Assignment 1

Task 1:

Goal:

The aim of this task is to fit three different GLMs based on various distributions, namely Gamma, Gaussian, and Rayleigh distributions, to identify the most suitable one for ground-type detection in a SAR image for San Francisco and Karlskrona. The ground types include urban, sea, and forest.

Tools:

- Matlab

Method:

Started by defining test regions with a size of 20x20 for each ground type in every image. The data behavior was visually checked to verify the suitability of the considered regression models for fitting the data. This was done by creating box and histogram plots for each ground type after vectorizing the data.

Additionally, two dummy covariates, x2 and x3, were created. x2 is defined as one for the forest region and zero for the other regions, and x3 is defined as one for the urban region and zero for the other regions if both x2 and x3 are zeros it represents the sea region. The regression models were then fitted, and detection theory was performed.

To identify any obvious patterns in the models, a residual test was conducted. This involved creating index, histogram, and QQ plots. Furthermore, the relationship between the mean of y[n] and the dummy covariates was visually examined by plotting the means in a barplot. The determination coefficient for the fitted models was also verified by computing the R-squared value for each model.

Results:

San Francisco image results are presented as follows.

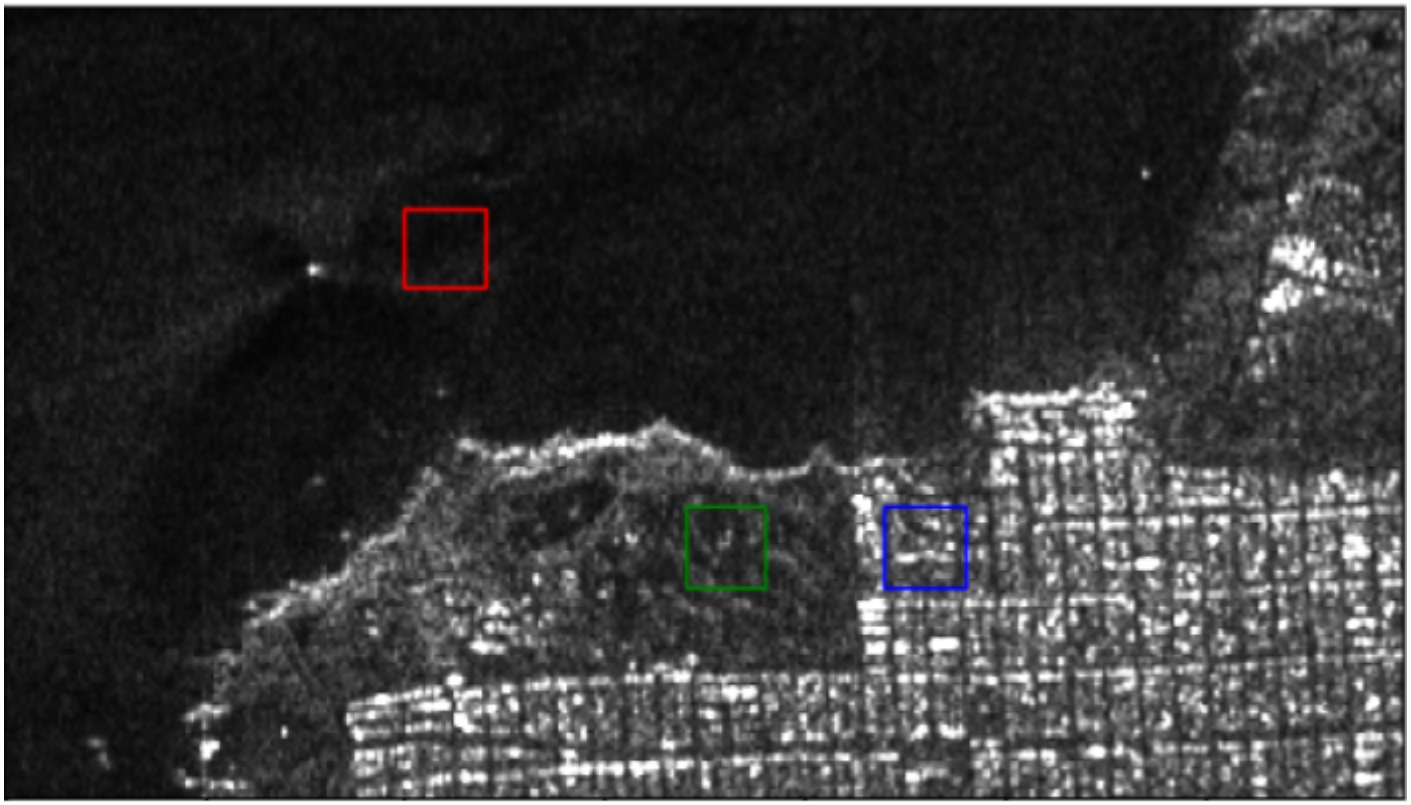

figure 1: San Francisco SAR image with the chosen 20x20 regions marked (red:sea, green:forest, blue:urban)

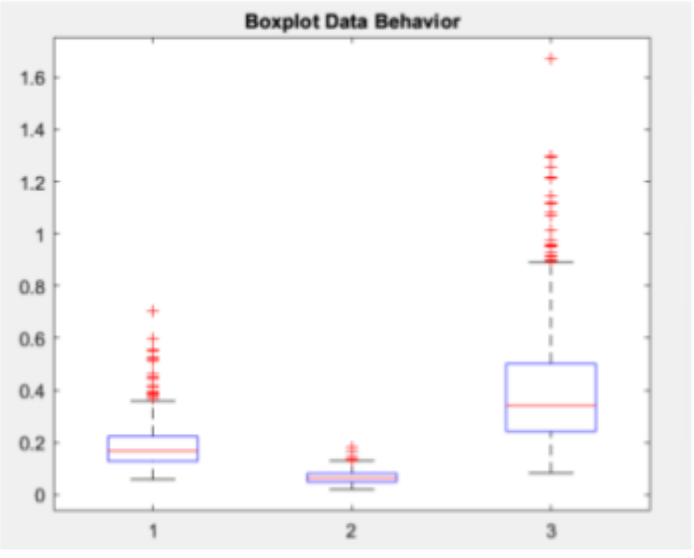

figure 2: Box plot data behavior

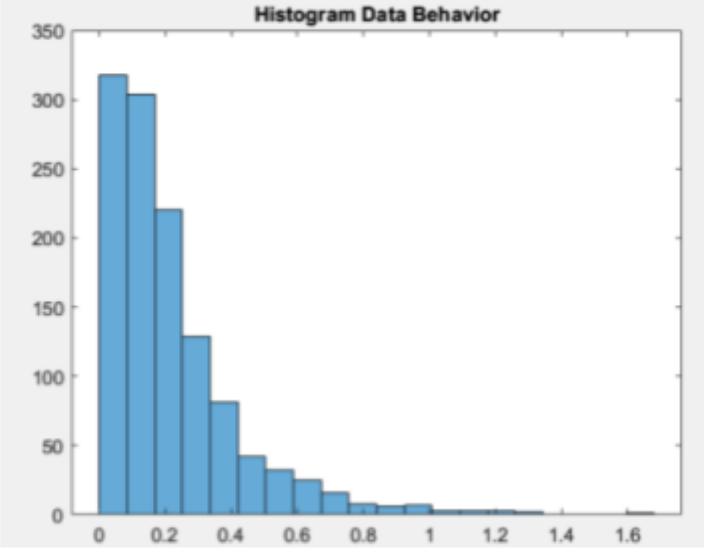

figure 3: Histogram data behavior

figure 4: Rayleigh distribution model, residual Index-plot

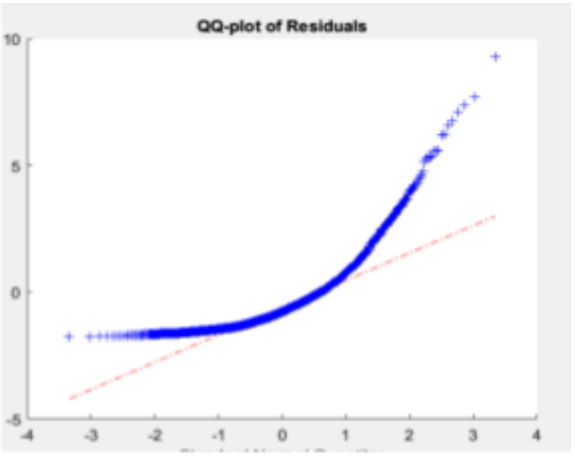

figure 5: Rayleigh distribution model, residual QQ-plot

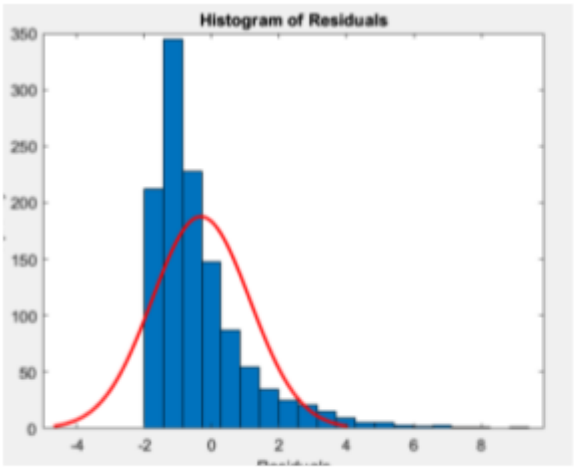

figure 6: Rayleigh distribution model, residual Histogram

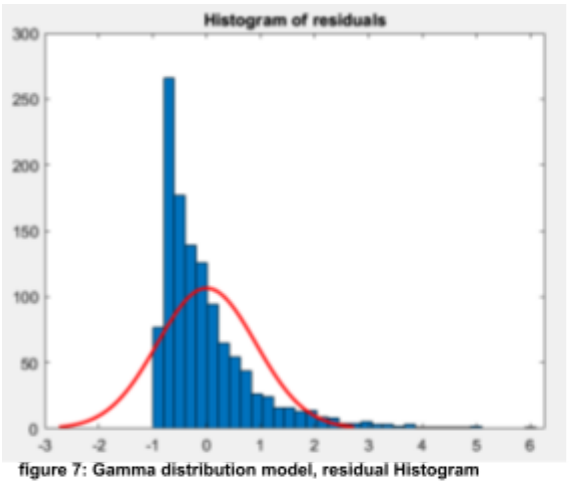

figure 7: Gamma distribution model, residual Histogram

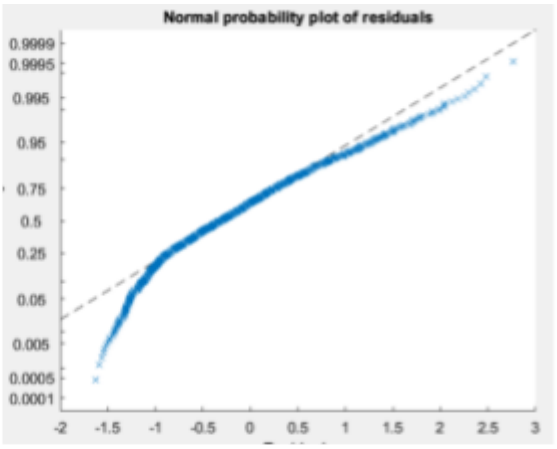

figure 8: Gamma distribution model, residual QQ-plot

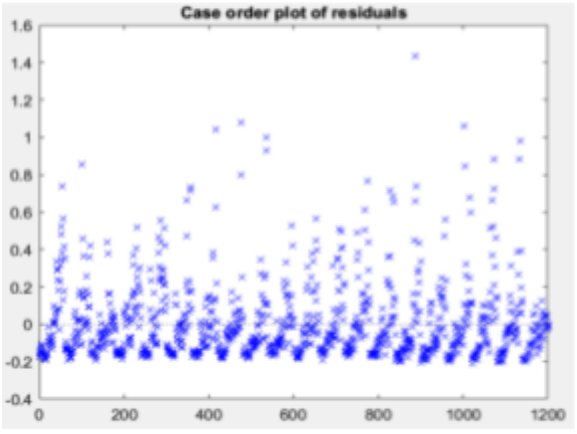

figure 9: Gamma distribution model, residual Index-plot

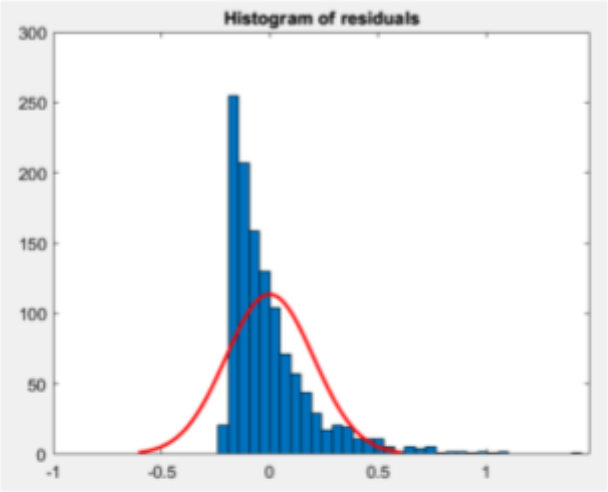

figure 10: Gaussian distribution model, Histogram

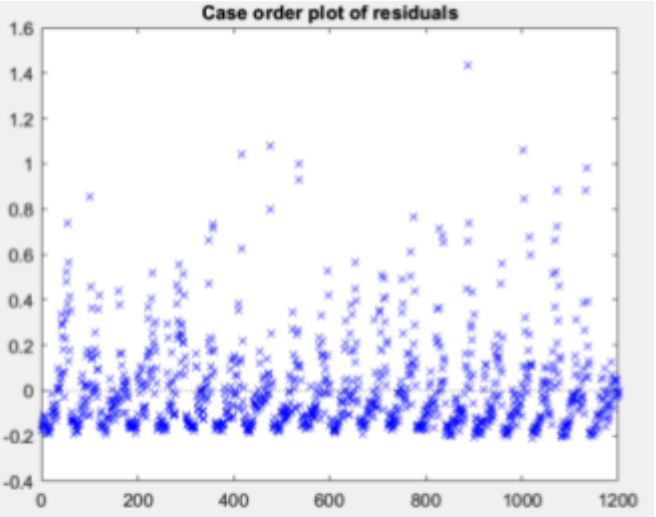

figure 11: Gaussian distribution model, Index-plot

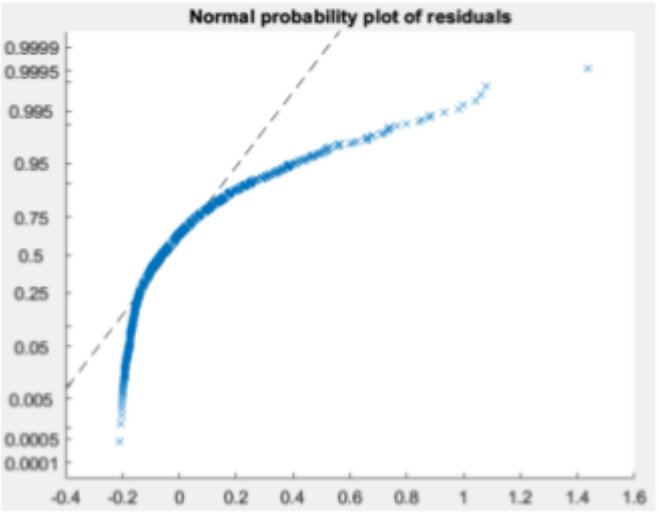

figure 12: Gaussian distribution model, QQ-plot

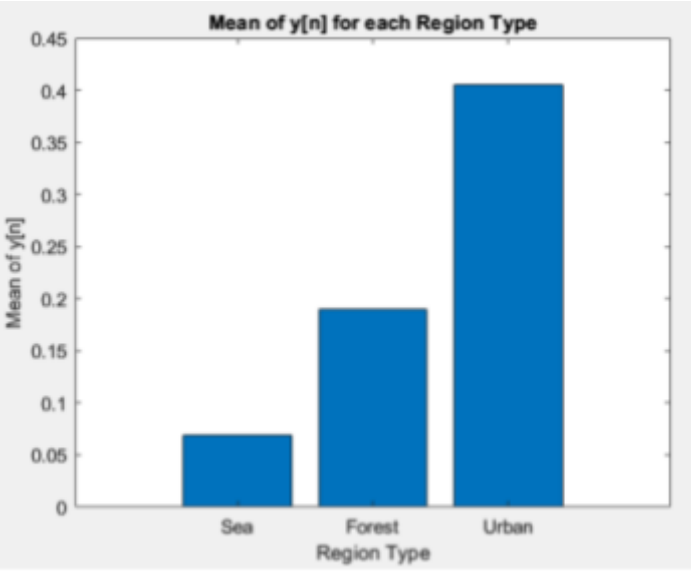

figure 13: Mean of y[n] for each region

p-value	Gamma-Distribution	Gaussian-Distribution	Rayleigh-Distribution
x1	0.79891	0.80582	0.5979
x2	0.17042	0.16319	0.0005

table 1: P-values result for each one of the models

Fitted-model	Gamma-Distribution	Gaussian-Distribution	Rayleigh-Distribution
R-squared	0.0026	0.0026	0.0121

table 2: Determination coefficient, R-squared for each model

Karlskrona image results are presented as follows.

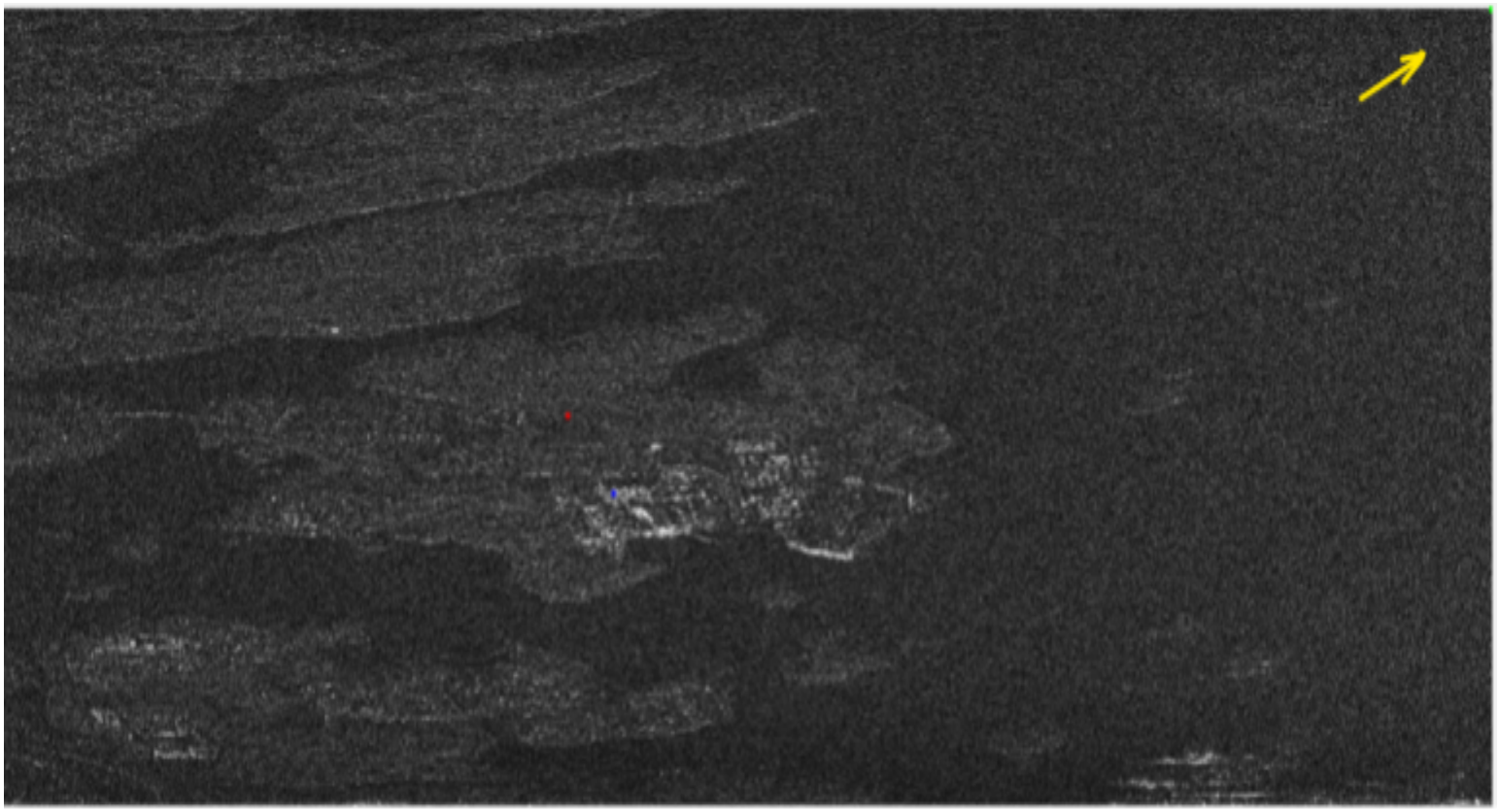

figure 14: Karlskrona SAR-image, the green(sea), red(forest) and blue(urban) dots represents the chosen regions

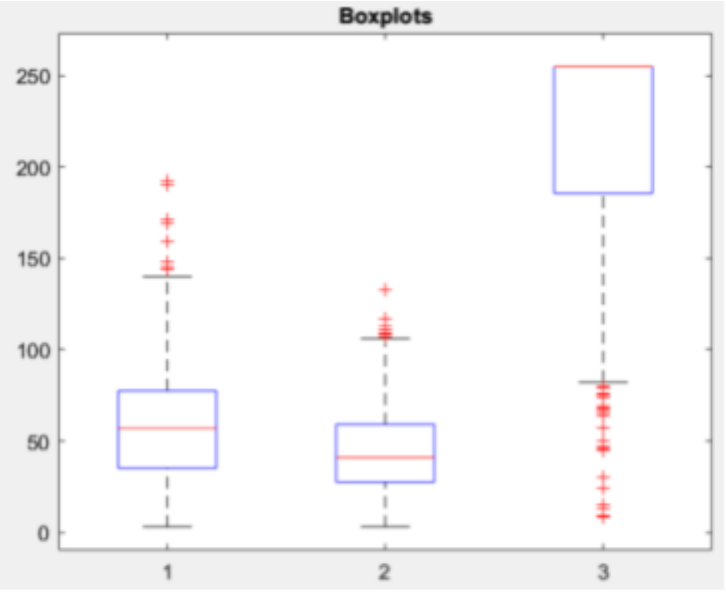

figure 15: Box plot data behavior

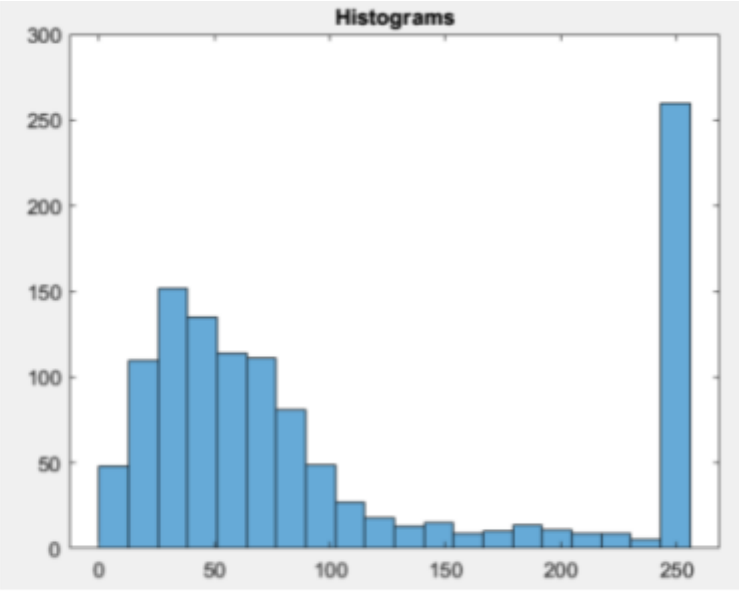

figure 16: Histogram data behavior

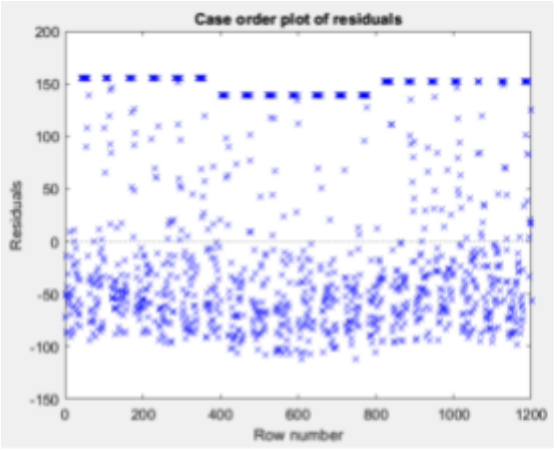

figure 17: Gaussian distribution model, residual Index-plot

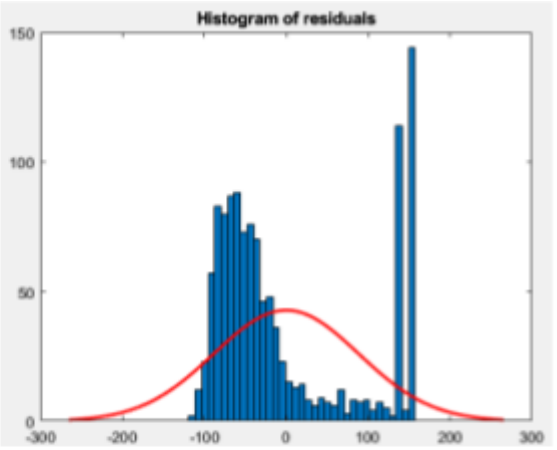

figure 18: Gaussian distribution model, residual Histogram

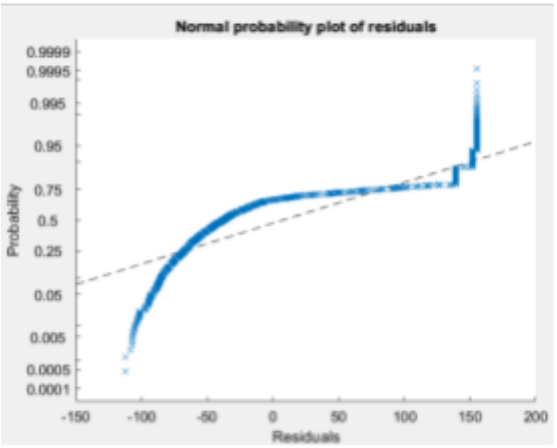

figure 19: Gaussian distribution model, residual QQ-plot

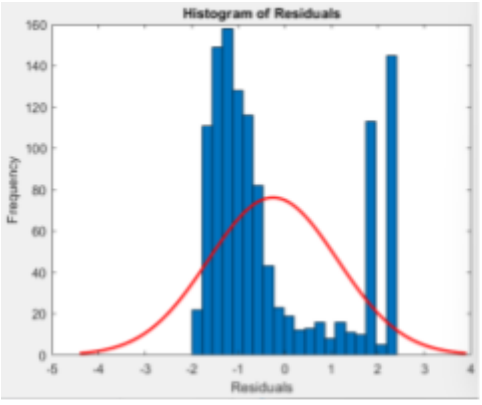

figure 21: Rayleigh distribution model, residual Histogram

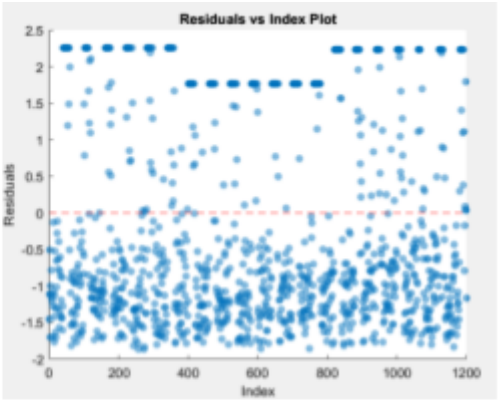

figure 20: Rayleigh distribution model, residual Index-plot

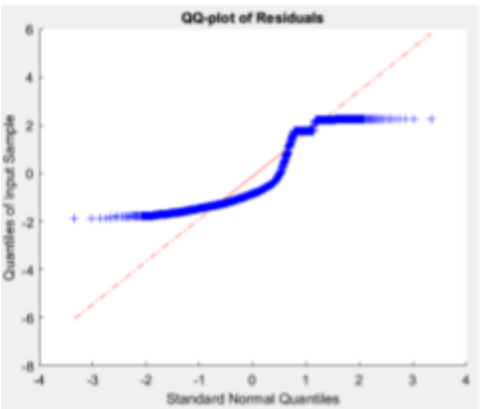

figure 22: Rayleigh distribution model, residual QQ-plot

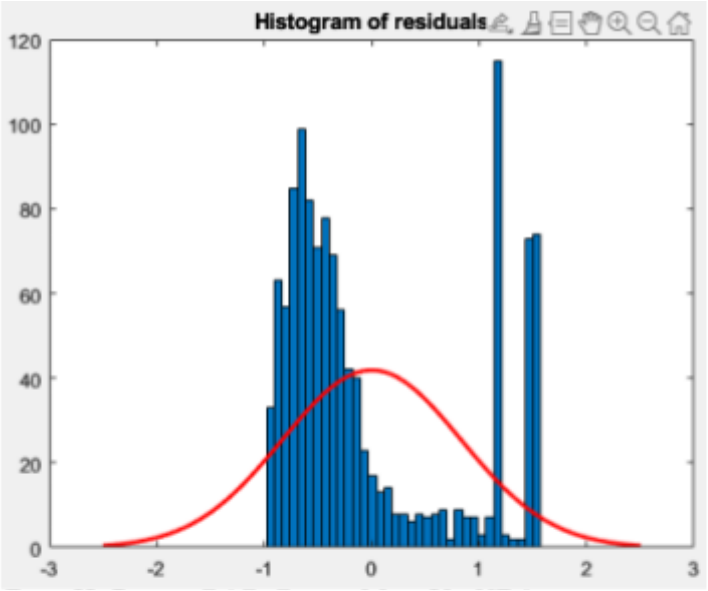

figure 23: Gamma distribution model, residual Histogram

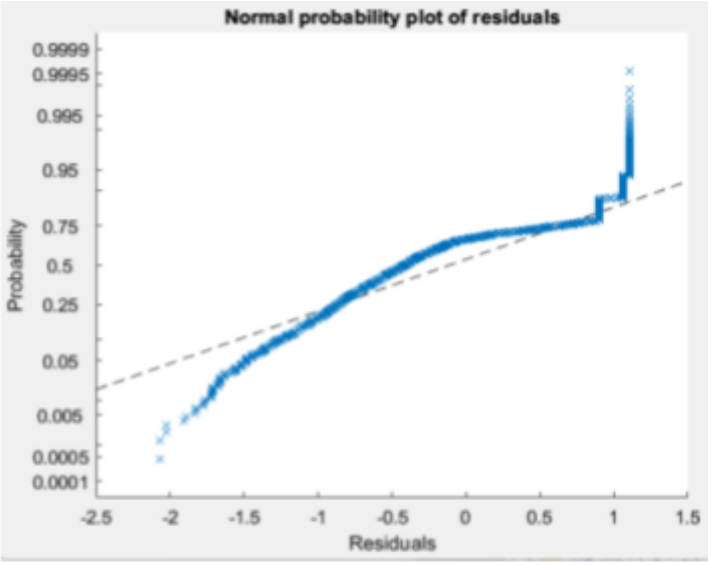

figure 24: Gamma distribution model, residual QQ-plot

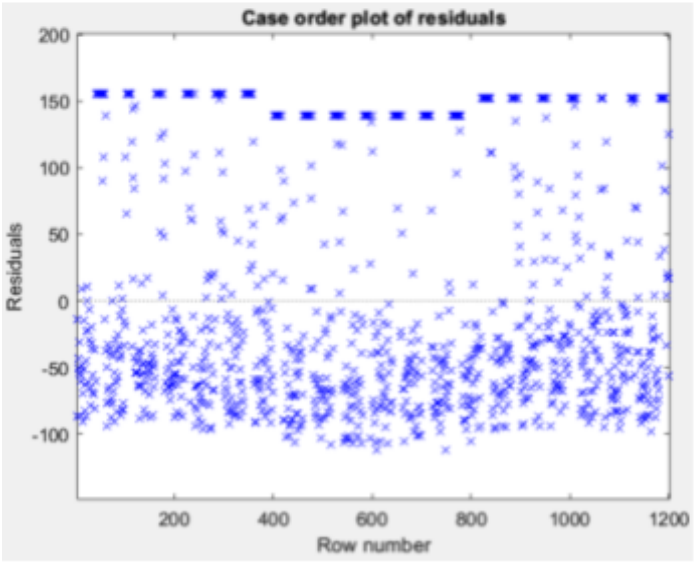

figure 25: Gamma distribution model, residual Index-plot

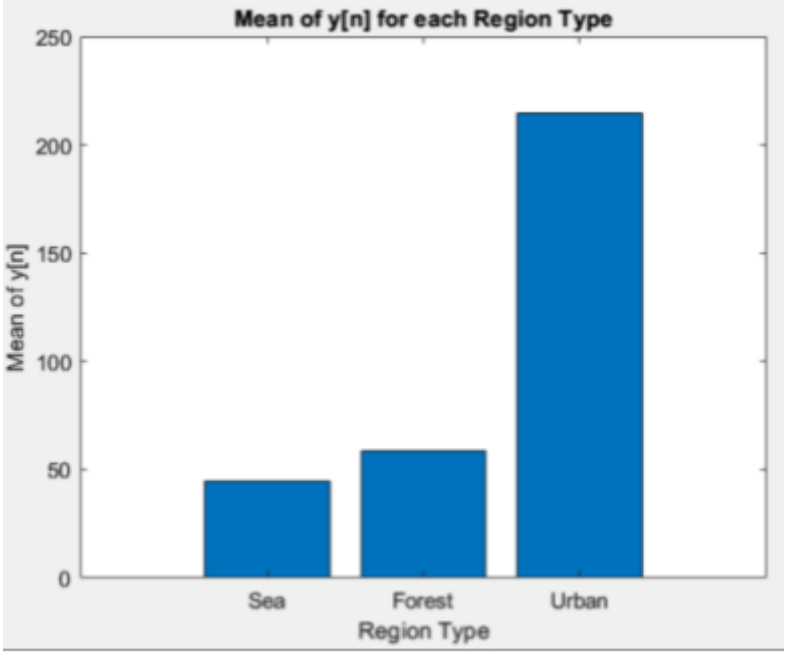

figure 26: Mean of y[n] for each region

p-value	Gamma-Distribution	Gaussian-Distribution	Rayleigh-Distribution
x1	0.0093522	0.0083363	0.0004
x2	0.60522	0.62235	0.8744

table 3: P-values result for each one of the models

p-value	Gamma-Distribution	Gaussian-Distribution	Rayleigh-Distribution
R-squared	0.0066	0.0066	0.0135

table 4: Determination coefficient, R-squared for each model

Discuss:

San Francisco-SF and Karlskrona-KN:

The regions selected for each type of ground, as portrayed in Figure 1 for SF and Figure 14 for KN, were diligently chosen to ensure they do not cross paths with other ground types. This deliberate selection was instrumental in mitigating the risk of potential overlaps, which could consequently lead to deceptive conclusions.

Upon visual inspection of the data behavior, as depicted in Figure 2 for SF and Figure 15 for KN, a discernible discrepancy in the medians across the regions is apparent. This noteworthy observation suggests that the pixel values within these regions could potentially follow distinct distributions. Moreover, Figure 3 for SF and Figure 16 for KN suggest a positive skewness in the data, leaning towards the right. This finding implies that both the Gamma and Rayleigh distributions could be appropriate fits for the data in both images.

The fundamental principles of detection theory emphasize that the strongest correlation between the predictor and response variables is found in the Rayleigh fitted model. This conclusion is expanded upon in Table 1 for SF and Table 3 for KN, providing a more detailed perspective.

The residuals of the fitted models are graphically represented in Figures 6, 7, and 10 for SF, and Figures 18, 21, and 23 for KN. These histograms exhibit almost bell-shaped curves, hinting at a close resemblance to a normal distribution. This suggests that all models are performing adequately in SF. However, in the case of the KN image, the Rayleigh model stands out by producing the most bell-shaped curve compared to the other models. Contrarily, the Quantile-Quantile (QQ) plots in

Figures 5, 8, and 12 for SF and Figures 19, 22, and 24 for KN dispute this observation, as they deviate from the typical linear pattern of a normal distribution. This discrepancy emerges as the plotted points fail to consistently align along a straight line.

Additionally, Figures 4, 9, and 11 for SF, representing the residuals of index plots, suggest that the residuals are not purely random. They indicate the presence of a recognizable pattern, an observation that is also applicable to Figures 17, 20, and 25 for KN.

Finally, Figures 13 and 26 for both images shows a difference between the mean values of each region. This suggests that the detected signal, y[n], demonstrates distinct traits within each region. The increased mean value observed in the urban region implies that urban regions within the images exhibit a stronger intensity of the SAR signal in comparison to both sea and forest regions.

The results, as presented in Table 2 for SN and Table 4 for KN, reveal that the Rayleigh model achieved the highest R-squared value. This finding further reinforces the notion that the Rayleigh model is an effective and suitable model for this specific type of SAR-image analysis.

Task 2:

Goal:

The goal of this task is to explore, model, and forecast hidden unemployment rates and relative humidity using time series models. This includes the utilization of Gaussian-, beta-, and Kumaraswamy-based time series models. The data sets are derived from the conditions in two Brazilian cities.

In this task, we aim to delve into the exploration, modeling, and prediction of hidden unemployment rates and relative humidity. To achieve this, we will employ various time series models such as Gaussian, beta, and Kumaraswamy models.

Tools:

- Matlab and R-studio

Method:

The first step involves reading the data from a text file. Initially, the data is read as character strings. To perform mathematical and statistical operations on this data, it is necessary to convert these character strings into numeric data. Once the data is in an appropriate format, the next crucial step is to convert it into a time series. After that, the issue of missing values is addressed. In any dataset, the presence of missing or NA values can lead to inaccurate results or errors during analysis. To ensure

the integrity of the data, we first check for NA values and then remove them from the dataset, effectively cleaning the data for further analysis. The final significant operation is splitting the data into training and testing sets.

Firstly, an ARMA model is fitted to the training data using the `auto.arima` function from the forecast library. This function automatically selects the best ARMA model that fits the data by comparing different models based on the Akaike Information Criterion (AIC), a measure of the relative quality of statistical models.

Next, we examine the residuals of the fitted model. Residuals are the differences between the observed and predicted values, and analyzing them is an important diagnostic tool to assess the quality of the model. We check the residuals using the "check residuals" function, which provides a plot of the residuals, a histogram with an overlaid density plot, and a Ljung-Box test of the residuals.

Finally, the model generates forecasts for the next 10 periods and plots the forecasted values alongside the actual values for comparison.

The next task is to fit a beta autoregressive moving average (BARMA) model to the training data using the BARMA functions from Canvas. The function does not automatically select the best AR and MA parameters, so we pass the ar = c(1,2,3,4,5) and ma = c(NA) to specify the orders of the autoregressive and moving average parts of the model. We run the model with orders 1 to 5 and select the best model. To ensure that we choose the best parameters for the model, we checked the residuals, AFC, and PACF.

The final task is to fit a Kumaraswamy distribution (KARMA) model. We applied the same theory as in the beta distribution but used the KARMA functions from Canvas.

Results:

the relative humidity data set

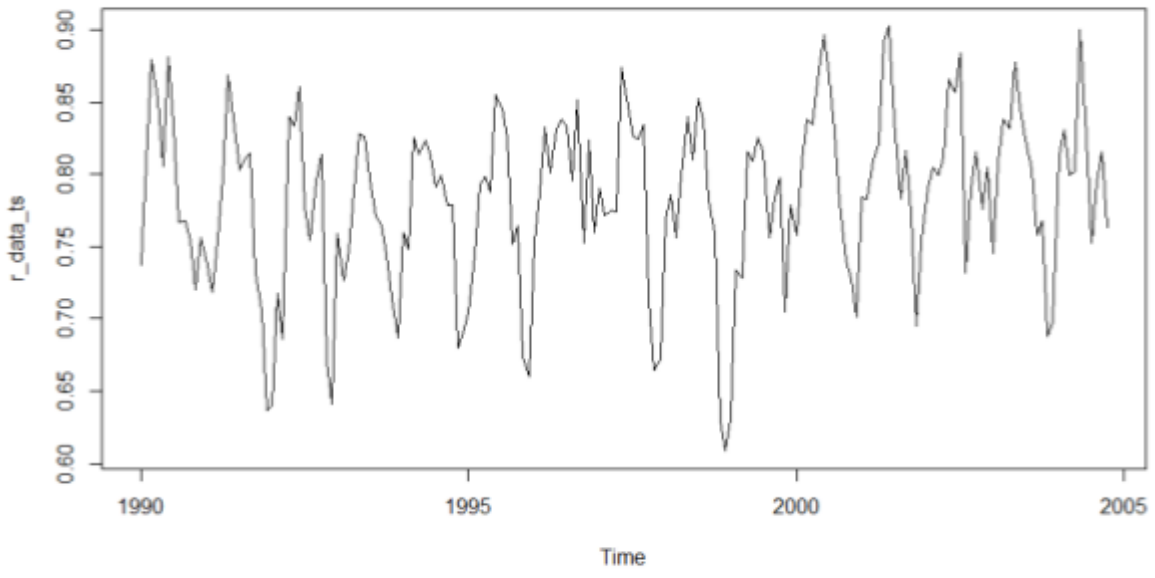

figure 0: relative humidity dataset time-series

ARMA model:

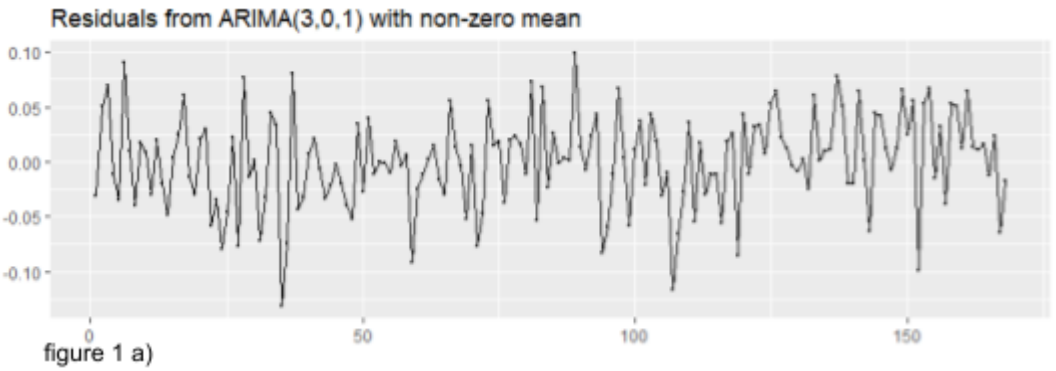

figure 1 a)

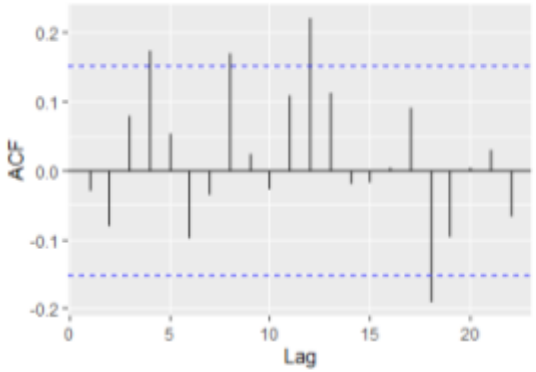

figure 1 b)

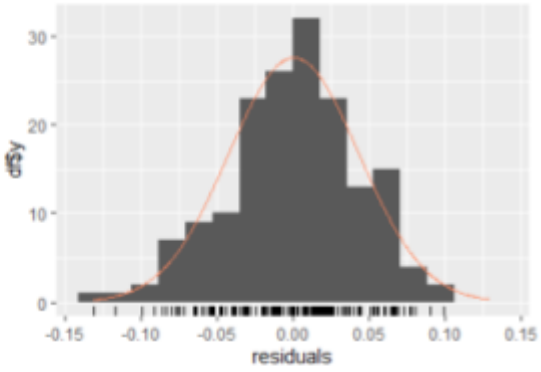

figure 1 c)

figure 1 a), b) and c) Residuals

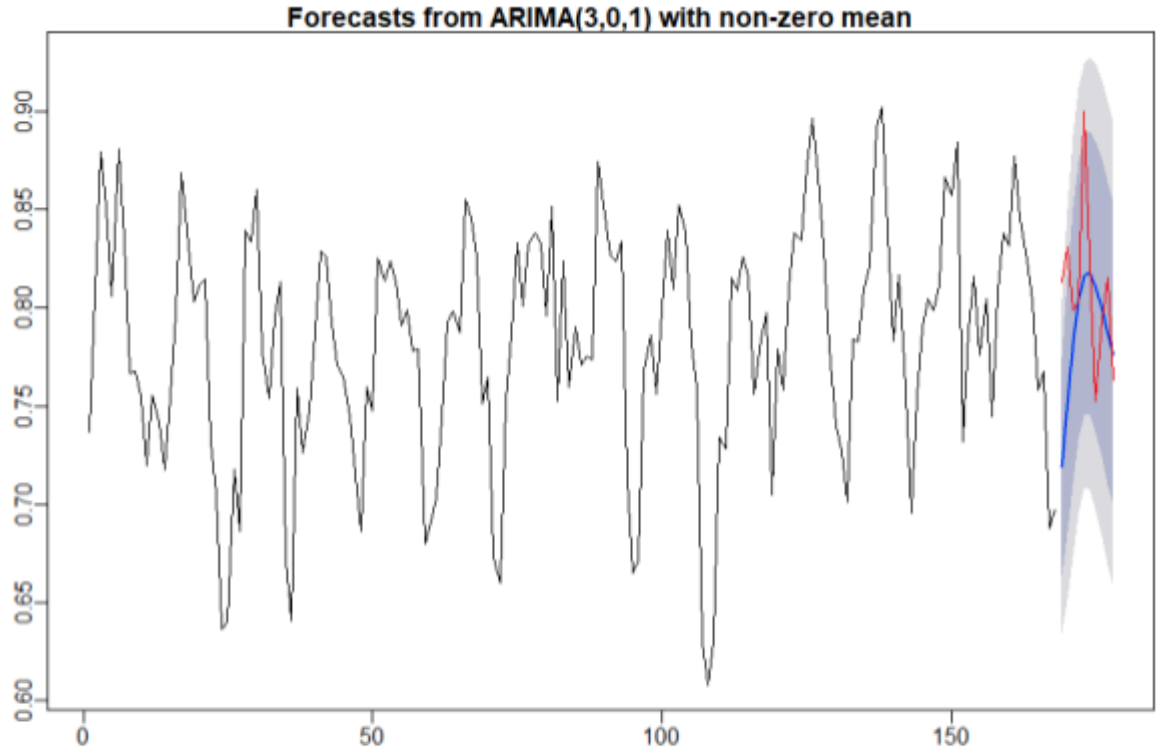

figure 2: ARMA model prediction

BARMA:

```
Estimate Std. Error z value Pr(>|z|)
alpha       0.9564     0.0900 10.6304   0.0000
phi1        0.5468     0.0578  9.4549   0.0000
phi2        0.0683     0.0674  1.0133   0.3109
phi3       -0.0283     0.0672  0.4213   0.6736
phi4       -0.0472     0.0673  0.7014   0.4831
phi5       -0.2832     0.0578  4.9010   0.0000
precision  94.7279    10.4577  9.0582   0.0000
[1]
[1] Log-likelihood: 299.5267
[1] Number of iterations in BFGS optim: 83
[1] AIC:      -585.0534  BIC:     -563.1856
[1] Residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-3.80145 -0.75573  0.06614  0.09502  0.79512  3.31906
```

report 2: BARMA model

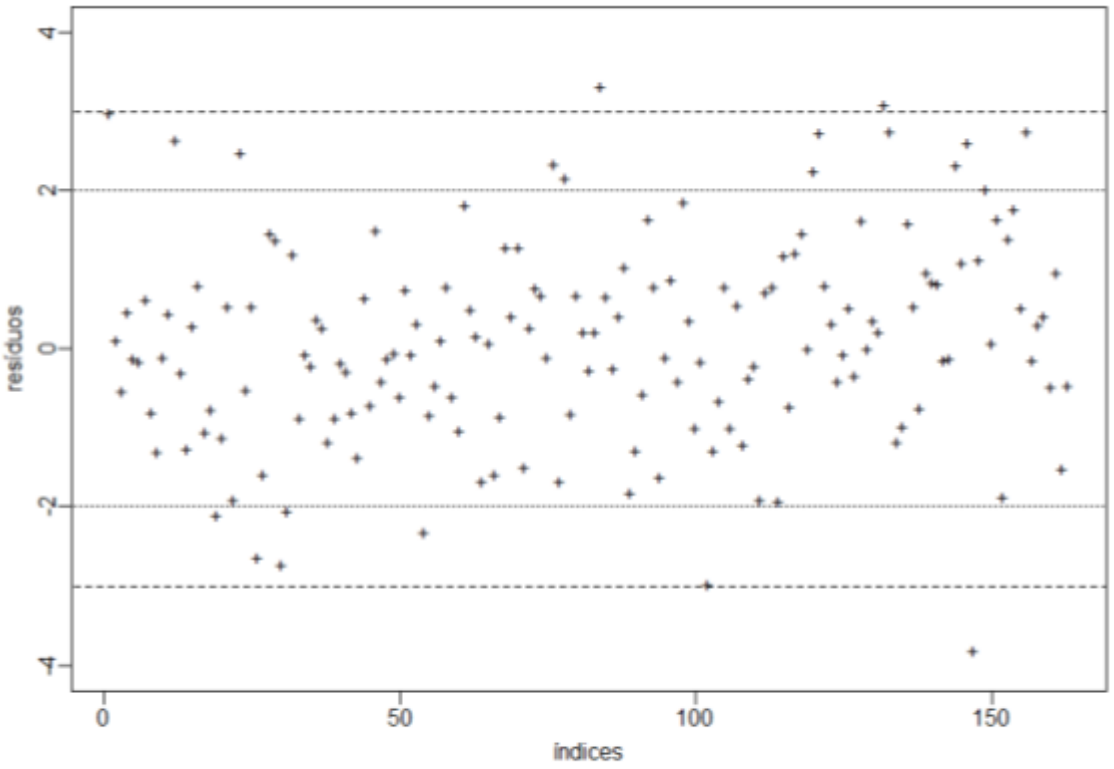

figure 3: BARMA model residuals

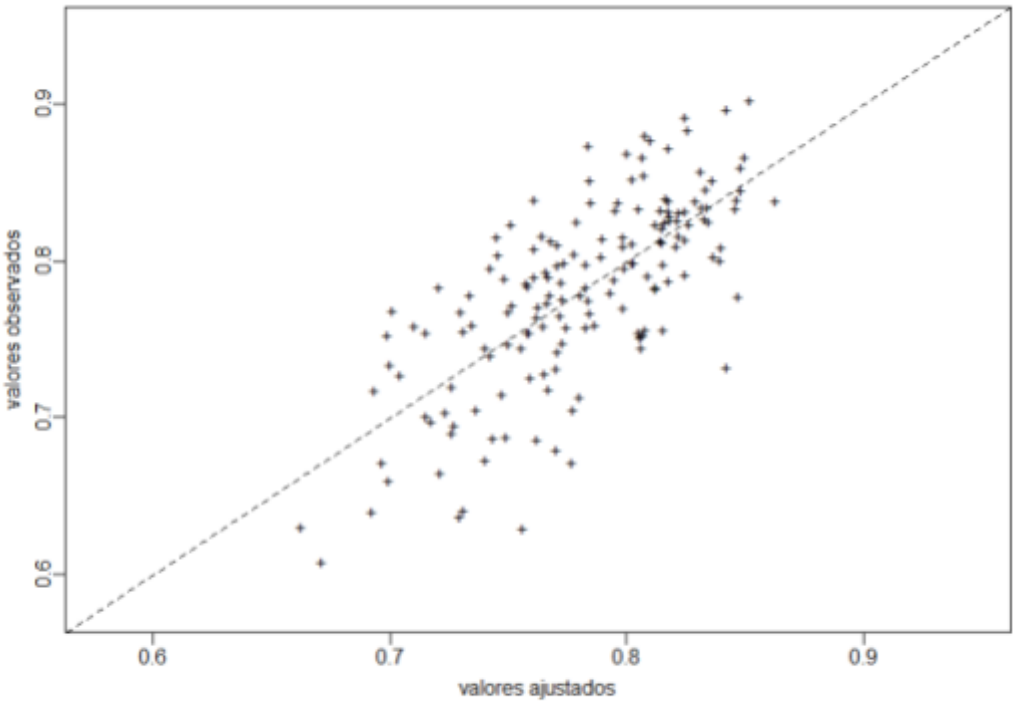

figure 4: BARMA model residuals

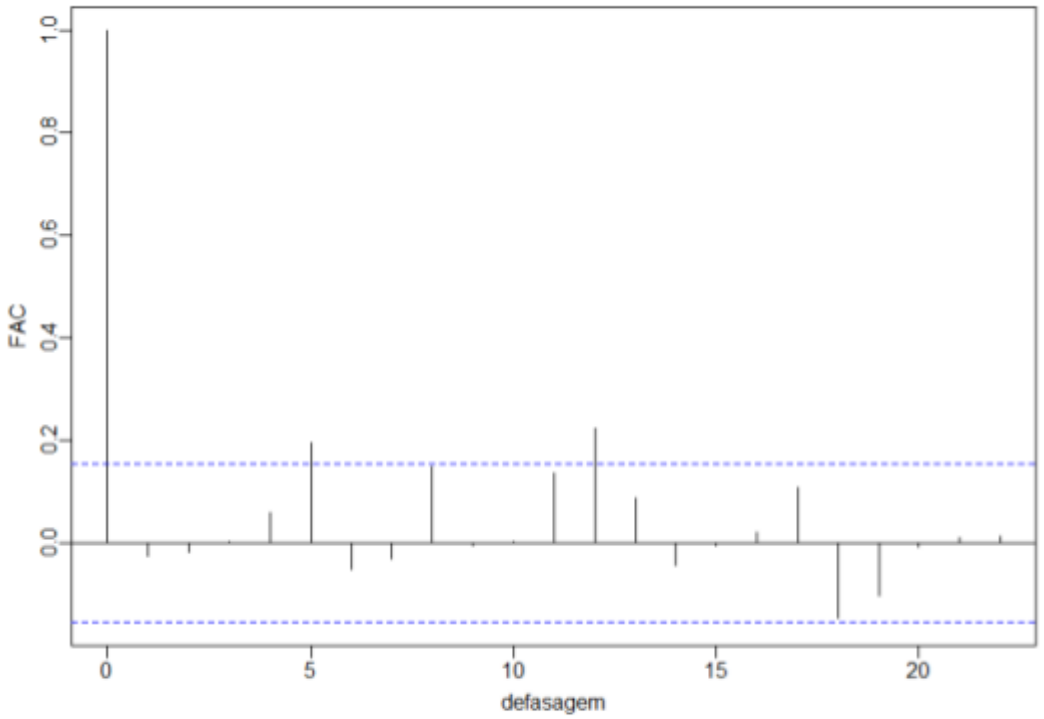

figure 5: BARMA model, ACF

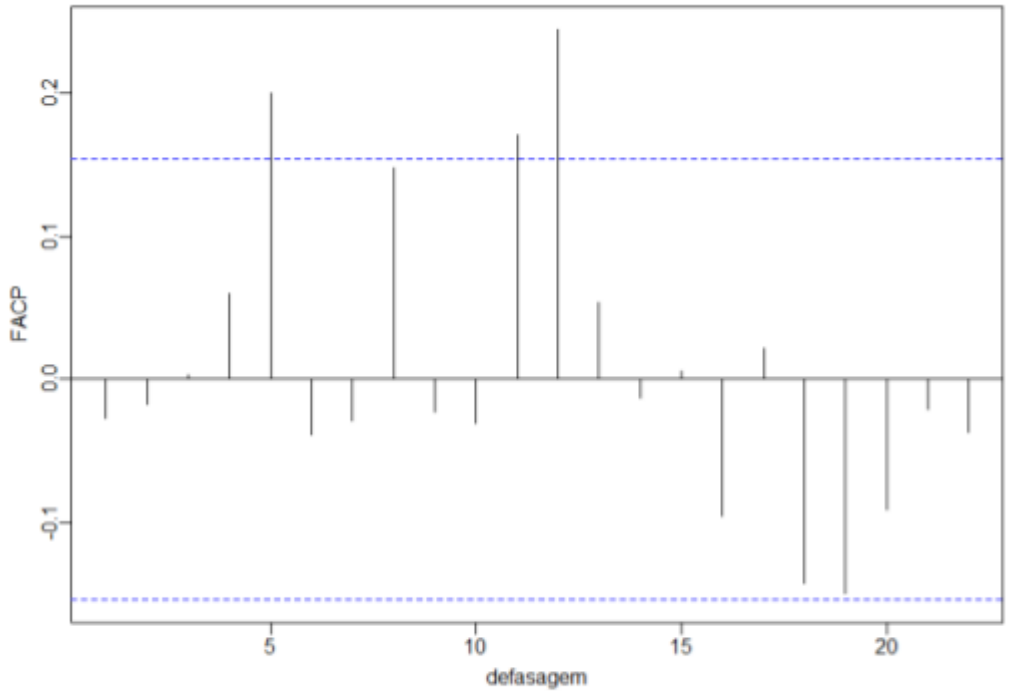

figure 6: BARMA model, FACP

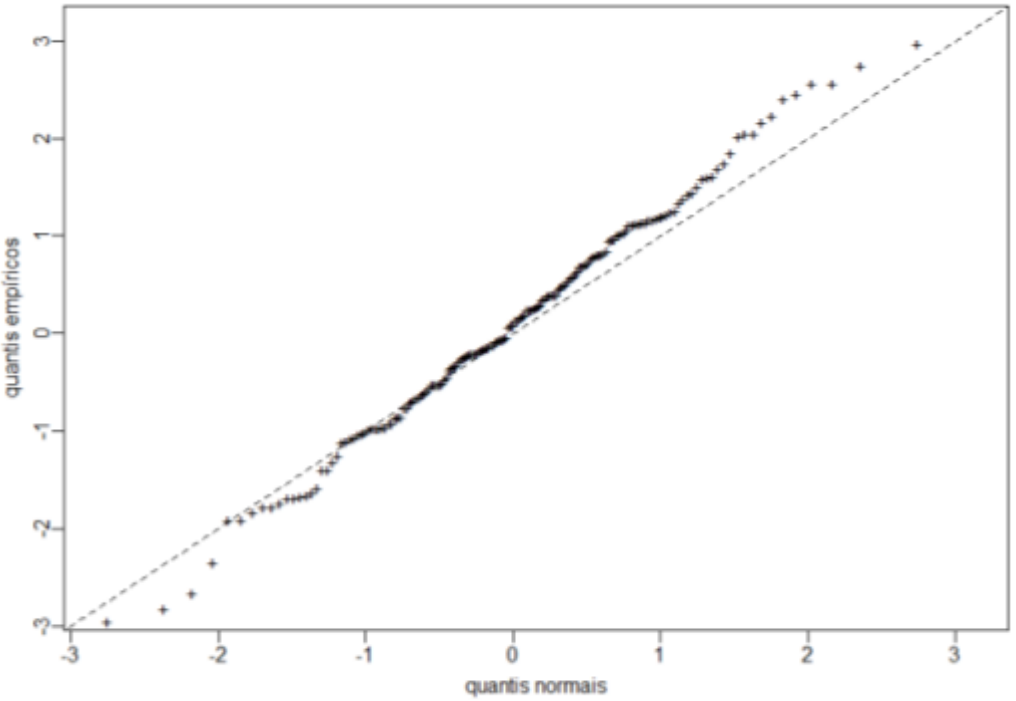

figure 7: BARMA model residuals

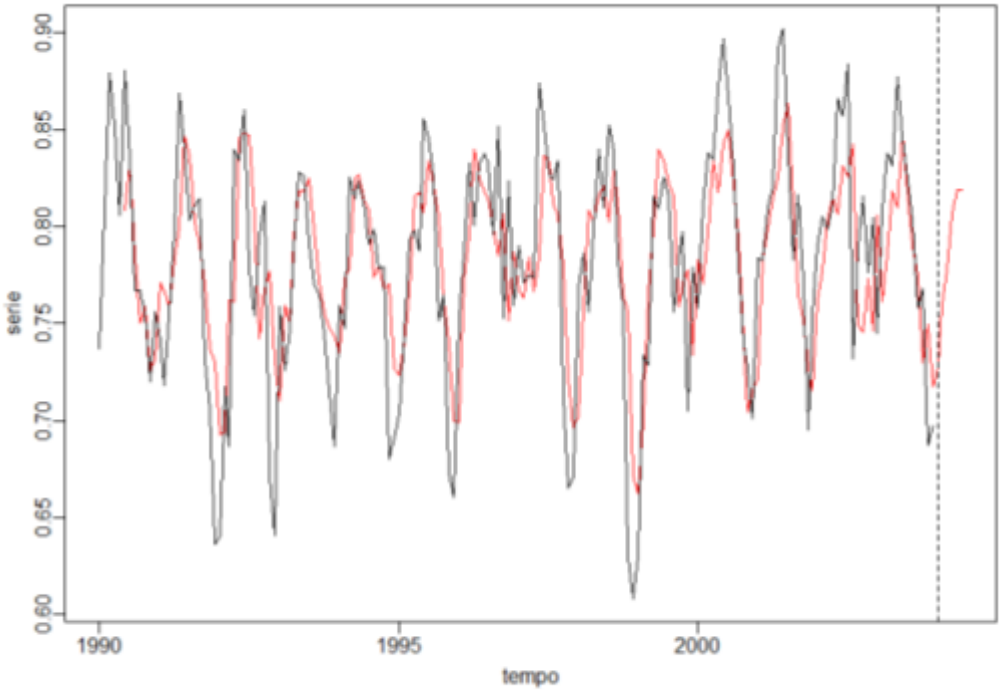

figure 8: BARMA model prediction

KARMA:

```
Estimate Std. Error z value Pr(>|z|)
alpha       1.0445     0.1012 10.3225   0.0000
phi1        0.4898     0.0655  7.4743   0.0000
phi2        0.0773     0.0753  1.0259   0.3049
phi3        0.0132     0.0747  0.1766   0.8598
phi4       -0.0825     0.0752  1.0970   0.2726
phi5       -0.2828     0.0649  4.3573   0.0000
precision  21.2784     1.2991 16.3799   0.0000
[1]
[1] Log-likelihood: 286.594
[1] Number of iterations in BFGS optim: 83
[1] AIC:      -576.7705  SIC:     -554.9028  HQ:      -579.333
[1] Residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-2.277001 -0.703306 -0.047636  0.001417  0.538619  2.867412
```

report 3: KARMA model

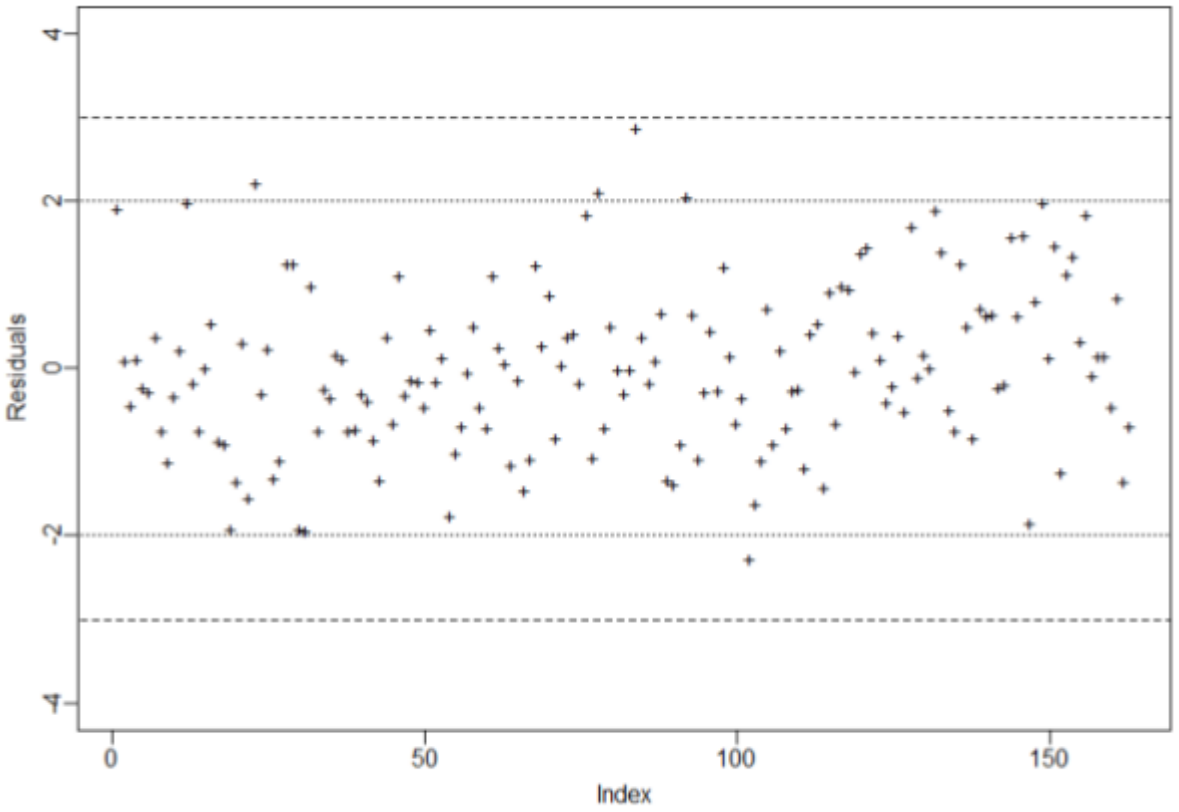

figure 9: KARMA model, residuals

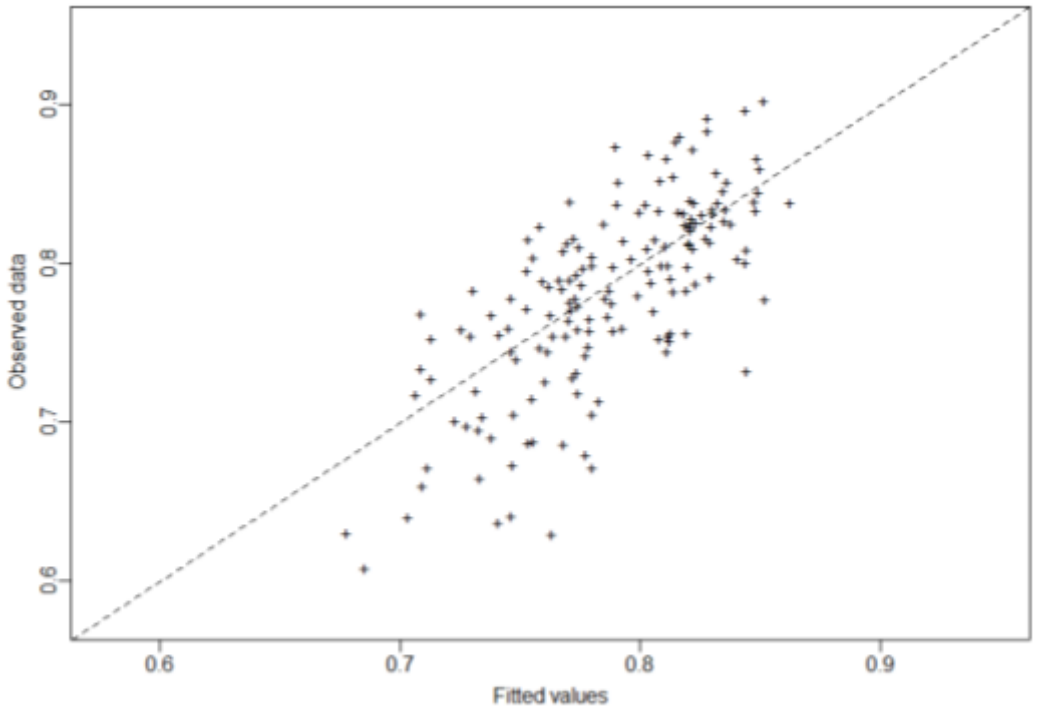

figure 10: KARMA model, residuals

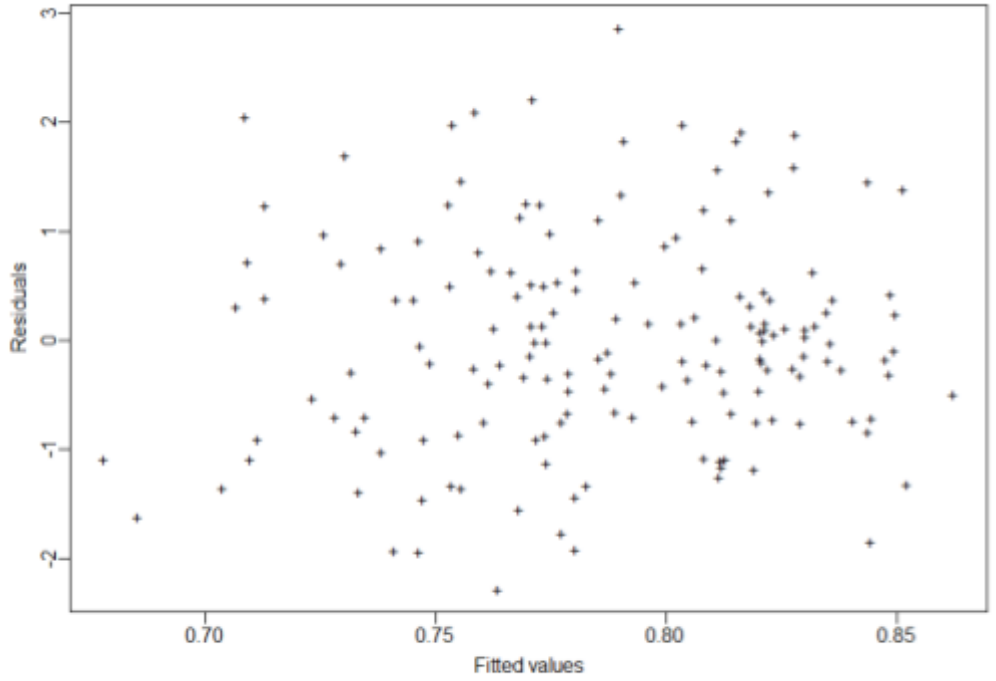

figure 11: KARMA model residuals

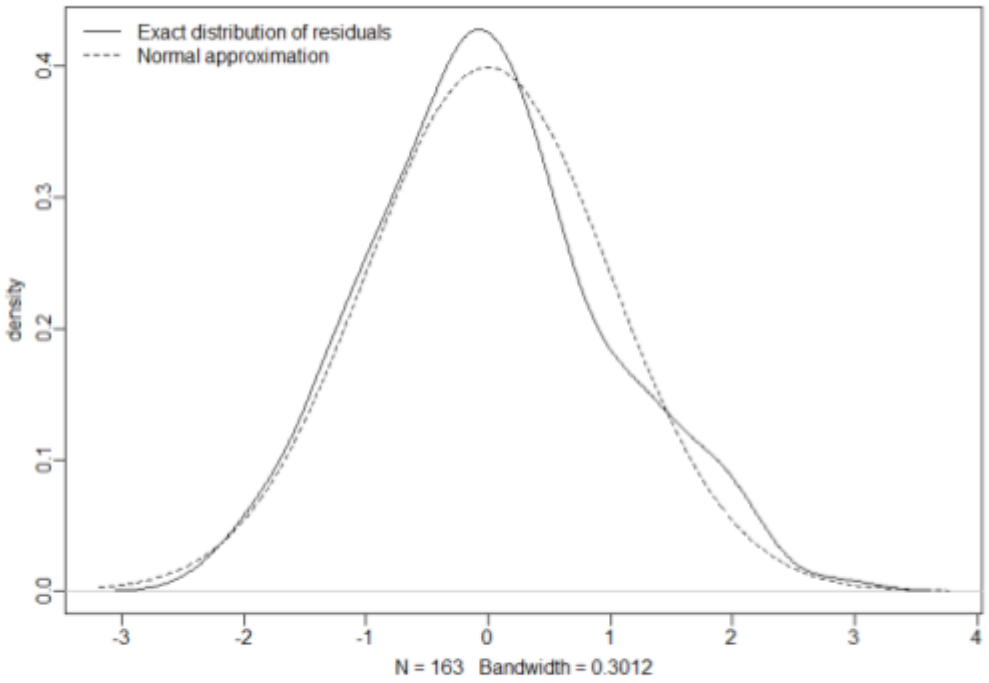

figure 12: KARMA model, residuals

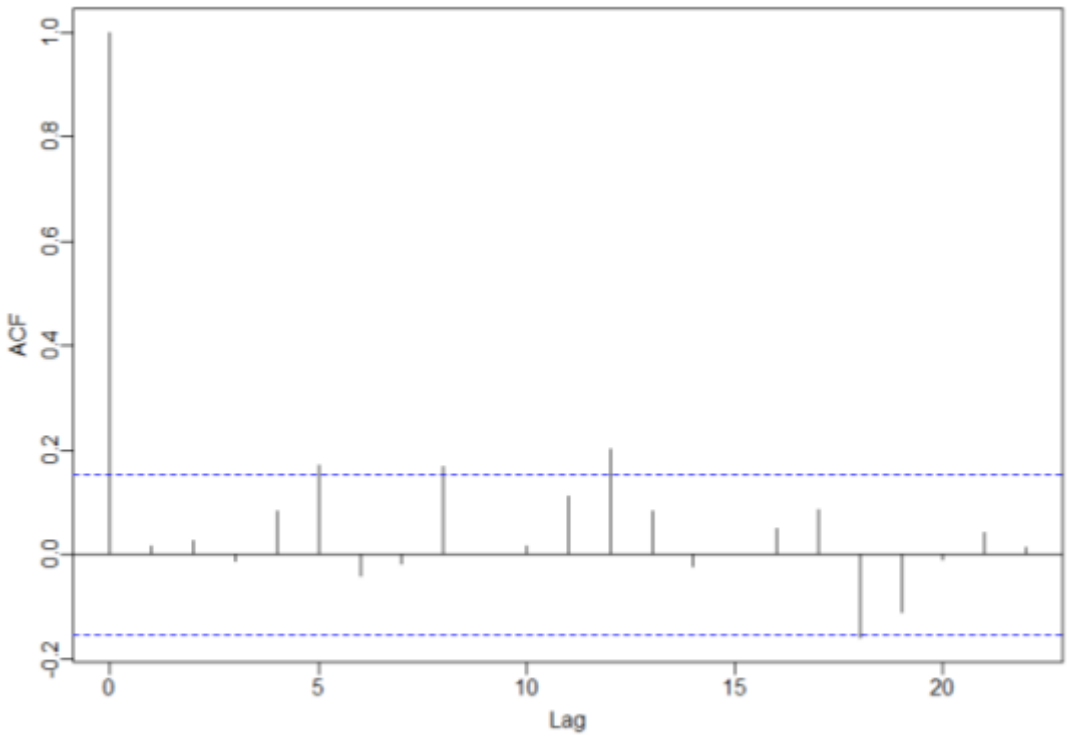

figure 13: KARMA model, ACF

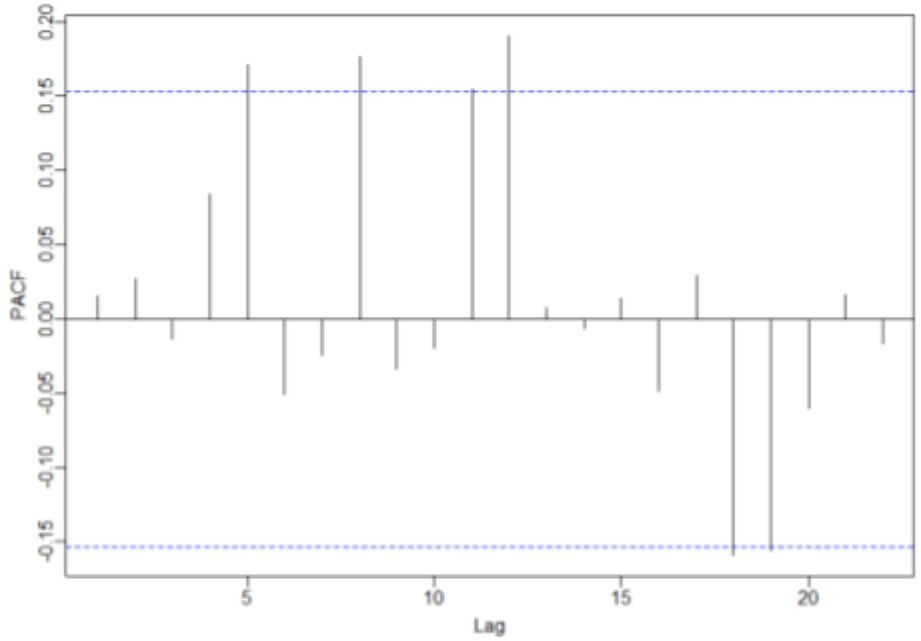

figure 14: KARMA model, PACF

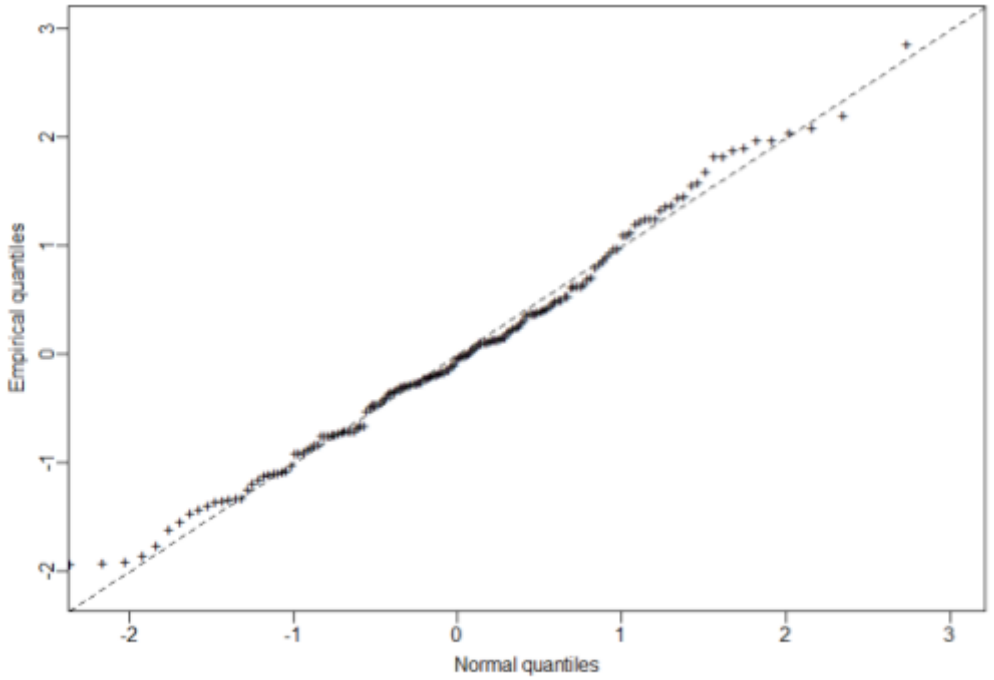

figure 15: KARMA model, residuals

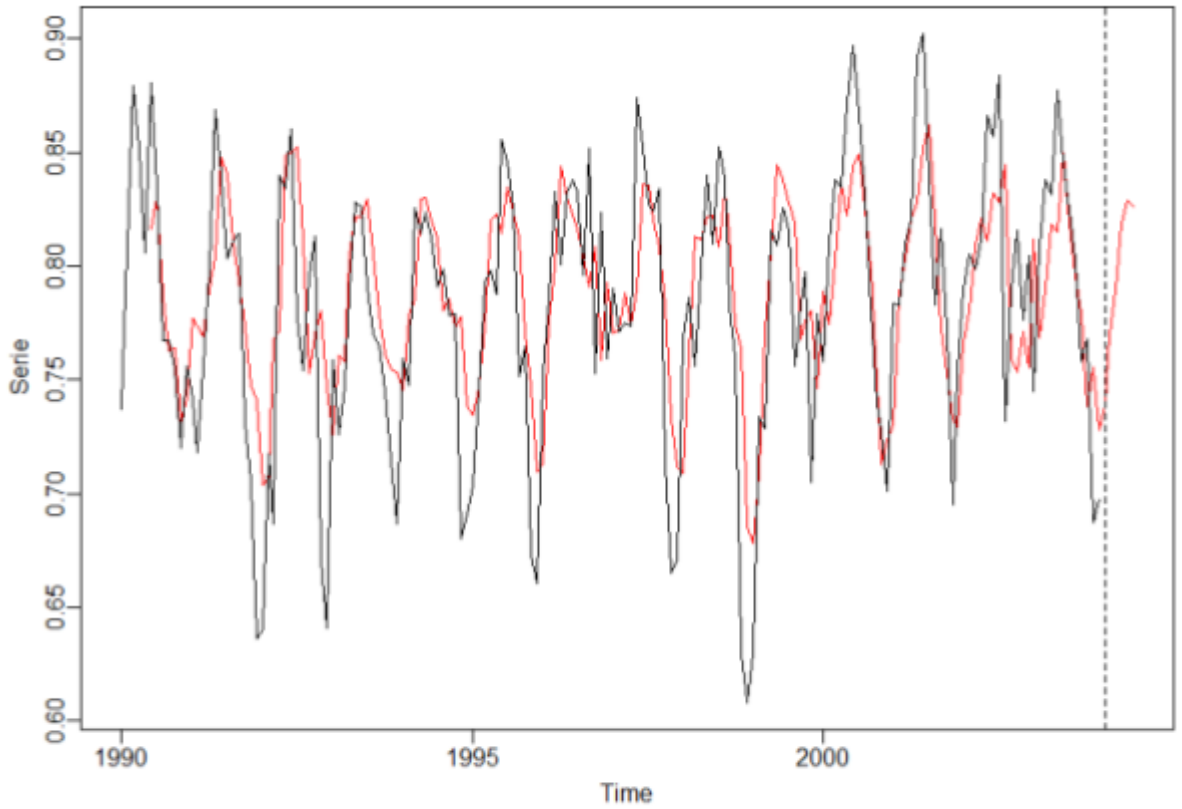

figure 16: KARMA model prediction

unemployment rate data set

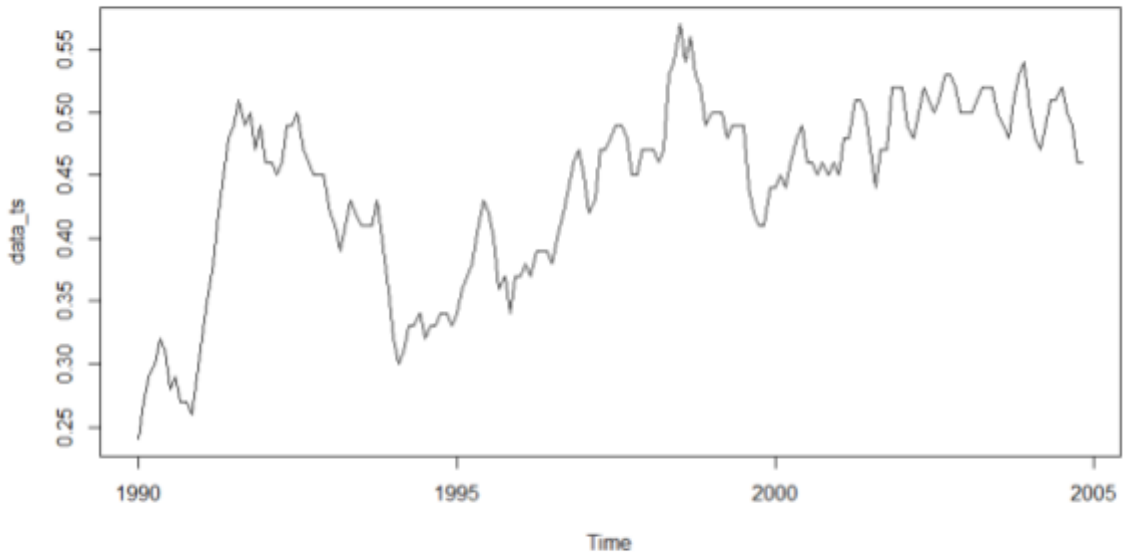

figure 1: Unemployment-rate dataset time-series

ARMA:

```
Series: train_data
ARIMA(1,1,3)

Coefficients:
          ar1     ma1     ma2      ma3
      0.0763  0.2380  0.3958  -0.4869
s.e.  0.1735  0.1514  0.1081   0.1181

sigma^2 = 0.0002862:  log likelihood = 447.91
AIC=-885.82   AICc=-885.45   BIC=-870.2



Ljung-Box test

data:  Residuals from ARIMA(1,1,3)
Q* = 2.2358, df = 6, p-value = 0.8968

Model df: 4.   Total lags used: 10
```

report 1: ARMA model

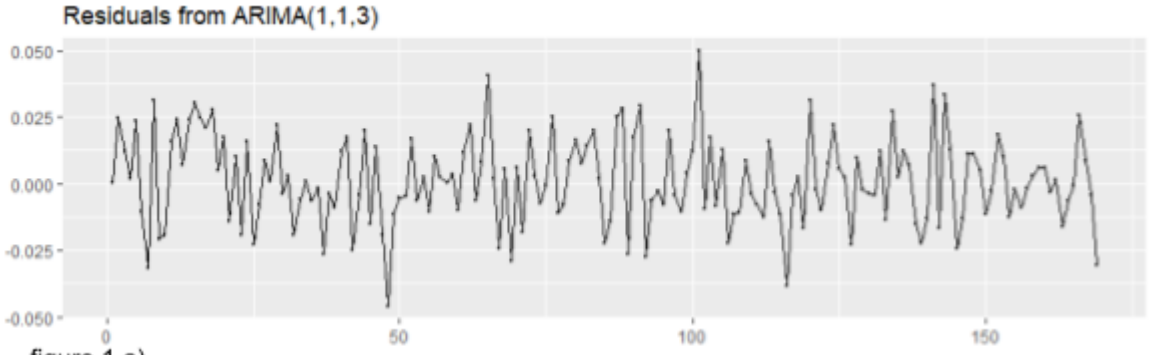

figure 1 a)

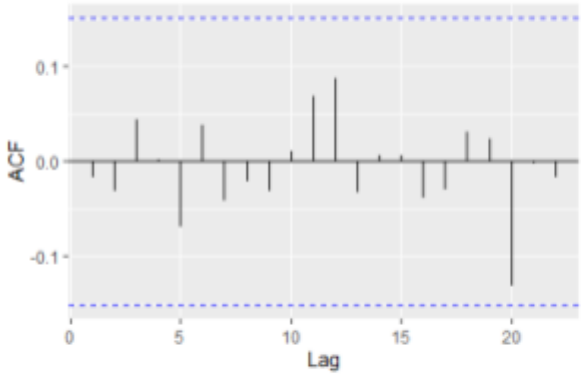

figure 1 b)

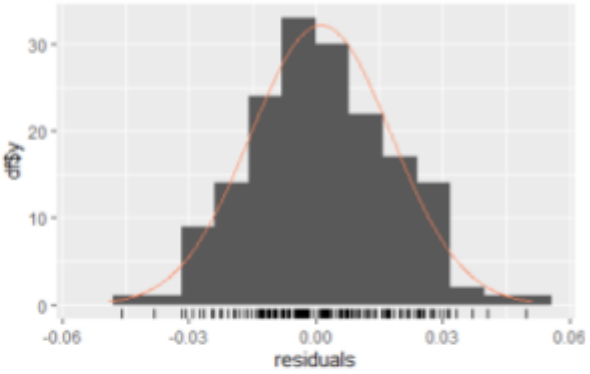

figure 1 c)

figure 1 a), b) and c) ARMA model, residuals

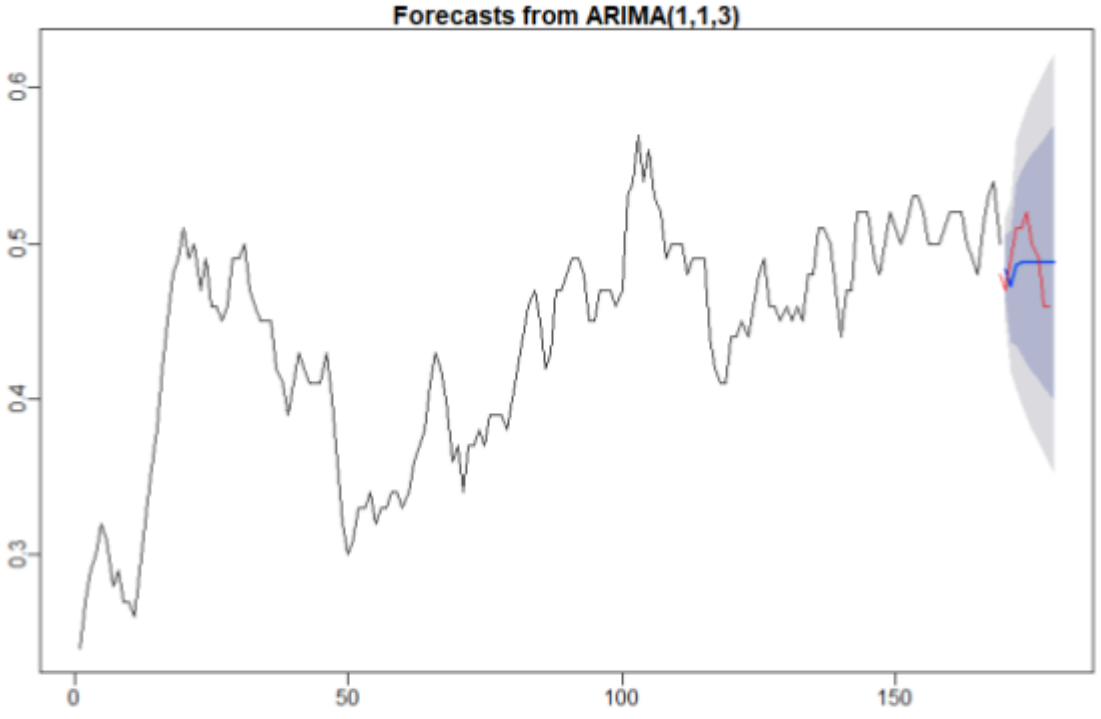

figure 2: ARMA model prediction

BARMA:

```
Estimate Std. Error z value Pr(>|z|)
alpha      -0.0082     0.0081  1.0199   0.3078
phi1        1.2125     0.0801 15.1340   0.0000
phi2       -0.0469     0.1212  0.3868   0.6989
phi3       -0.6007     0.1117  5.3793   0.0000
phi4        0.4908     0.1211  4.0517   0.0001
phi5       -0.1055     0.0787  1.3415   0.1798
precision 762.0410    84.0995  9.0612   0.0000
[1]
[1] Log-likelihood: 444.5942
[1] Number of iterations in BFGS optim: 142
[1] AIC:      -875.1883  BIC:     -853.279
[1] Residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-2.54592 -0.64789 -0.03926 -0.00856  0.69473  3.15261
```

report 2: BARMA model

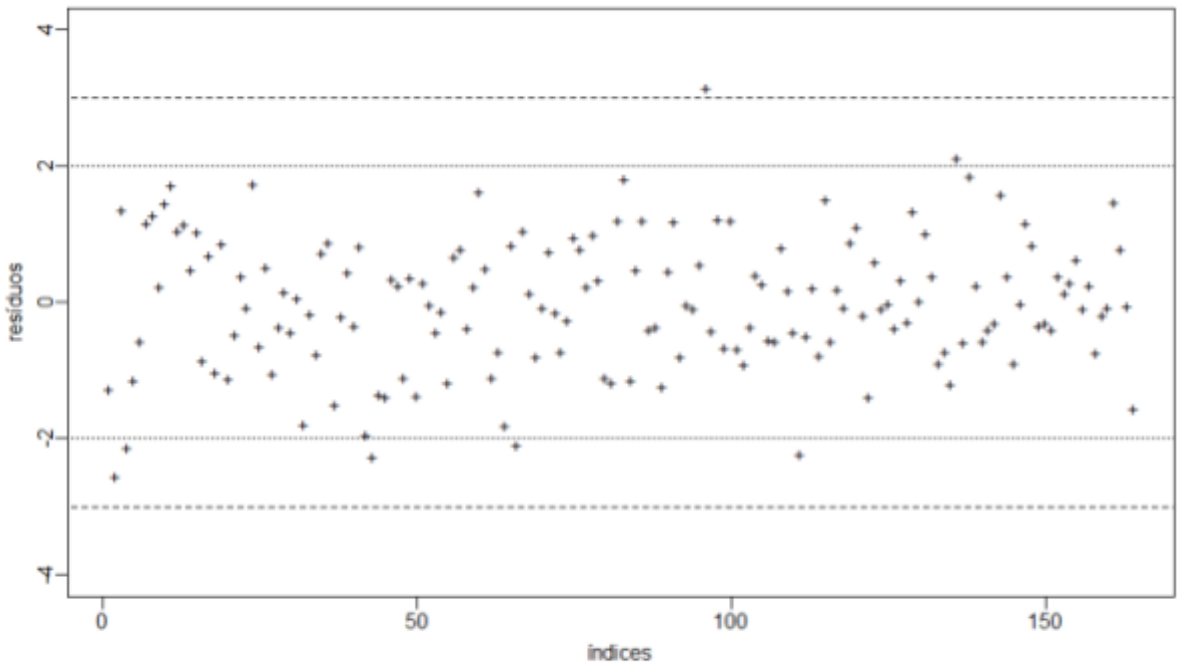

figure 3: BARMA model residuals

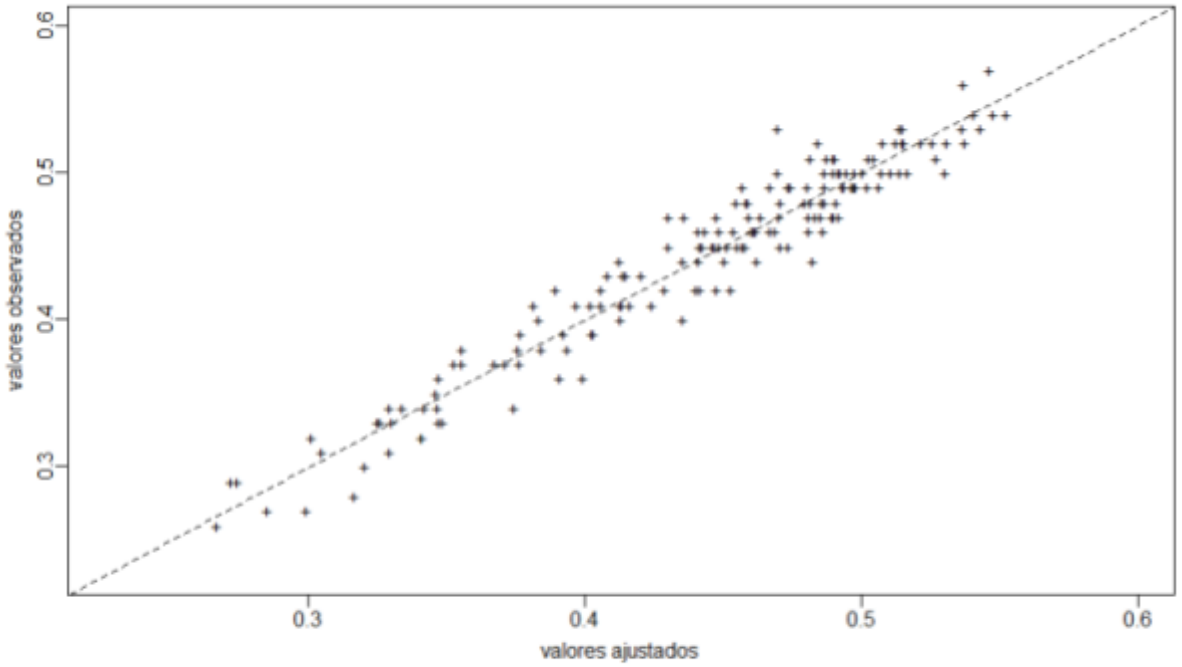

figure 4: BARMA model, residuals

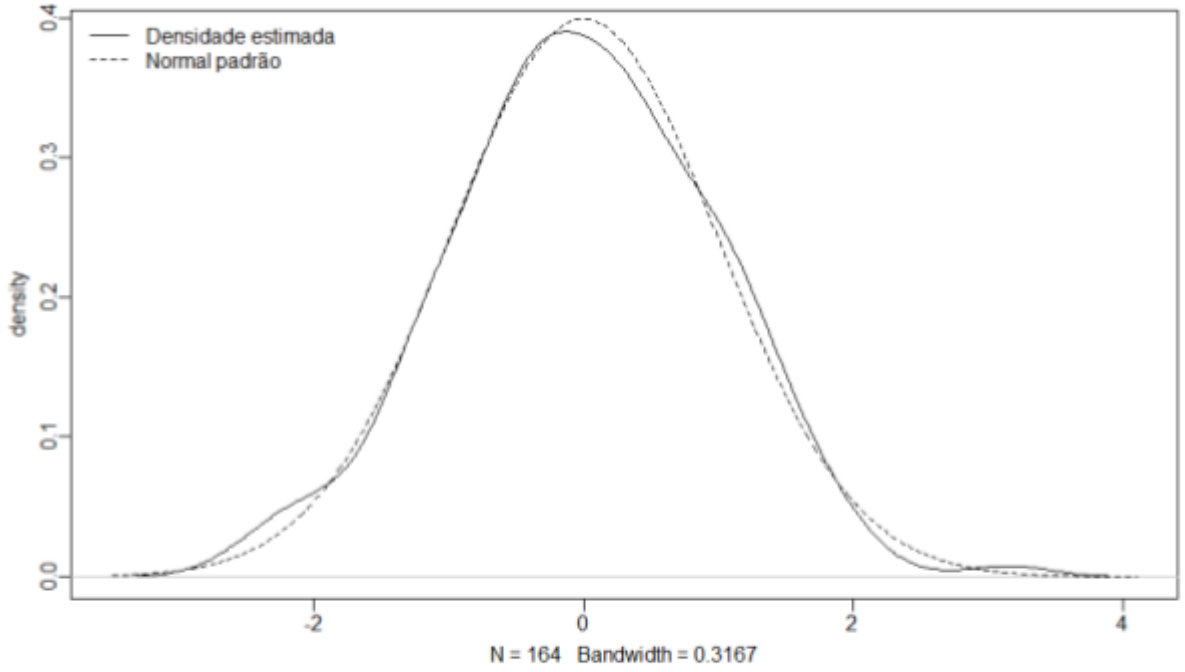

figure 5: BARMA model, residuals

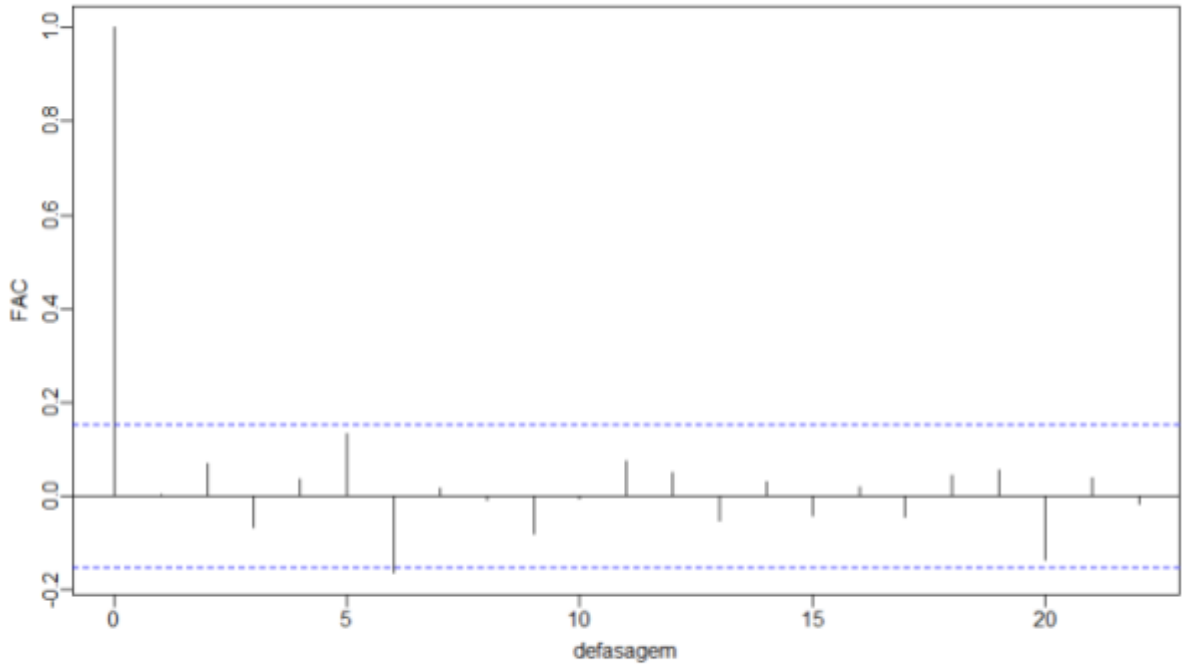

figure 6: BARMA model, ACF

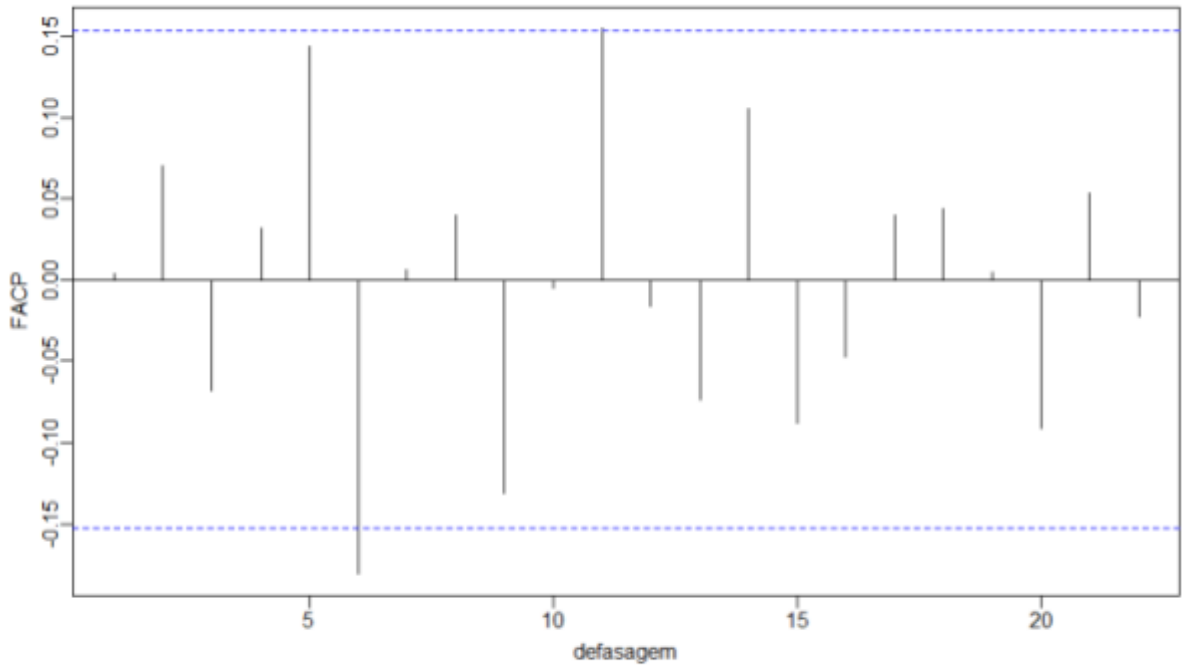

figure 7: BARMA model, FACP

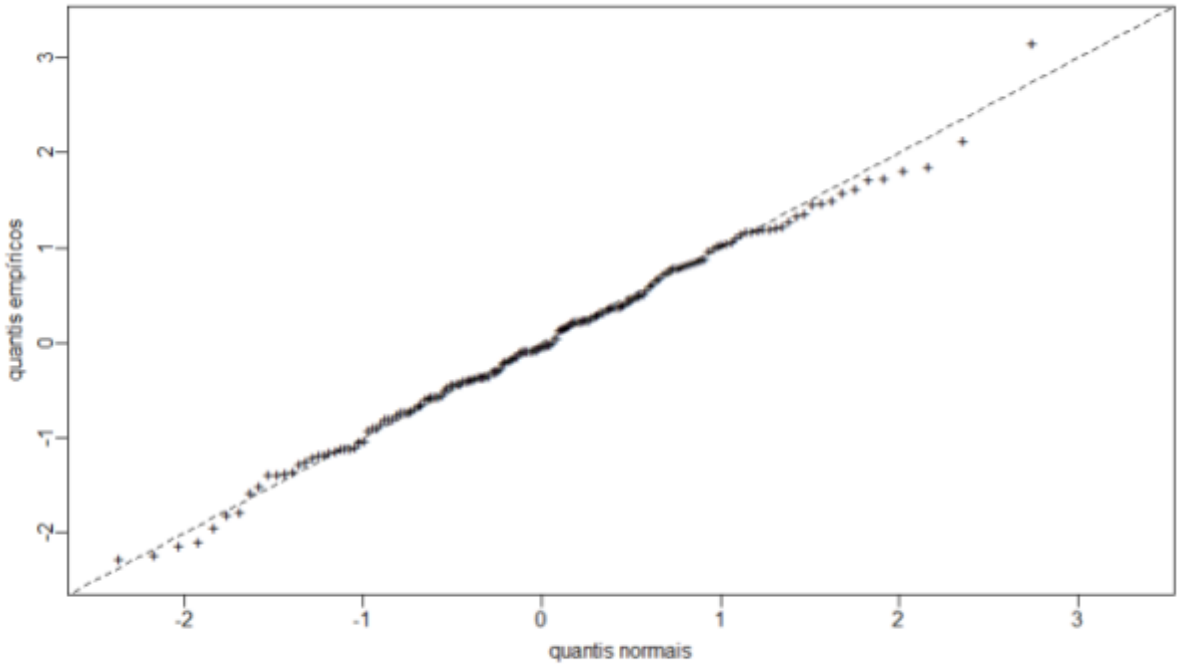

figure 8: BARMA model, residuals

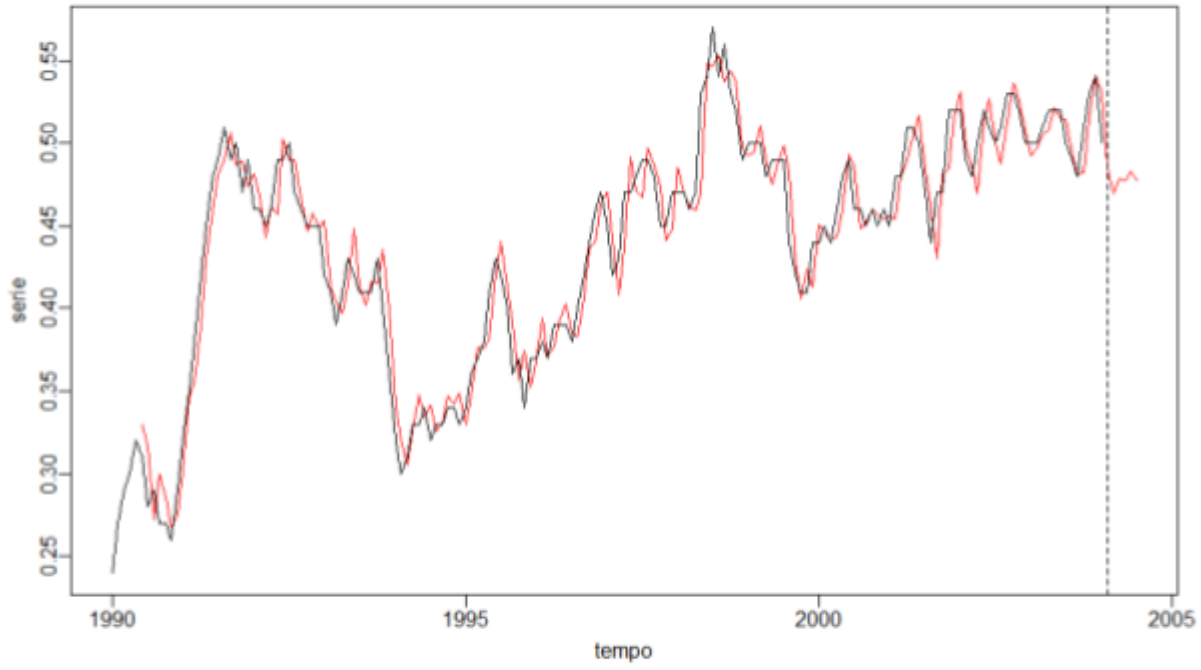

figure 9: BARMA model, residuals

KARMA:

```
Estimate Std. Error z value Pr(>|z|)
alpha      -0.0038     0.0088  0.4257   0.6704
phi1        1.2363     0.0759 16.2816   0.0000
phi2       -0.1302     0.1156  1.1262   0.2601
phi3       -0.5438     0.1069  5.0852   0.0000
phi4        0.5104     0.1153  4.4260   0.0000
phi5       -0.1368     0.0743  1.8397   0.0658
precision  25.0855     1.3846 18.1172   0.0000
[1]
[1] Log-likelihood: 417.283
[1] Number of iterations in BFGS optim: 142
[1] AIC:      -846.01    SIC:     -824.1007  HQ:      -848.5644
[1] Residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.96721 -0.59193 -0.10993  0.01632  0.58250  4.48075
```

report 3: KARMA model

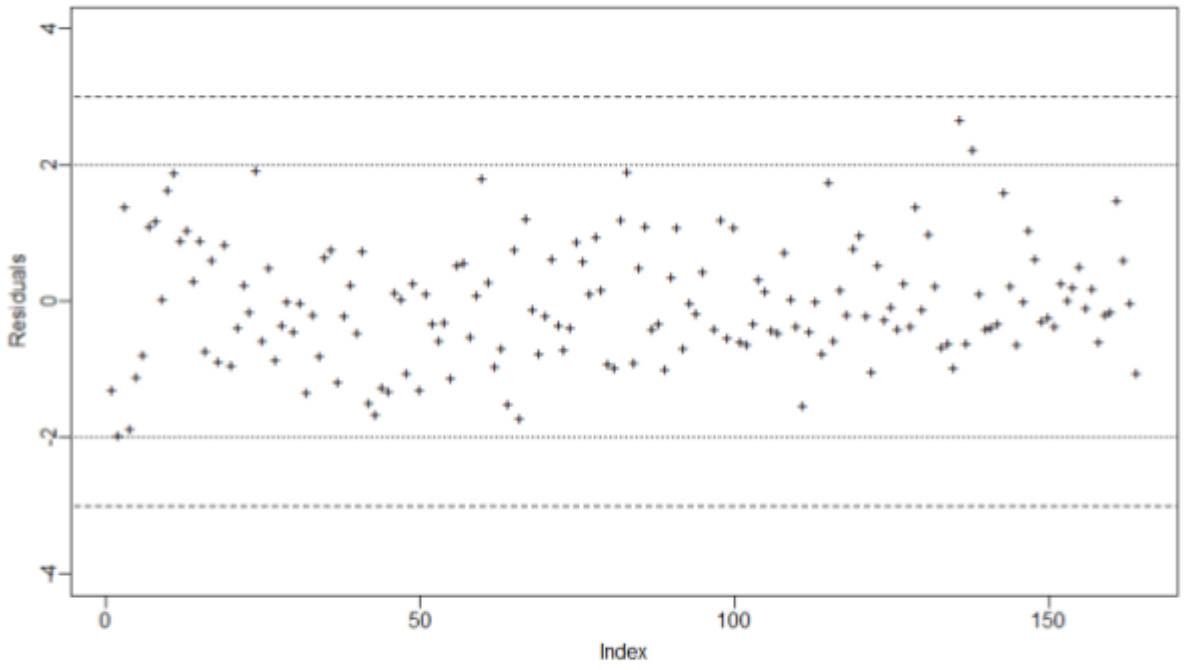

figure 10: KARMA model, residuals

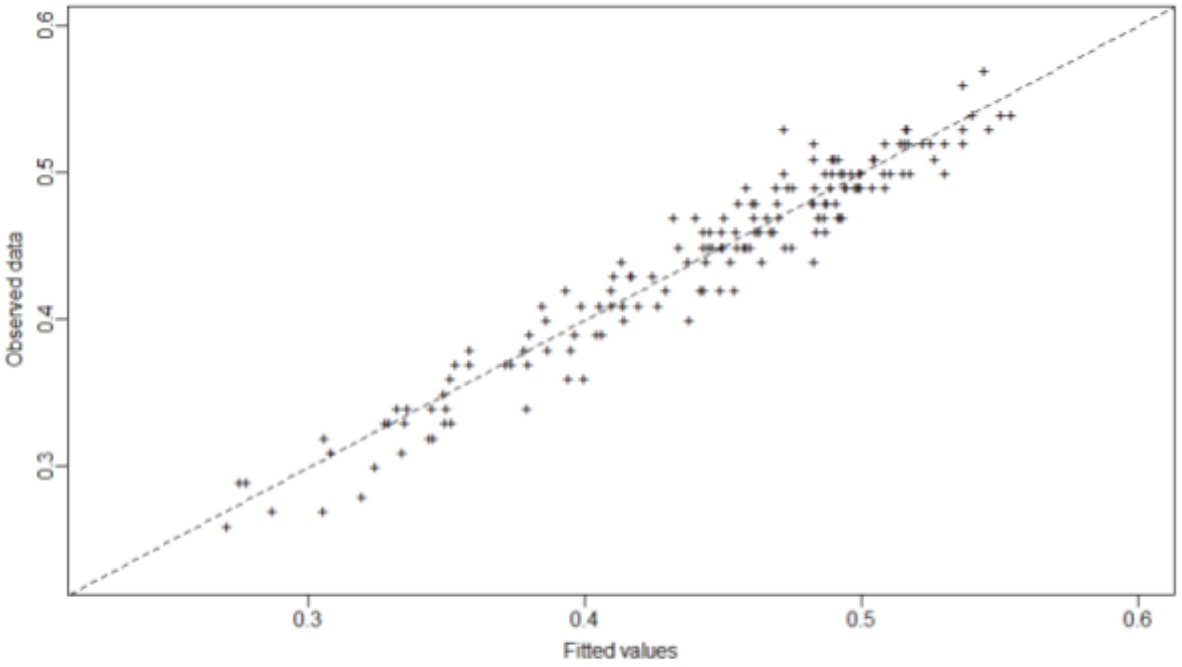

figure 11: KARMA model, residuals

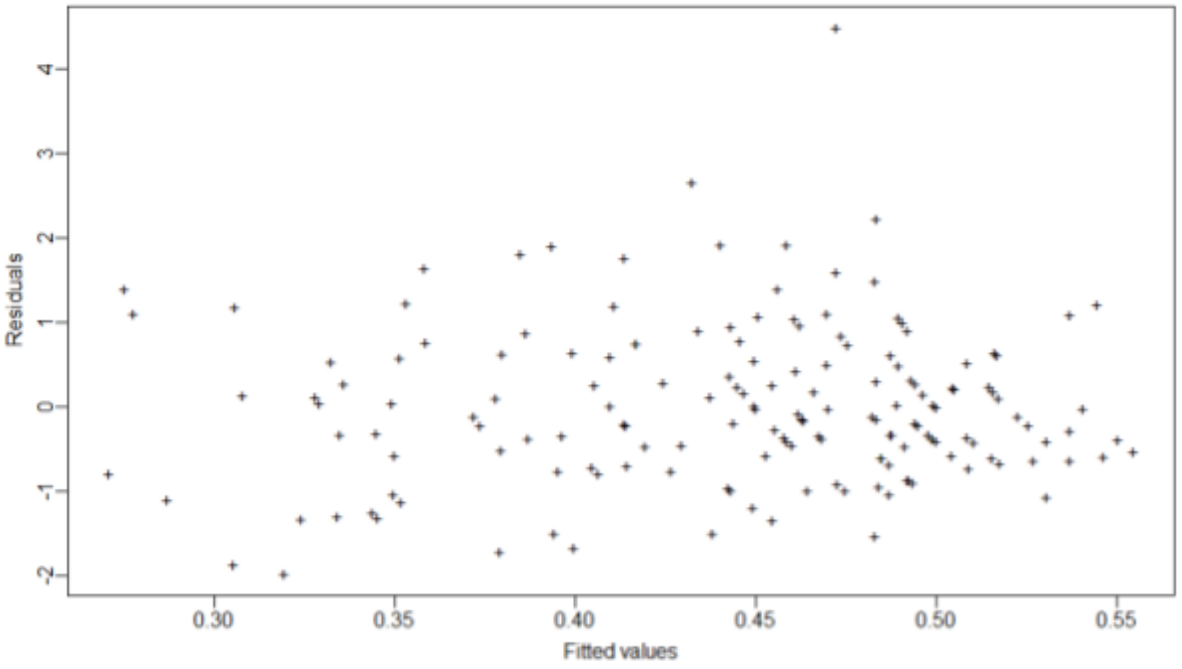

figure 12: KARMA model, residuals

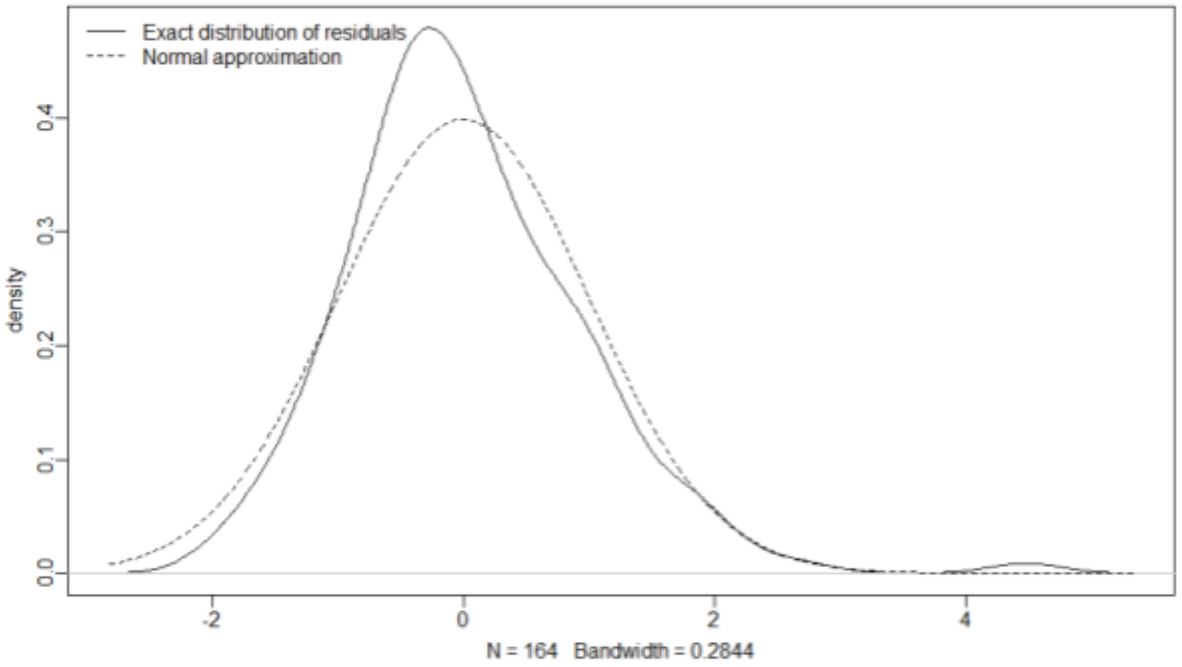

figure 13: KARMA model, residuals

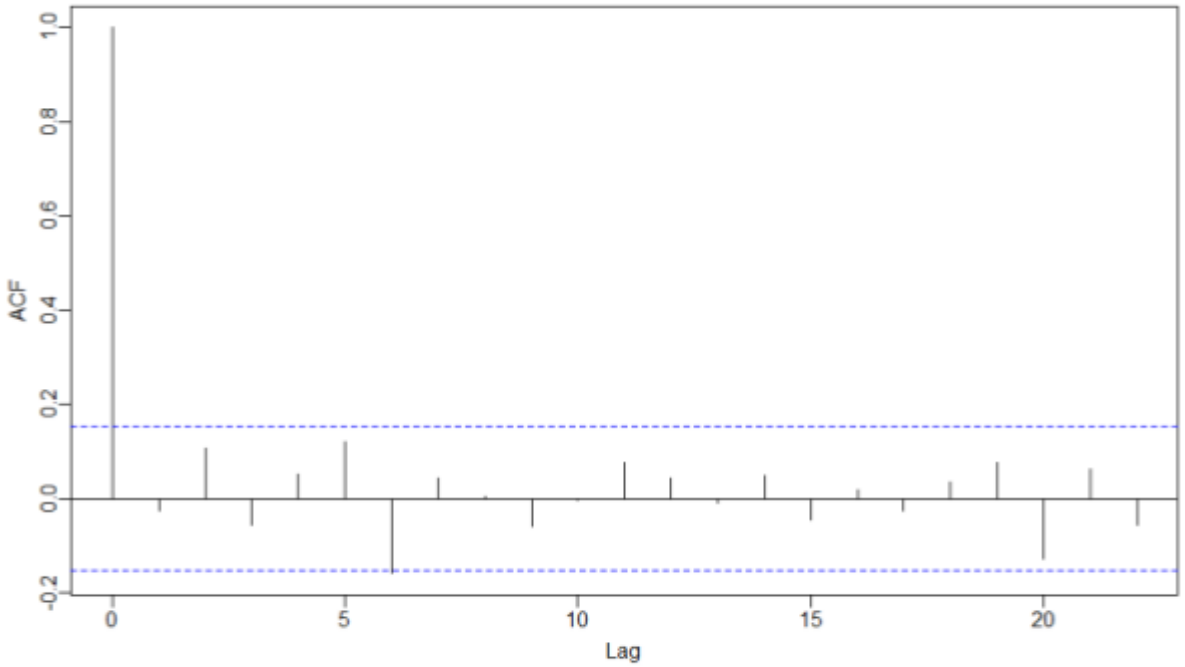

figure 14: KARMA model, ACF

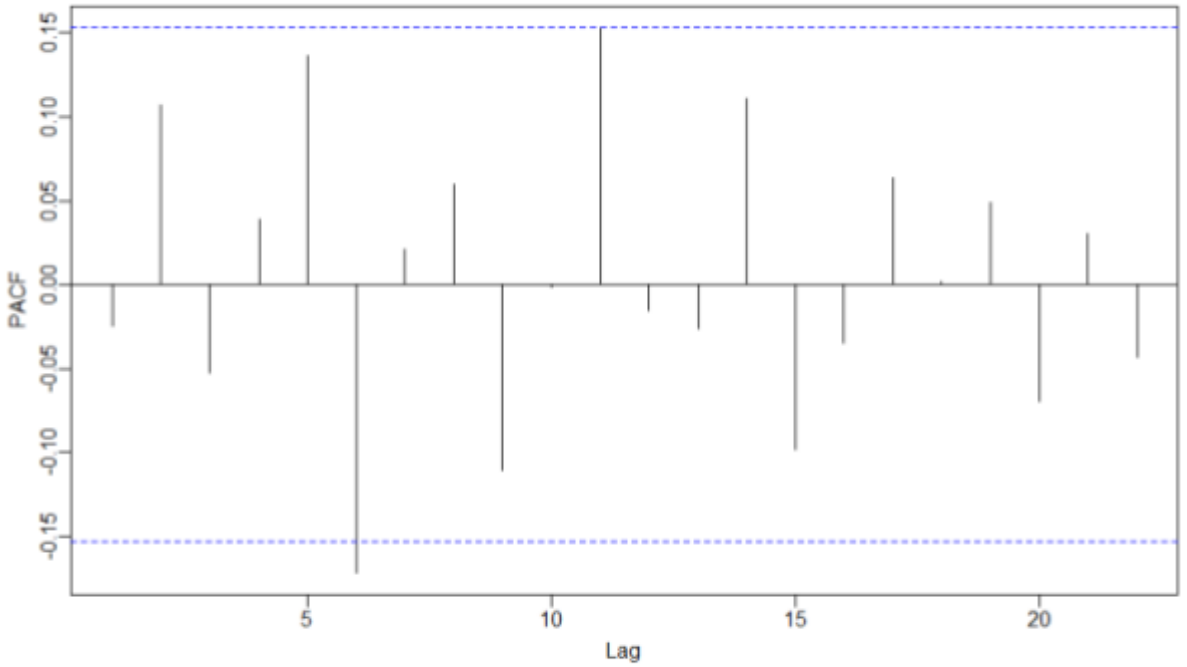

figure 15: KARMA model, PACF

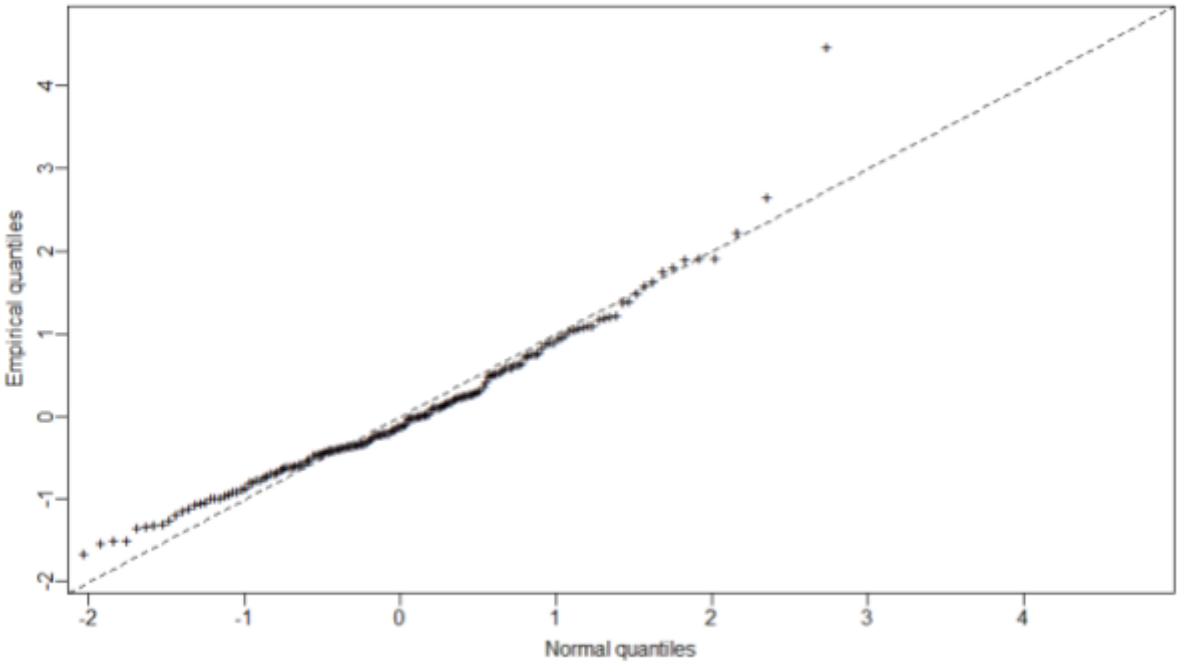

figure 16: KARMA model, residuals

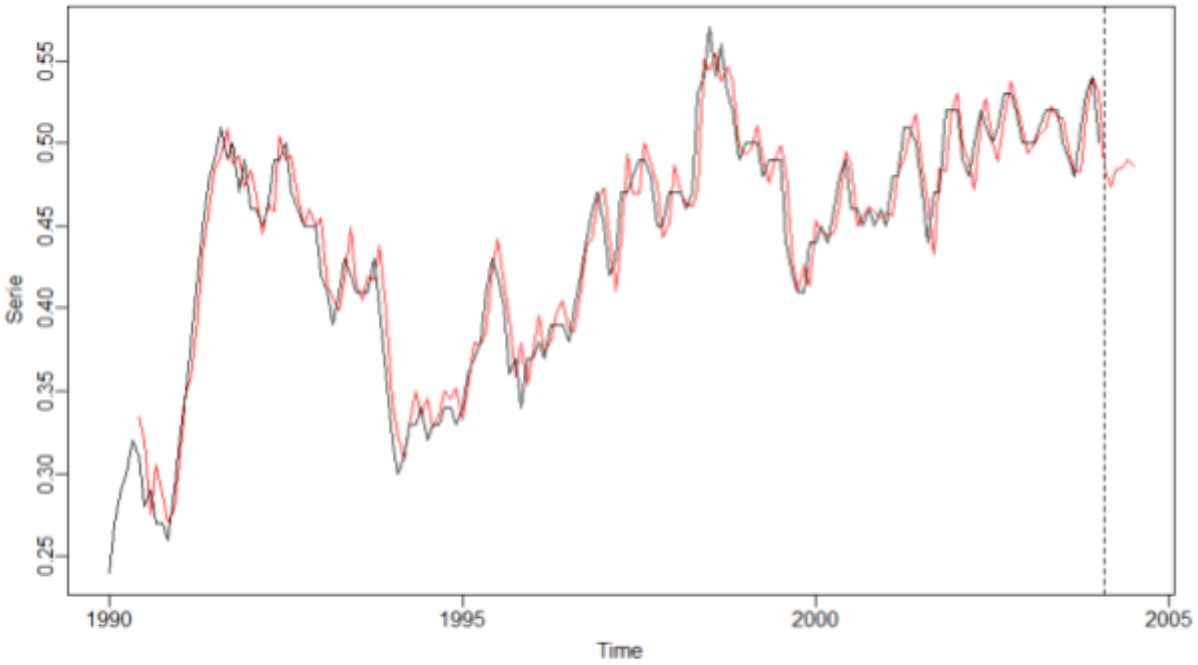

figure 17: KARMA model prediction

Discuss:

the relative humidity data set

The ARMA model demonstrates a reasonable fit to the data, as evidenced by its AIC, BIC, and log-likelihood values. However, the presence of autocorrelation in the residuals, as implied by the Ljung-Box test in ARMA-model 1, indicates that the model may be failing to capture some aspects of the data, suggesting potential room for improvement.

The forecasted outcomes for the ARMA model, depicted in figure 2, exhibit a rise and fall pattern similar to the actual data. However, an exact match isn't achieved, which suggests that while the model is able to capture the overall pattern, there is still room for enhancing its precision.

The auto arima function, which is used in our analysis, automatically estimates the AR and MA parameters for the model and applies differentiation to make the data stationary. This stationarity ensures a constant variance, effectively removing any existing trends or seasonality, such as those seen in the unemployment dataset. Should we decide to use another function, it would be incorporated before fitting the ARMA model.

When comparing the ARMA model with the BARMA model, the log-likelihood value serves as a good indicator of fit, with higher values signifying better fit. In our case, the BARMA model's log-likelihood of 299.53 (as per report 2 BARMA model) is greater than the ARMA model's log-likelihood of 290.6 (as per report 1 ARMA model), hinting at a better fit for the BARMA model.

AIC and BIC are measures of the relative quality of statistical models for a given dataset, with lower values indicating a better fit. Here, the BARMA model (report 2 BARMA model) has lower AIC and BIC values (-585.05 and -563.19, respectively) than the ARMA model (-569.21 and -550.47) (report 1 ARMA model), further supporting the notion that the BARMA model is a superior fit.

Analyzing the residuals can provide insights into the model's fit. We desire residuals that are normally distributed and centered around zero. The residuals for the BARMA model appear to be centered around 0.09502 (report 2 BARMA model), which is favorable. The residual plots in figures 3, 4, and 7 suggest a normal distribution (figure 7) and slight centering around zero and randomness (figure 3). However, some pattern is still evident, though not significant, as depicted in figures 4 and 3. For comparison, the ARMA model's residuals (figure 1c) also show normal distribution but appear to have some autocorrelation, considering the improved prediction results from the BARMA model.

The Ljung-Box test, which assesses the independence of residuals (i.e., the absence of autocorrelation), reveals a p-value of 0.01764 in the ARMA model (report 1 ARMA model). This value, being less than 0.05, indicates significant autocorrelation in the residuals, suggesting the model might not be a good fit.

The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots, depicted in figures 5 and 6 respectively, show almost no significant spikes, generally indicating no autocorrelation in the residuals. This suggests the model has adequately captured the time-dependent structure in the data, leaving minimal information in the residuals that could be used to improve the model's predictions.

Comparing the BARMA and KARMA models, the log-likelihood, AIC, and BIC values all suggest that the BARMA model performs better. Visual inspection of the residuals in both models, as represented in figures 3-4-7 (BARMA model) and 9-10-11-12-15 (KARMA model), supports this. While the residuals for the KARMA model in figure 9 (KARMA model) are randomly spread and lie within the significance coefficients, the BARMA residuals, shown in figure 3 (BARMA model), perform better. Furthermore, in figure 10 (KARMA model), the points do not follow the diagonal, similar to the BARMA residual figure 4 (BARMA model), whereas figure 11 (KARMA model) reveals no pattern.

The ACF and PACF in figures 13 and 14 (KARMA model) exhibit some sort of pattern, which might negatively impact its predictive accuracy compared to the BARMA model. This is reflected in the prediction results depicted in figure 16 (KARMA model). The results also show a close similarity in the predictions between the BARMA and KARMA models due to certain resemblances in their residuals. However, the BARMA model still delivers the most accurate prediction among the three models.

In conclusion, based on our evaluation of the ARMA, BARMA, and KARMA models using various statistical measures and visual analysis of residuals, the BARMA model appears to provide the best fit and most accurate predictions for our dataset.

unemployment rate data set

The ARMA model, similar to the previous dataset, has been applied. The results from (report 1) are as follows:

The estimated coefficients for the AR and MA components of the model are provided, along with their standard errors (s.e.). The model has an autoregressive (AR) term of order 1 (ar1 = 0.0763) and moving average (MA) terms of order 3 (ma1 = 0.2380, ma2 = 0.3958, ma3 = -0.4869).

The estimated variance of the residuals is 0.0002862. A smaller value indicates a better fit.

The log-likelihood of the model is 447.91. In this case, the AIC is -885.82, the AICc is -885.45, and the BIC is -870.2. Ljung-Box Test. The test statistic (Q*) is 2.2358 with 6 degrees of freedom, and the p-value is 0.8968. A high p-value (generally above 0.05) indicates that we fail to reject the null hypothesis that the residuals are independently distributed, suggesting that the model adequately captures the underlying structure in the data.

The residuals have been shown to be normally distributed in (figure 1c), which is good for the model. The Autocorrelation Function (ACF) in (figure 1b) shows no significant spikes, indicating no autocorrelation in the residuals. However, despite the good results of both the residuals and the log-likelihood, the model does not perform well in terms of prediction accuracy, as seen in figure 2.

The ARMA model, as reported in (report 1), has a log-likelihood of 447.91, which is higher than the BARMA model's log-likelihood of 444.5942 (report 2). This suggests that the ARMA model might be a better fit for the data.

The ARMA model also has lower AIC and BIC values (-885.82 and -870.2, respectively) compared to the BARMA model (-875.1883 and -853.279, respectively) as shown in (report 1) and (report 2). This further suggests that the ARMA model might be a better fit for the data, considering both the goodness of fit and the complexity of the model.

In the BARMA model, the residuals have a mean of -0.00856, suggesting a slight systematic underestimation by the model. The minimum and maximum values are -2.54592 and 3.15261, respectively, indicating larger errors. The 1st Quartile, Median, and 3rd Quartile provide additional insights into the spread and potential skewness of the residuals, as shown in (figure 3 and report 2).

The residual plots in (figures 8, 5, and 4) indicate that the residuals are normally distributed but exhibit significant patterns in the data, resulting in a loss of information.

Comparing KARMA and BARMA, based on the criteria in reports 2 and 3, the BARMA model seems to perform better. It has a higher log-likelihood, lower AIC and BIC values, and higher precision.

The residual plots in figures 10, 11, 12, 13, and 16 show that the KARMA model also exhibits some normality and is centered around zero, but it displays more patterns compared to the BARMA model, which can affect the model's performance.

The ACF and PACF plots in the KARAMA model, as seen in figures 14 and 15, are identical to the BARMA model. This explains why the KARMA model's prediction, as seen in figure 17, is also identical to the BARMA model's prediction.

In conclusion, the performance of the KARMA and BARMA models is almost identical. However, the KARMA model exhibits more noticeable patterns compared to the BARMA model. Despite this, on average, the ARMA model performs the best among them.

.