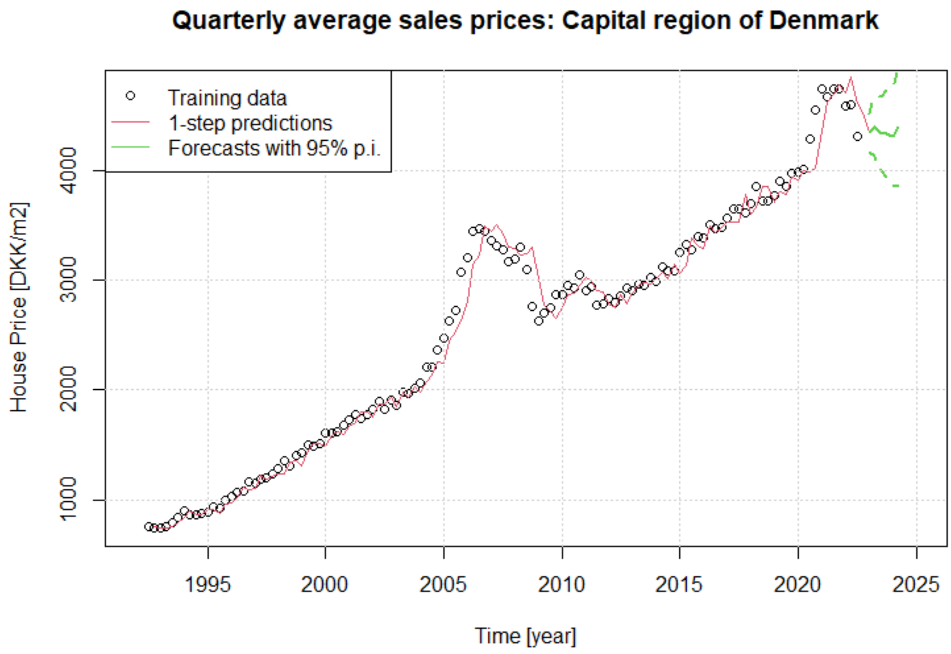

Figure 21: Quarterly average sales prices in the capital region of Denmark

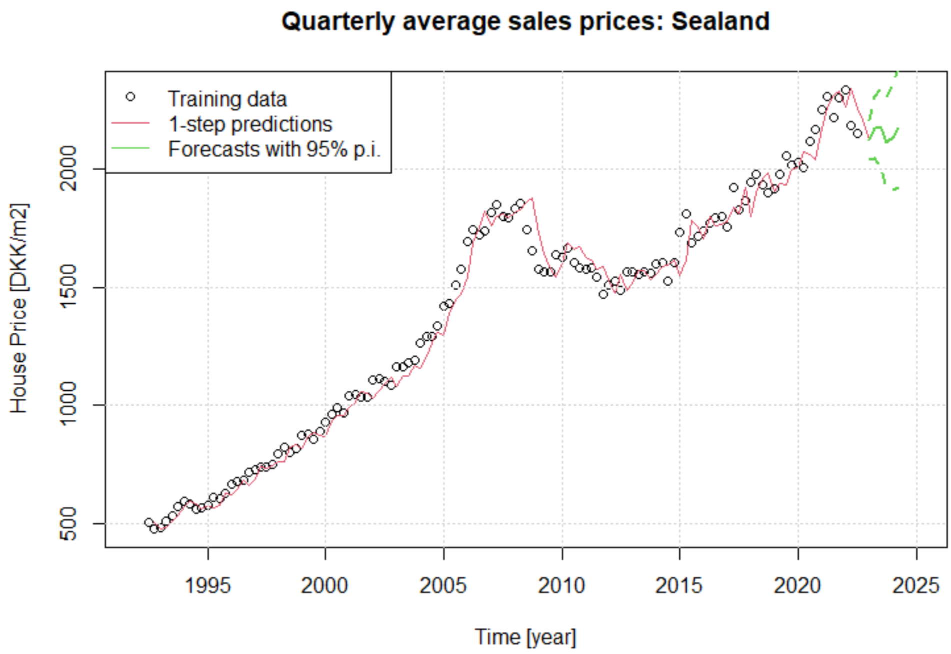

Figure 22: Quarterly average sales prices in Sealand

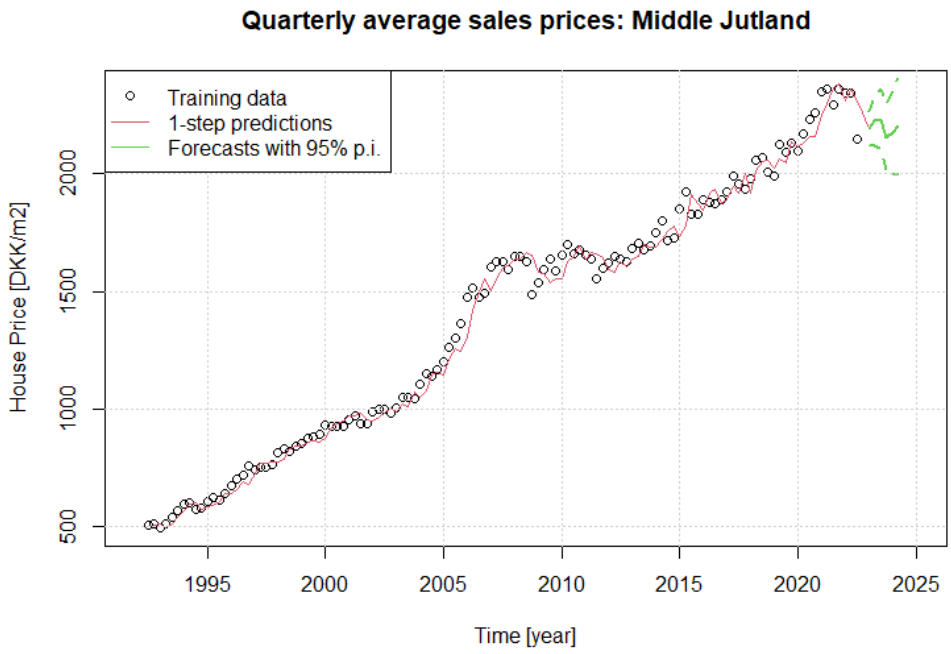

Figure 23: Quarterly average sales prices in Middle Jutland

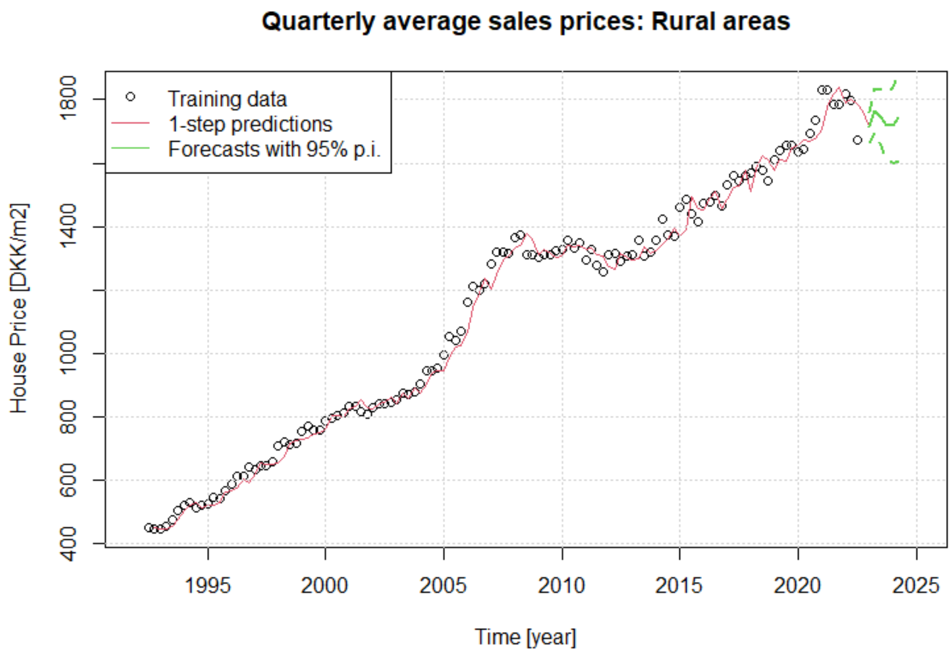

Figure 24: Quarterly average sales prices in the rural areas

Question 4.14: Conclusions - II

To sum up, it can be stated that the MARIMAX model predicts more stable conditions in the market of house prices in comparison to the ARIMAX model, which predicted a large decrease in the average house sales prices in the near future for the whole Denmark. Relatively speaking, we would trust more the predictions of the MARIMAX model as it is a region oriented model and more internal factors that determine the different region time series data behaviour are taken into consideration. Furthermore, the derived model coefficients showed also the relationship for specific time series with external inputs enhancing the initial intuition that there should be spatial variability regarding the way different regions are affected by the external inputs and their respective house price resilience. Lastly, the initial intuition that there is seasonality in house prices during the year is also captured by the model and depicted in the forecasts.

To give a more specific proposal regarding a wise economic investment in relation to the timing of buying a house the fourth quarter of 2023 is recommended for the capital region and the rural areas and the third quarter of 2023 for Sealand and Middle Jutland. Regarding the region, the chosen model showed that buying a house in Sealand will not be affected by the near future inflation and interest rate, thus leading to a safer investment. On the other hand, the forecasted decreasing house price in the capital region of Denmark can be a potential opportunity to buy at low price for a long term investment that in the future can generate bigger profit, but that would require higher budget in comparison to the other regions. So, it is always a trade-off between risk and potential profit.

In conclusion, it should be stated that in no case the constructed model is perfect and of course there are limitations, as well as the additional uncertainty that is by default generated by the unknown future conditions.

Code

Listing 4: Part 1 - Univariate models (initialization)

```
library(tseries)
library(forecast)
library(car)
library(aTSA)

data <- read.csv("A3Data.csv")

# Initialize an empty list to store the results
ts_list <- list()

#Convert the data to quarterly Time series
ts <- ts(data, frequency = 4)

# Assign the values to separate variables
time <- seq(1992.5,2024.25,0.25)
denmark <- na.omit(ts[,2])
capital <- na.omit(ts[,3])
sealand <- na.omit(ts[,4])
midjutland <- na.omit(ts[,5])
rural <- na.omit(ts[,6])
interest <- na.omit(ts[,7])
inflation <- na.omit(ts[,8])
```

Listing 5: Question 4.1 - Presenting the data

```
# plots
par(mfrow = c(1,1))

# plot the quarterly average sale prices
plot(time[1:122],denmark, type='l',
     main="Quarterly average sales prices in Denmark",
     xlab='Time [year]' ,ylab="House Price [DKK/m2]")
grid()
abline(h=mean(denmark), lty=2, col="#8B0000")
legend("topleft", legend="Average sales prices", col=1, lty=1, bty
  ="n", cex=0.8)

# plot the interest rate
plot(time[1:124],interest[1:124], type='l',
     main="Interest Rate",
```

```
     xlab='Time [year]', ylab="Interest Rate [%]")
grid()
abline(h=mean(interest), lty=2, col="#8B0000")
legend("topright", legend="Interest Rate", col=1, lty=1, bty="n",
   cex=0.8)

# plot the inflation rate
plot(time[1:124],inflation[1:124], type='l',
     main="Inflation Rate",
     xlab='Time [year]', ylab="Inflation Rate [%]")
grid()
abline(h=mean(inflation), lty=2, col="#8B0000")
legend("topleft", legend="Inflation Rate", col=1, lty=1, bty="n",
   cex=0.8)
```

Listing 6: Question 4.2 - ACF and PACF

```
# ACF and PACF of the house prices
par(mfrow = c(1,1))
acf(denmark, main = "Autocorrelation funtion of the raw time
   series data")
pacf(denmark, main = "Partial autocorrelation fucntion of the raw
   time series data")

# Differencing
dif_den <- diff(denmark)
dif_cap <- diff(capital)
dif_seal <- diff(sealand)
dif_midj <- diff(midjutland)
dif_rul <- diff(rural)

# ACF and PACF for the transformed data
par(mfrow = c(1,1))
acf(dif_den, main = "Autocorrelation funtion of the differenced
   time series data")
pacf(dif_den, main = "Partial autocorrelation fucntion of the
   differenced time series data")

# cross-correlation function between the four time series of the
   house prices.
par(mfrow = c(2,3)) # raw series
ccf(capital, sealand, main = "CCF: 'Capital' & 'Sealand' (raw)")
ccf(capital, midjutland, main = "CCF: 'Capital' & 'MidJutland' (
   raw)")
```

```
ccf(capital, rural, main= "CCF: 'Capital' & 'Rural' (raw)")
ccf(sealand, midjutland, main = "CCF: 'Sealand' & 'MidJutland' (raw)")
ccf(sealand, rural, main= "CCF: 'Sealand' & 'Rural' (raw)")
ccf(midjutland, rural, main = "CCF: 'MidJutland' & 'Rural' (raw)")

# Pre-whitening
# Define the ARIMA model for Denmark data
arima_model = arima(denmark, order = c(0, 1, 1), seasonal = list(order = c(2, 0, 0), period = 4))

# Prewhiten the Denmark data
pw_den <- residuals(arima_model)

# Prewhiten the capital data
pw_cap <- residuals(update(arima_model, x = capital))

# Prewhiten the sealand data
pw_seal <- residuals(update(arima_model, x = sealand))

# Prewhiten the midjutland data
pw_midj <- residuals(update(arima_model, x = midjutland))

# Prewhiten the rural data
pw_rul <- residuals(update(arima_model, x = rural))

# Estimate the SCCF for the filtered series
acf(na.pass(cbind(pw_den, pw_cap, pw_seal, pw_midj, pw_rul)), type = "correlation")

# Different plot style
par(mfrow=c(2,2))
ccf(pw_den, pw_cap, lag.max = 20, type = "correlation", main = "CCF: 'Denmark' & 'Capital'")
ccf(pw_den, pw_seal, lag.max = 20, type = "correlation", main = "CCF: 'Denmark' & 'Sealand'")
ccf(pw_den, pw_midj, lag.max = 20, type = "correlation", main = "CCF: 'Denmark' & 'MidJutland'")
ccf(pw_den, pw_rul, lag.max = 20, type = "correlation", main = "CCF: 'Denmark' & 'Rural'")

par(mfrow=c(2,3))
ccf(pw_cap, pw_seal, lag.max = 20, type = "correlation", main = "CCF: 'Capital' & 'Sealand'")
```

```
ccf(pw_cap, pw_midj, lag.max = 20, type = "correlation", main = "CCF: 'Capital' & 'MidJutland'")
ccf(pw_cap, pw_rul, lag.max = 20, type = "correlation", main = "CCF: 'Capital' & 'Rural'")
ccf(pw_seal, pw_midj, lag.max = 20, type = "correlation", main = "CCF: 'Sealand' & 'MidJutland'")
ccf(pw_seal, pw_rul, lag.max = 20, type = "correlation", main = "CCF: 'Sealand' & 'Rural'")
ccf(pw_midj, pw_rul, lag.max = 20, type = "correlation", main = "CCF: 'MidJutland' & 'Rural'")
```

Listing 7: Question 4.4 - Residual diagnostics

```
# white noise
tsdiag(m1)
par(mfrow=c(1,1))
pacf(residuals(m1), main='PACF of Residuals')
cpgram(m1$residuals, main='Cumulative Periodogram of Residuals')
grid()
qqPlot(residuals(m1), main='Q-Q plot of Residuals', ylab='Residuals')
```

Listing 8: Question 4.5 - Forecasting the future house prices (I)

```
forecast_data <- predict(m1, 6)
print(forecast_data)
plot(time[1:122],denmark, xlim=c(1992.5,2025), type="l",
      xlab='Time [year]', ylab='House Price [DKK/m2]',
      main='Predictions without external inputs')
matlines(time[123:128], forecast_data$pred + 1.96*cbind(0, -forecast_data$se, forecast_data$se), col=2, lty=c(1,2,2))
grid()
legend("topleft", legend = c("Training data", "Predictions with 95% p.i."), lty=1, col=1:2)
```

Listing 9: Question 4.7 - Forecasting the future house prices (II)

```
interest.pred <- c(interest[123:124], replicate(4,interest[124]))
inflation.pred <- c(inflation[123:124], replicate(4,inflation[124]))
externalreg.pred <- cbind(cumsum(interest.pred) + externalreg[122,1], cumsum(inflation.pred) + externalreg[122,2])

m2.pred <- predict(m2, n.ahead = 6, newxreg = externalreg.pred)
print(m2.pred)
```

```

plot(time[1:122],denmark, xlim=c(1992.5,2025), type="l",
     xlab='Time [year]', ylab='House Price [DKK/m2]',
     main='Predictions with external inputs')
matlines(time[123:128], m2.pred$pred + 1.96*cbind(0, -m2.pred$se,
   m2.pred$se), col=2, lty=c(1,2,2))
grid()
legend("topleft", legend = c("Training data", "Predictions with
   95% p.i."), lty=1, col=1:2)
```

Listing 10: Part 2 - Multivariate models (initialization)

```
library(marima)
library(car)

data <- read.csv("A3Data.csv")

# Initialize an empty list to store the results
ts_list <- list()

# Convert the data to quarterly time series
ts <- ts(data, frequency = 4)

# Assign the values to separate variables
time <- seq(1992.5,2024.25,0.25)
denmark <- na.omit(ts[,2])
capital <- na.omit(ts[,3])
sealand <- na.omit(ts[,4])
midjutland <- na.omit(ts[,5])
rural <- na.omit(ts[,6])
interest <- na.omit(ts[,7])
inflation <- na.omit(ts[,8])
```

Listing 11: Question 4.9 - Re-presenting the data

```
par(mfrow=c(1,1))
plot(time[1:122],capital,type='l',
     ylab="House Price [DKK/m2]",xlab="Time [year]",
     main='Quarterly average sales prices for each of the four
         regions')
lines(time[1:122],sealand,col='steelblue')
lines(time[1:122],midjutland,col='tomato')
lines(time[1:122],rural,col='seagreen')
grid()
```

```
legend("topleft",c("Capital region","Sealand","Middle Jutland","Rural areas"),
       col=c("black","steelblue","tomato","seagreen"),lty=1,bty='n')

#Differences
dif_cap <- diff(capital)
dif_seal <- diff(sealand)
dif_midj <- diff(midjutland)
dif_rul <- diff(rural)

plot(time[2:122],dif_cap,type='l',
     ylab="House Price [DKK/m2]",xlab="Time [year]",
     main='Quarterly average sales prices for each of the four regions (differenced)')
lines(time[2:122],dif_seal,col='steelblue')
lines(time[2:122],dif_midj,col='tomato')
lines(time[2:122],dif_rul,col='seagreen')
grid()
legend("topleft",c("Capital region","Sealand","Middle Jutland","Rural areas"),
       col=c("black","steelblue","tomato","seagreen"),lty=1,bty='n')
```

Listing 12: Question 4.10 - ACF and PACF (mutlivariate)

```
# Raw data
four_series <- data.frame(ts[1:122,3:6])
par(mfrow=c(2,2))
for(i in 1:4) {
  series <- four_series[,i]
  acf(series, lag.max=25, main=colnames(four_series)[i])
}
par(mfrow=c(2,2))
for(i in 1:4) {
  series <- four_series[,i]
  pacf(series, lag.max=25, main=colnames(four_series)[i])
}
# Differenced data
dif_four_series <- cbind(dif_cap,dif_seal,dif_midj,dif_rul)
par(mfrow=c(2,2))
for(i in 1:4) {
  dif_series <- dif_four_series[,i]
  acf(dif_series, lag.max=25, main=colnames(four_series)[i])
```

```
}
par(mfrow=c(2,2))
for(i in 1:4) {
  dif_series <- dif_four_series[,i]
  pacf(dif_series, lag.max=25, main=colnames(four_series)[i])
}
```

Listing 13: Question 4.11 - Multivariate model selection

```
dif_data <- cbind(dif_cap, dif_seal, dif_midj, dif_rul, ts
    [2:122,7:8]) # don't take the first element because of
    differencing

# Model 1: just ma process with extrenal reg
ar<-c(0)
ma<-c(1)

Model1 <- define.model(kvar=6, ar=ar, ma=ma, reg.var=c(5,6))

par(mfrow=c(1,1))
M1 <- marima(dif_data, means=1, ar.pattern=Model1$ar.pattern,
             ma.pattern=Model1$ma.pattern, Check=FALSE, Plot="log.
                det")

short.form(M1$ar.estimates, leading=TRUE) # print estimates
short.form(M1$ma.estimates, leading=TRUE)

acf(t(M1$residuals)[,1:4],lag.max=25) # not good, we see 4-
    periodicity
pacf(t(M1$residuals)[,1:4],lag.max=25)

# Model 2: ma process with seasonality and with extrenal reg
# (0,0,1)(1,0,0)[4]
ar<-c(4)
ma<-c(1)

Model2 <- define.model(kvar=6, ar=ar, ma=ma, reg.var=c(5,6))
M2 <- marima(dif_data, means=1, ar.pattern=Model2$ar.pattern,
             ma.pattern=Model2$ma.pattern, Check=FALSE, Plot="log.
                det",penalty=2)

short.form(M2$ar.estimates, leading=TRUE) # print estimates
short.form(M2$ma.estimates, leading=TRUE)

```

```r
acf(t(M2$residuals)[,1:4],lag.max=25, na.action = na.omit) # better but could be improve especially for the 2 first series
pacf(t(M2$residuals)[,1:4],lag.max=25, na.action = na.omit) # I think it is fine

# Model 3: ma process with seasonality and with extrenal reg
# (0,0,1)(2,0,0)[4]
ar<-c(4,8)
ma<-c(1)

Model3 <- define.model(kvar=6, ar=ar, ma=ma, reg.var=c(5,6))
M3 <- marima(dif_data, means=1, ar.pattern=Model3$ar.pattern,
              ma.pattern=Model3$ma.pattern, Check=FALSE, Plot="log.det", penalty=2)


short.form(M3$ar.estimates, leading=TRUE) # print estimates
short.form(M3$ma.estimates, leading=TRUE)

acf(t(M3$residuals)[,1:4],lag.max=25, na.action = na.omit)
pacf(t(M3$residuals)[,1:4],lag.max=25, na.action = na.omit)


# Model 4: ma process with seasonality and with independant variable
# (0,0,1)(2,0,0)[4]
ar<-c(4,8)
ma<-c(1)

Model4 <- define.model(kvar=6, ar=ar, ma=ma, indep=c(5,6)) # just change indep param
M4 <- marima(dif_data, means=1, ar.pattern=Model4$ar.pattern,
              ma.pattern=Model4$ma.pattern, Check=FALSE, Plot="log.det",penalty = 2)

## Let's improve model 4

source("/Users/user/Desktop/DTU/1. Courses/4. Spring 23/02417 - Time Series Analysis/Week 9/step.slow.marima_2017.R")
M4 <- marima(dif_data, means=1, ar.pattern=Model4$ar.pattern,
              ma.pattern=Model4$ma.pattern, Check=FALSE, Plot="log.det")
M4_sl <- step.slow(M4, dif_data) #penalty = 2

```

```r

short.form(M4_sl$ar.estimates, leading=FALSE) # print estimates
short.form(M4_sl$ma.estimates, leading=FALSE)


acf(t(M4_sl$residuals)[,1:4],lag.max=25, na.action = na.omit)
pacf(t(M4_sl$residuals)[,1:4],lag.max=25, na.action = na.omit)


# Model 4 - Residual Diagnostics

par(mfrow = c(2,2))
titles <- c("Capital Region", "Sealand", "Middle Jutland", "Rural Areas")
for (i in 1:4){
  plot(M4_sl$residuals[i,], type='l', ylab = "Residual", main = titles[i])
}

# conduct the Ljung-Box test with 12 lags (one year)
for (i in 1:4){
    print(Box.test(M4_sl$residuals[i,], type = "Ljung-Box", lag = 12))
}
# good except maybe for 2 first one

# cpgram
par(mfrow = c(2,2))
for (i in 1:4){
    cpgram(M4_sl$residuals[i,], main = titles[i])
}
# good

# means
for (i in 1:4){
    print(mean(M4_sl$residuals[i,9:121]))
}
# good

# qqplot
par(mfrow = c(2,2))
for (i in 1:4){
    qqPlot(M4_sl$residuals[i,], ylab = "Residual", main = titles[i])
```

```
}
# fine

# t test
for (i in 1:4){
    print((t.test(M4_sl$residuals[i,])$p.value) )
}
# good

par(mfrow=c(1,1))
plot(M4_sl$trace,type='l')
# nice convergence
```

Listing 14: Question 4.13 - Forecasting the future house prices (III)

```
## Predicting 6 quarters ahead
pred.data <- t(data[2:128,3:8])
# add future value of regressors as in part 1
pred.data[5,124:127] <- rep(data[124,7],4)
pred.data[6,124:127] <- rep(data[124,8],4)
time <- seq(1992.5,2024.25,0.25)

#Differences poly
difference <- matrix(c(1,1,2,1,3,1,4,1), nrow=2)
Y <- define.dif(ts[1:122,3:8], difference=difference)
names(Y)
str(Y)
dif_data <- Y$y.dif
y.lost <- Y$y.lost
dif.poly <- Y$dif.poly
averages <- Y$averages

# make prediction
pred <-  arma.forecast(pred.data, nstart=121, nstep=6, marima=M4_
    sl, dif.poly = dif.poly, check = TRUE)

## Time to plot
par(mfrow=c(1,1))
plot(time[1:121], pred.data[1,1:121], xlim=c(1992,2025),
     xlab = 'Time [year]', ylab = 'House Price [DKK/m2]',
     main = 'Quarterly average sales prices: Capital region of
        Denmark')
lines(time[2:128], pred$forecasts[1,], col=2)
pred.int <- pred$forecasts[1,122:127] + cbind(rep(0, 6), -1, 1)*
```

```
    qnorm(0.975)*sqrt(pred$pred.var[1,1,])
matlines(time[123:128], pred.int, lty=c(1,2,2), col=3, lwd=2 )
grid()
legend("topleft", legend = c("Training data", "1-step predictions"
    , "Forecasts with 95% p.i."),
       lty = c(0, 1, 1), col = 1:3, pch = c(1, NA, NA))

plot(time[1:121], pred.data[2,1:121], xlim=c(1992,2025),
     xlab = 'Time [year]', ylab = 'House Price [DKK/m2]',
     main = 'Quarterly average sales prices: Sealand')
lines(time[2:128], pred$forecasts[2,], col=2)
pred.int <- pred$forecasts[2,122:127] + cbind(rep(0, 6), -1, 1)*
    qnorm(0.975)*sqrt(pred$pred.var[2,2,])
matlines(time[123:128], pred.int, lty=c(1,2,2), col=3, lwd=2 )
grid()
legend("topleft", legend = c("Training data", "1-step predictions"
    , "Forecasts with 95% p.i."),
       lty = c(0, 1, 1), col = 1:3, pch = c(1, NA, NA))

plot(time[1:121], pred.data[3,1:121], xlim=c(1992,2025),
     xlab = 'Time [year]', ylab = 'House Price [DKK/m2]',
     main = 'Quarterly average sales prices: Middle Jutland')
lines(time[2:128], pred$forecasts[3,], col=2)
pred.int <- pred$forecasts[3,122:127] + cbind(rep(0, 6), -1, 1)*
    qnorm(0.975)*sqrt(pred$pred.var[3,3,])
matlines(time[123:128], pred.int, lty=c(1,2,2), col=3, lwd=2 )
grid()
legend("topleft", legend = c("Training data", "1-step predictions"
    , "Forecasts with 95% p.i."),
       lty = c(0, 1, 1), col = 1:3, pch = c(1, NA, NA))

plot(time[1:121], pred.data[4,1:121], xlim=c(1992,2025),
     xlab = 'Time [year]', ylab = 'House Price [DKK/m2]',
     main = 'Quarterly average sales prices: Rural areas')
lines(time[2:128], pred$forecasts[4,], col=2)
pred.int <- pred$forecasts[4,122:127] + cbind(rep(0, 6), -1, 1)*
    qnorm(0.975)*sqrt(pred$pred.var[4,4,])
matlines(time[123:128], pred.int, lty=c(1,2,2), col=3, lwd=2 )
grid()
legend("topleft", legend = c("Training data", "1-step predictions"
    , "Forecasts with 95% p.i."),
       lty = c(0, 1, 1), col = 1:3, pch = c(1, NA, NA))
```

Extra Plots

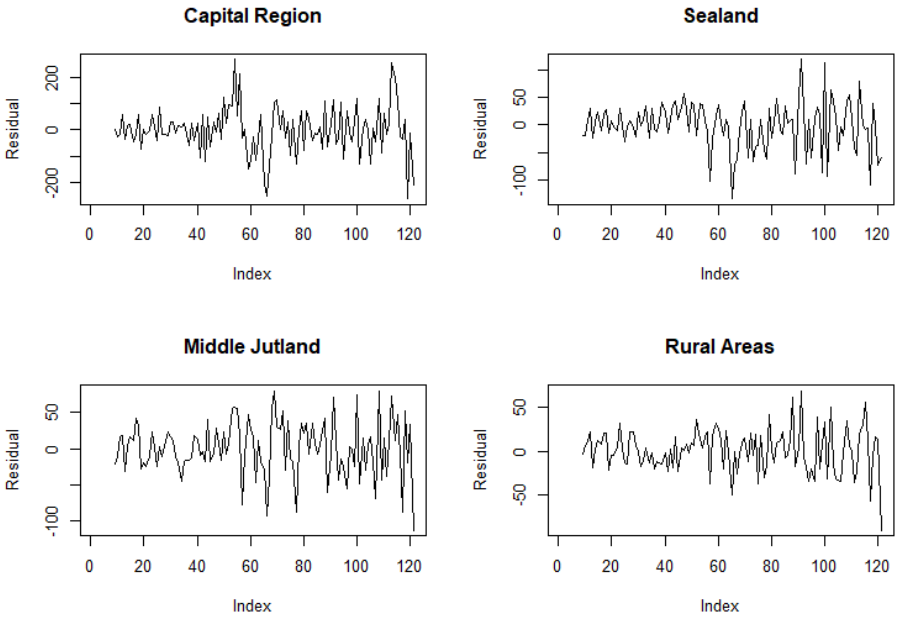

Figure 25: Multivariate Model - Residual Diagnostics (I))

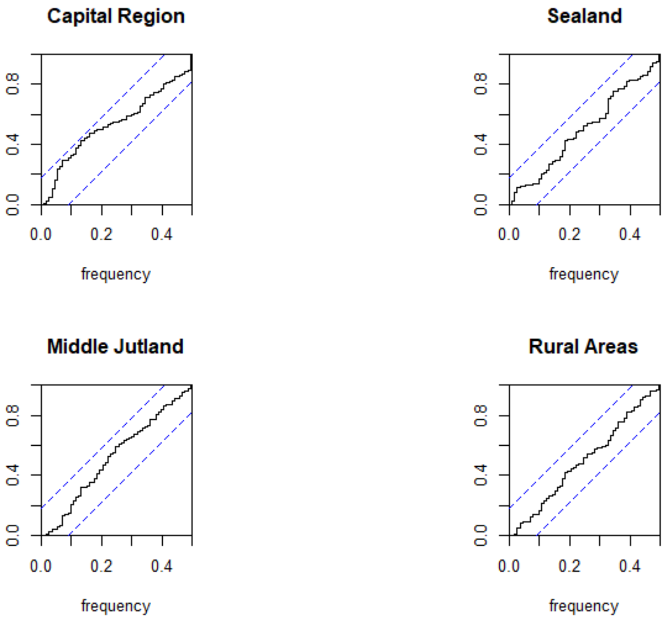

Figure 26: Multivariate Model - Residual Diagnostics (II))

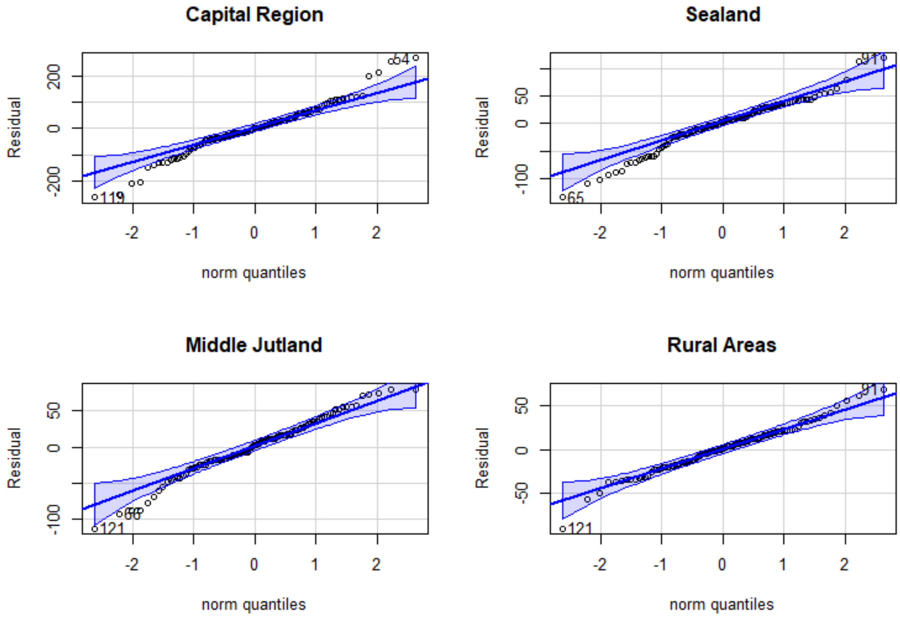

Figure 27: Multivariate Model - Residual Diagnostics (III))