STAT-485 & 685 Applied Time Series Analysis
Project

Executive Summary

Description:

We are given a record of hourly temperature checks of equipment usage over a 243 hour interval with the objective of estimating the probability of damage to the equipment to allow the company of the equipment to determine whether or not they should spend to replace the current equipment or not.

Goals of the Report:

The objective of the report is to provide a decision for whether or not the high temperatures that the equipment operates at is suitable for the equipment to not be damaged over an interval of 243 hours from a sample of 243 observations.

Key assumption is that the equipment is most likely to take damage when used over 72 hours in high temperatures (45+ degrees Celsius).

Findings:

We find that the probability of the equipment being damaged when in high temperature for over 72 hours is approximately 0.23. This value would suggest that there is a 23% chance that the equipment would be damaged upon extensive usage.

Conclusion:

Based on the findings on the report, it is recommended that the company does not buy new equipment and retain the current equipment.

1.Introduction

This project aims to demonstrate the application of time series analysis on real-life data. The data that our group is analyzing contains the data of the temperature of a piece of equipment that is being recorded every hour. This data contains 243 observations and 2 variables which are hour and temperature (celsius).

Our group applied a structured methodology to complete this project. We start by coming up with a problem that we want to answer through this project. The main goal of this project is to estimate the probability of damage to the equipment. We wanted to demonstrate our understanding of concepts in time series analysis and apply it to the equipment data to come up with an answer as to whether the company should replace the piece of equipment. We decided that for this project we will assume a machine that goes over a high temperature (above 45 degree Celsius) for over 72 hours (3 days) in an interval of 243 hours will likely be broken.

With the purpose to come up with an answer to the project's objective. Our group starts by understanding the data. The initial step is to determine whether our data is a stochastic process or a deterministic process. Followed by data preparation we do it by differentiating our data. Next, by using the concepts we learned about time series analysis as a reference, our group fits the data to different time series models and applies various model selection methods. In addition to that, our group also ran simulations to come up with an estimation of the probability that the equipment could be damaged from high temperatures, followed by some explanation and reasoning behind our conclusion.

2. Analysis

2.1 Pre Data Analysis

2.1.1 Stationary / Non Stationary Data

In order to work with the data, we must first make observations and apply any necessary change to it if needed. The first condition is whether or not the data is stationary, as the time series applications can only be used if the data is stationary. Looking at the time series plot of the data (Figure 1.), we observe that the data's mean is not constant and its standard deviation fluctuates, as denoted by the different peak heights. Therefore, we conclude that there is a possibility of the data not being stationary. With the data being non-stationary, we now have to convert the data into being stationary before proceeding to find a model.

2.1.2 Trend of the Data

We apply techniques to the values in our data to make it stationary. We first must know which trend the graph shows to apply the correct technique. We determine that the data is stochastic

with evidence from Figure 1. as there is no general trend and there are some hour intervals where the values are close to each other, meaning there is no change in temperature for a couple of hours. We can apply the appropriate method to our data. We apply backward differencing to the data set to make it stationary. Figure 2. shows a much better graph of the data after it has been applied with backward differencing. Now that the data is stationary, we can now proceed with finding a model for it. Since we applied backward differencing, we will be looking for a certain ARIMA (p,1,q) model that is most suitable for the data. Since we did one backward differencing, the value of d is 1.

2.1.3 Assumptions and Limitations

One thing to consider before proceeding into the next steps. Our first assumption is that at this stage, our data is assumed to be stationary now after applying backward differencing. This statement is to ensure that we can use the tools to find our ARIMA model and other important information.

2.2 Model Estimation

2.2.1 Box-Jenkins method

As the Box-Jenkins method is a widely known approach, we assume that our data is stationary, and we will identify the order of the ARIMA model by comparing AIC and BIC of fitting potential models.

2.2.2 Autocorrelation Function(ACF) and Partial Autocorrelation Function(PACF)

The sample ACF graph shows that the ACFs are really significant at lag 1, and 2, and moderately significant at lag 3, 6, and from 10 to 16, and have a gradual decrease over time rather than cutoff at certain lag. We assume that there is an AR component, but still need to study the ACF around lag 12. The sample PACF shows significant spikes at lags 1 and 2 and small spikes after that, although there are spikes at different lags. We might need MA because of the spikes at lags 6,7, and 14, but we assume the order of AR to be 2. To check the ACF around lag 12, we might need to consider AR at lag 10. However, we are focusing on the temperature of equipment, and considering lag 10 would be unnecessary because studying the temperature of the 10-hour gap might not be meaningful. For the significant values at the lags that are not 1 or 2 in the PACF graph, the values would be significant in 5 % of chance, so we assume that we can ignore them this time. Thus, we expect our model to be ARIMA(2, 1, 0) based on ACF and PACF.

2.2.3 Systematic Model Selection(AIC and BIC)

From the EACF table provided for differentiated data(Figure 5), we assume that the models could be ARIMA(2,1,0), ARIMA(2,1,1), ARIMA(2,1,2), ARIMA(0,1,3), and ARIMA(1,1,3). The minimum AIC among these models is 698.3832 and the minimum BIC is 715.8279 for

ARIMA(2,1,2). The second smallest for AIC and BIC are 705,6563 and 716.1231 which are the values of ARIMA(2,1,0).

2.2.4 Model Selection(autocorrelation)

By comparing the autocorrelation of each model and ACF, we can see that ARIMA(1,1,3), ARIMA(0,1,3), and ARIMA(2,1,1) do not align well with the original data. Although ARIMA(2,1,2) seems closer to the original one, the autocorrelation of ARIMA(2,1,0) also fits well and the original values are inside the confidence interval of the autocorrelation.

2.2.5 Result

With the analysis thorough AIC, BIC, and autocorrelation, either ARIMA(2,1,2) or ARIMA(2,1,0) would be the candidates for our model. Applying the principle of parsimony, we prefer choosing ARIMA(2,1,0) which is a simpler model. As the mean value for first-order differentiation is 0, the ∇Yt would be shown below.

$$\nabla Yt - 0.1 = 1.09\Phi_1(\nabla Yt-1 - 0.1) - 0.39\Phi_2(\nabla Yt-2 - 0.1) + et$$

2.3. Model Diagnostic

2.3.1 Residual Analysis

In the ACF of residuals for ARIMA(2,1,0) we observed that the sample autocorrelation value for lag 6 is outside of the 2 standard deviation line from zero. This does not satisfy the condition for white noise, for that reason, we are going to compare it with the other models. To further compare all the models we compare the residuals of the model using the diagnostic tools.

We compare the ACF residual plot (Figure 7) of the 5 models. In the ACF of residuals for ARIMA(2,1,1), we observed that the sample autocorrelation value for lag 6 is outside of the 2 standard deviation line from zero. In the ACF of residuals for ARIMA(0,1,3) we observed that the sample autocorrelation value for lag 12 is outside of the 2 standard deviation line from zero. In the ACF of residuals for ARIMA(1,1,3) we observed that the sample autocorrelation value for lag 6 is outside of the 2 standard deviation line from zero. The ACF of residuals for ARIMA(2,1,2) we observed that the sample autocorrelation value for lag 12 is outside of the 2 standard deviation line from zero. The other models also have autocorrelation values that are outside of the 2 standard deviations from zero, so we are still going to proceed with ARIMA(2,1,0).

Another plot that we are looking at is the residual plot over time and the residual vs fitted value plot. When comparing the residual over time plots and residual vs fitted value for all the models, the distribution of the residuals looks similar to each other. For all models, we observed the points are scattered over the plot and do not indicate any trend.

Normality assumption can be tested by using histogram and QQ plot. From observing all the histograms for all five of our models we can see that the distribution of the residuals are all similar. The histogram is approximately centered around zero, symmetric, and even though it is not perfectly bell-shaped, we can still assume that the distribution of the residuals based on the histogram is similar to a normal distribution.

From the QQ plots of the five models, we observed that the residuals are close to the line. Even though there are deviations from the line at both ends, the residuals at the center are relatively close to the line, with this we can assume that the normality assumption of the residuals from observing the QQ plots is acceptable.

2.3.2 Result from residual analysis

In conclusion, from the residual analysis that we did, the residuals from all the models can be considered normally distributed and the residuals do not show any pattern in both residual over time plot and residual against fitted value plot, but from checking the ACF of residual, each of the five models have autocorrelation that is slightly outside of two standard deviations from zero. This could indicate that the independence assumption may be violated. Since all the models are similar to each other based on normality and independence, we are still going to choose ARIMA (2,1,0) as our best model.

2. 4 Calculating the Probability

2.4.1 Assumptions

The objective of this research is to find the probability of damaging the equipment. To do this we have to simulate similar ARIMA models many times and find the percentage of it going over the limit for some time. For this project, we assume that the equipment will be damaged when the temperature exceeds 45°c for 3 days (72 hours).

2.4.2 Procedure

Since we decided on ARIMA(2,1,0) as the best model for this data, we simulated 100 ARIMA(2,1,0) models with Φ_1=1.0936 and Φ_2= -0.3976. Those numbers were gathered from the summary of the ARIMA model. We shifted the simulated series by 20.59 so that they started from the same point as our original data. After that, we put the simulations into a list of lists and turned the inner list into a data frame with the first column as the hour and the second column as the simulated temperature.

The purpose of turning it into a data frame with an index row as one of the columns is so that when we filter the data, we can keep track of the original index. After turning it into a data frame, we filter all the data frames inside the list, keeping only rows with temperatures more than 45°c.

Now we have a list of data frames with temperatures more than 45°c and their time. The last step is to mark the simulation as true if the data frame has 72 consecutive data. We do this by searching for 72 sequential index numbers in the filtered data frames. If there is an instance of 72 consecutive index numbers, then that simulated machine exceeded 45°c for 72 hours or more. Now we have a list boolean, we can use this list to calculate the probability by counting the number of trues divided by the number of simulations, 100. The code is attached in *Figure 15*.

2.4.3 Result

The probability of damaging the equipment from 100 simulations is 0.23. This means the probability of damaging the equipment is low and we should keep this equipment and buy the same model when it is broken. We increased the number of simulations to 10,000 to check the precision of the probability and we ended up with 0.239. This means the true probability is around 0.23-0.24, which is still a low probability.

2.4.4 Limitation

The drawback of this simulation is that some of the simulations decrease in temperature as time goes on, reaching negative value, instead of increasing. This is not an accurate representation of reality because machines rarely cool down until reaching negative values when used. This problem can be seen in *Figure 14*.

3. Conclusion

In this project, we studied the 243 hourly observations of equipment temperature to see if the temperature could cause damage to the equipment. We assumed that the condition to damage equipment is when the temperature is more than 45 degrees in three consecutive days. To find probability, we started by studying the basic statistics of the data. By looking at the observation closely, we found that the data has non-stationarity and stochastic trends. Since the data has a stochastic trend, we made the data stationary by differentiating. After the first differentiation, we assumed the data was stationary, and started modeling the data. The good model helped us with the simulation to find the probability for machines to have severe damage.

We have concluded that the replacement of the equipment in the company is unnecessary. The probability of critical damage to equipment is 23 %, and we suggest that we should still use the equipment. This analysis also made us realize that the manufacture of the equipment is reliable in terms of quality for this machine. As time passes, though the result could change, our findings will play a role in future decision-making related to equipment replacement within the organization.

Appendix

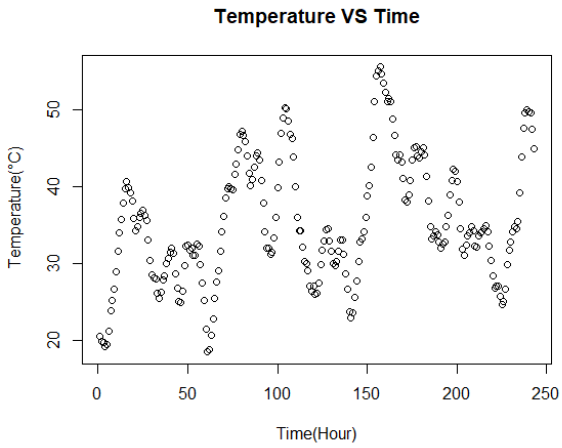

Figure 1 Plot of the original data

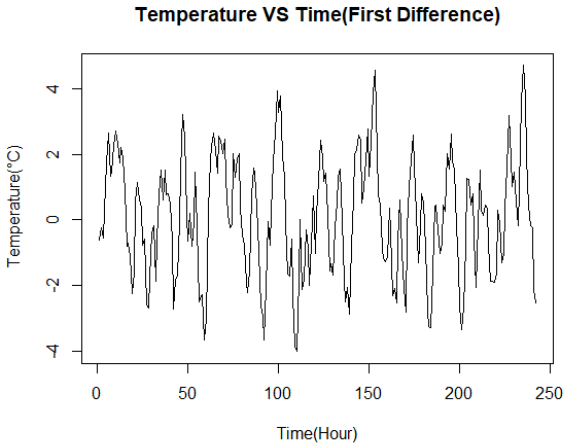

Figure 2 Plot of the first difference

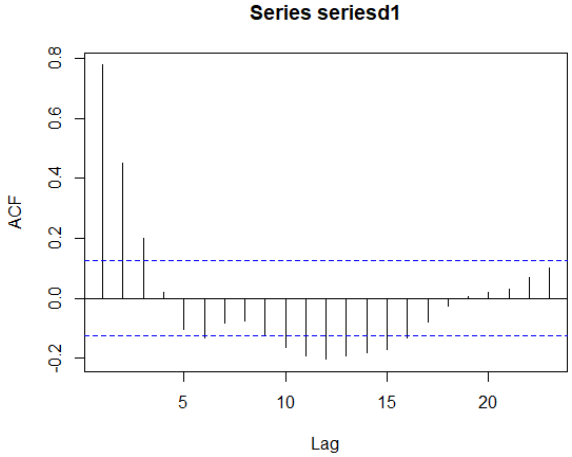

Figure 3 ACF of the first difference

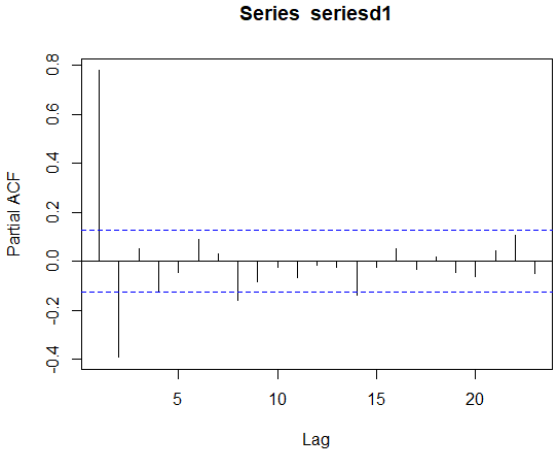

Figure 4 PACF of the first difference

AR/MA

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	x	o	o	x	o	o	o	x	x	x	x	x
1	x	x	x	o	x	x	o	o	o	o	o	o	o	o
2	o	x	o	x	o	x	o	o	o	o	o	o	o	o
3	x	x	o	x	o	x	o	o	o	o	o	o	o	o
4	x	x	o	o	o	x	o	o	o	o	o	o	o	o
5	x	x	o	o	o	x	o	o	o	o	o	o	o	o
6	x	x	x	x	o	x	o	o	o	o	o	o	o	o
7	o	x	x	x	o	x	o	o	o	o	o	o	o	o

Figure 5 EACF of the first difference

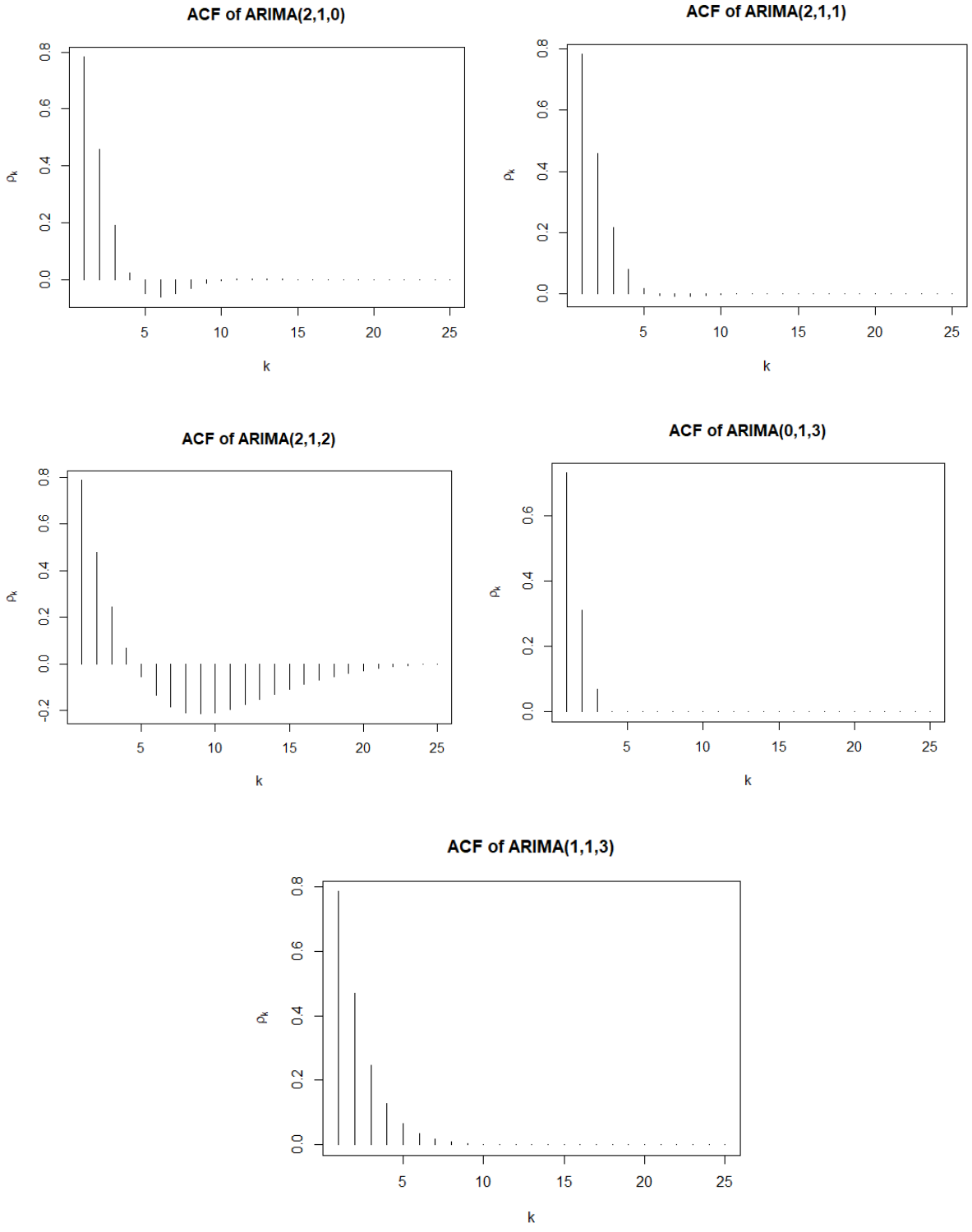

Figure 6 ACF of the possible models

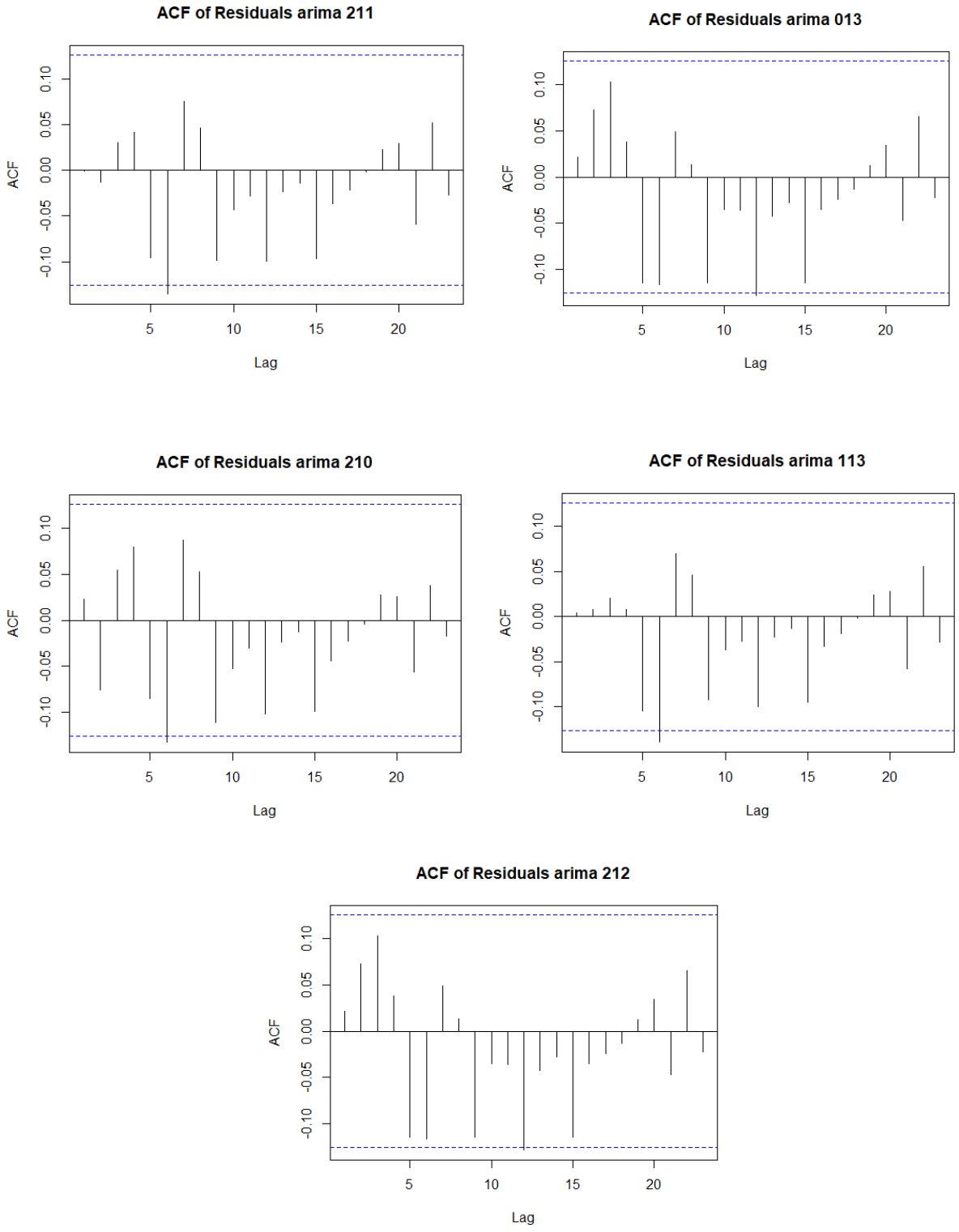

Figure 7 Residual ACF of the possible models

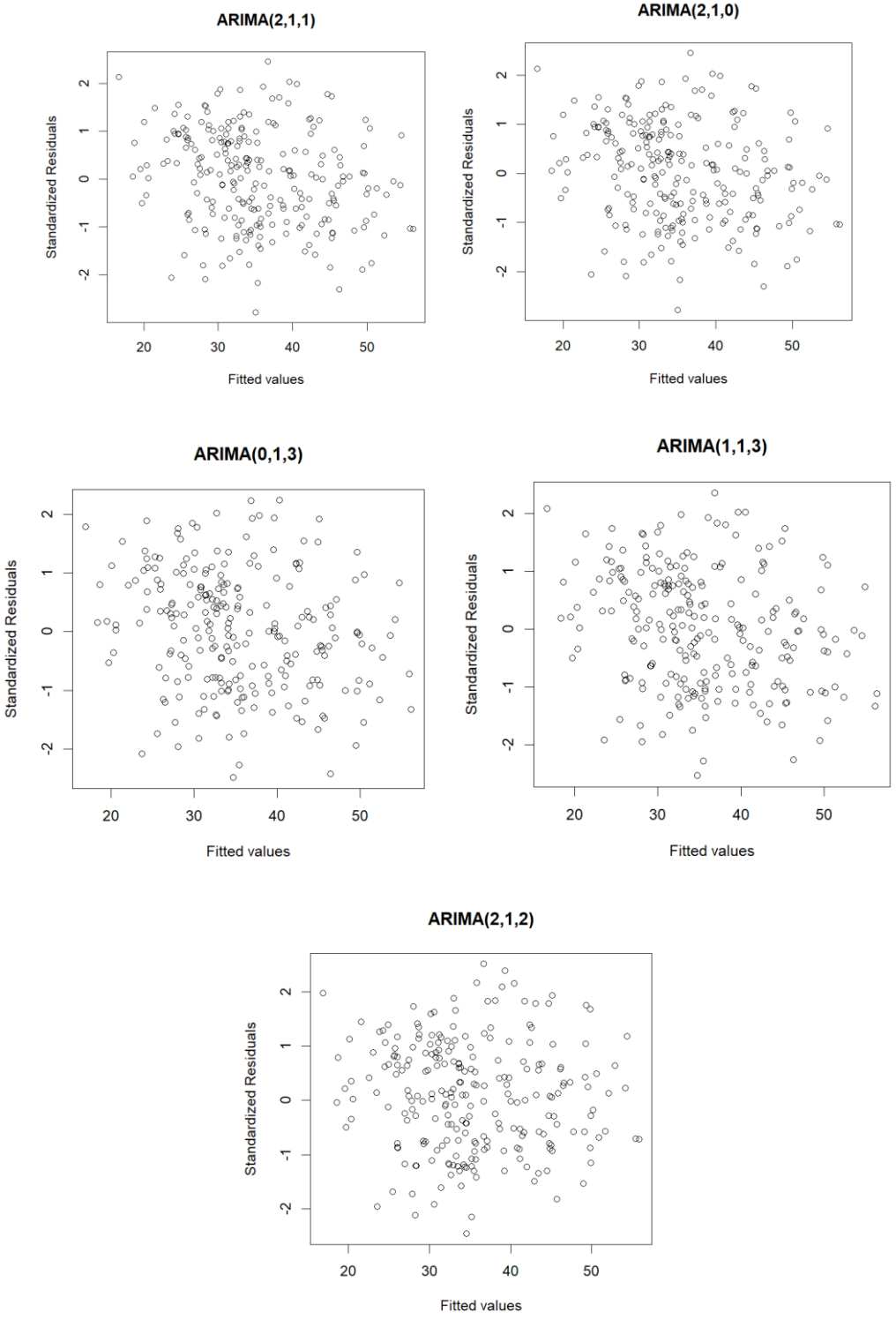

Figure 8 Residual VS fitted model of the possible models

```
Call:
arima(x = series, order = c(2, 1, 1))

Coefficients:
         ar1      ar2     ma1
      0.9215  -0.2623  0.2032
s.e.  0.1677   0.1409  0.1747

sigma^2 estimated as 1.044:  log likelihood = -349.29,  aic = 704.58
```

Figure 9 Summary of ARIMA(2,1,1) model

```
Call:
arima(x = series, order = c(0, 1, 3))

Coefficients:
         ma1     ma2     ma3
      1.1214  0.6745  0.2054
s.e.  0.0635  0.0797  0.0533

sigma^2 estimated as 1.069:  log likelihood = -352.14,  aic = 710.27
```

Figure 10 Summary of ARIMA(0,1,3) model

```
Call:
arima(x = series, order = c(2, 1, 2))

Coefficients:
         ar1      ar2      ma1      ma2
      1.6371  -0.6862  -0.5746  -0.4082
s.e.  0.0542   0.0543   0.0663   0.0654

sigma^2 estimated as 0.9958:  log likelihood = -344.19,  aic = 696.38
```

Figure 11 Summary of ARIMA(2,1,2) model

```
Call:
arima(x = series, order = c(2, 1, 0))

Coefficients:
         ar1      ar2
      1.0936  -0.3976
s.e.  0.0588   0.0589

sigma^2 estimated as 1.049:  log likelihood = -349.83,  aic = 703.66
```

Figure 12 Summary of ARIMA(2,1,0) model

```
Call:
arima(x = series, order = c(1, 1, 3))

Coefficients:
         ar1     ma1     ma2     ma3
      0.5161  0.6094  0.1778  0.0163
s.e.  0.1377  0.1465  0.1478  0.0920

sigma^2 estimated as 1.047:  log likelihood = -349.61,  aic = 707.23
```

Figure 13 Summary of ARIMA(1,1,3) model

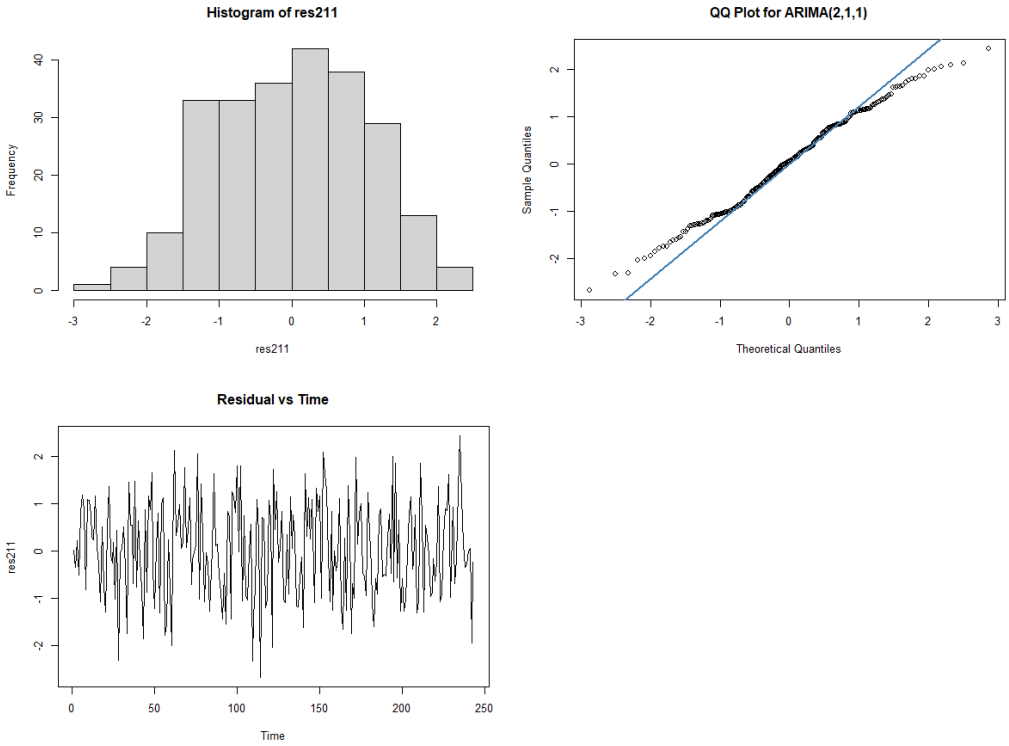

Figure 14 Model diagnostic of ARIMA(2,1,1)

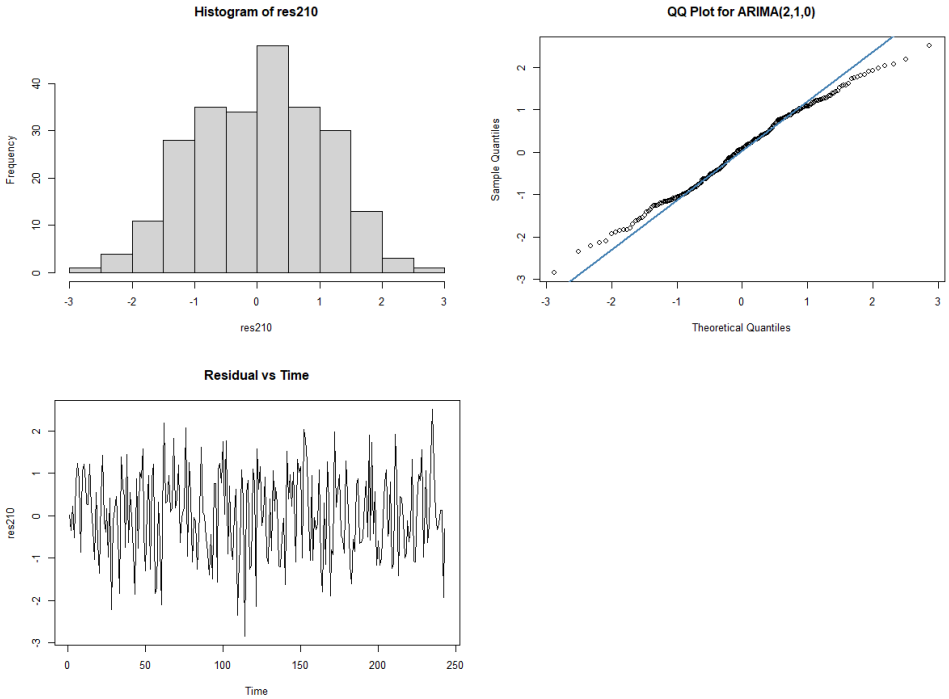

Figure 15 Model diagnostic of ARIMA(2,1,0)

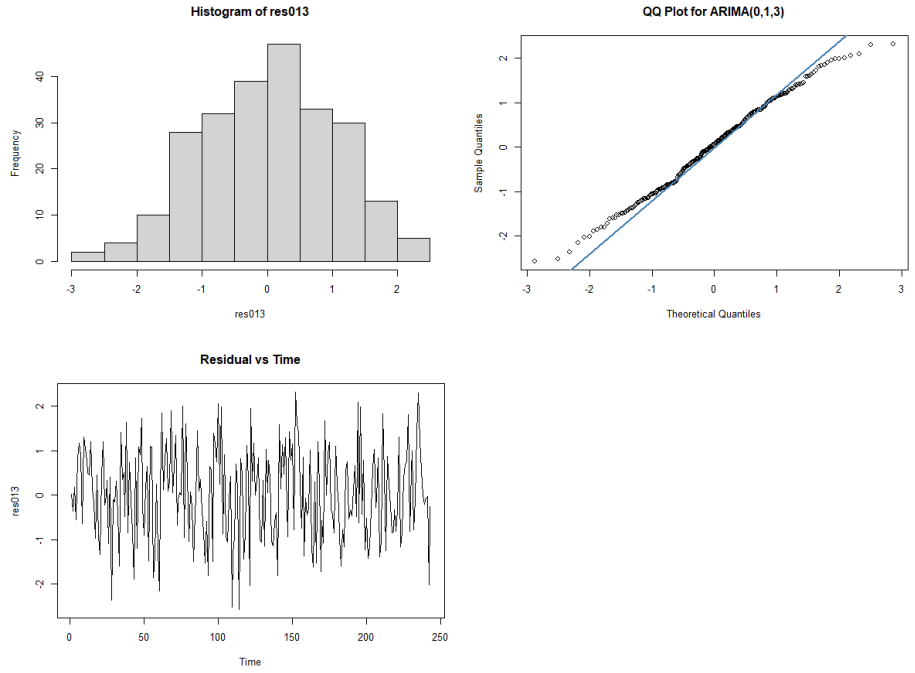

Figure 16 Model diagnostic of ARIMA(0,1,3)

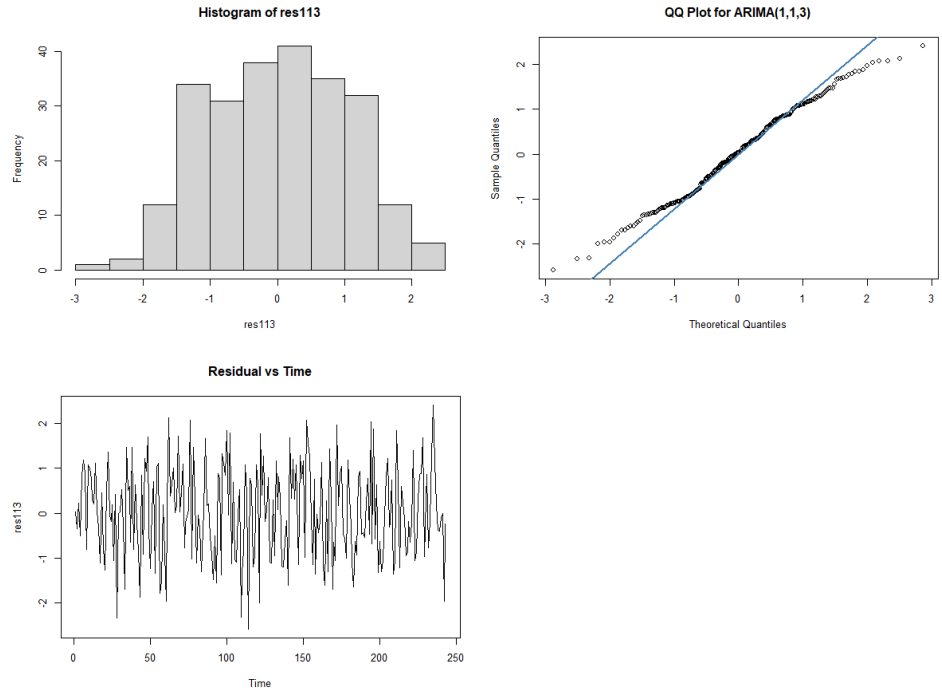

Figure 17 Model diagnostic of ARIMA(1,1,3)

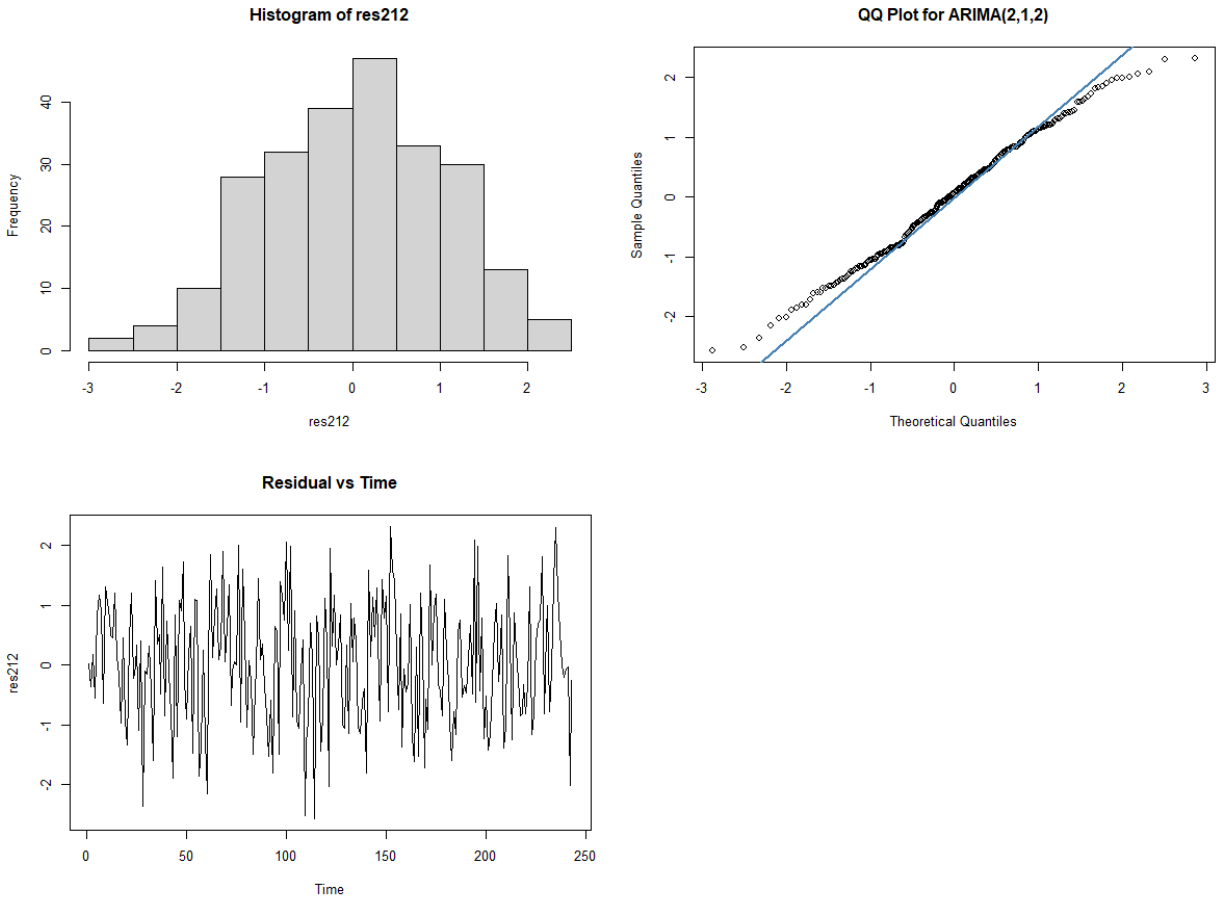

Figure 18 Model diagnostic of ARIMA(2,1,2)

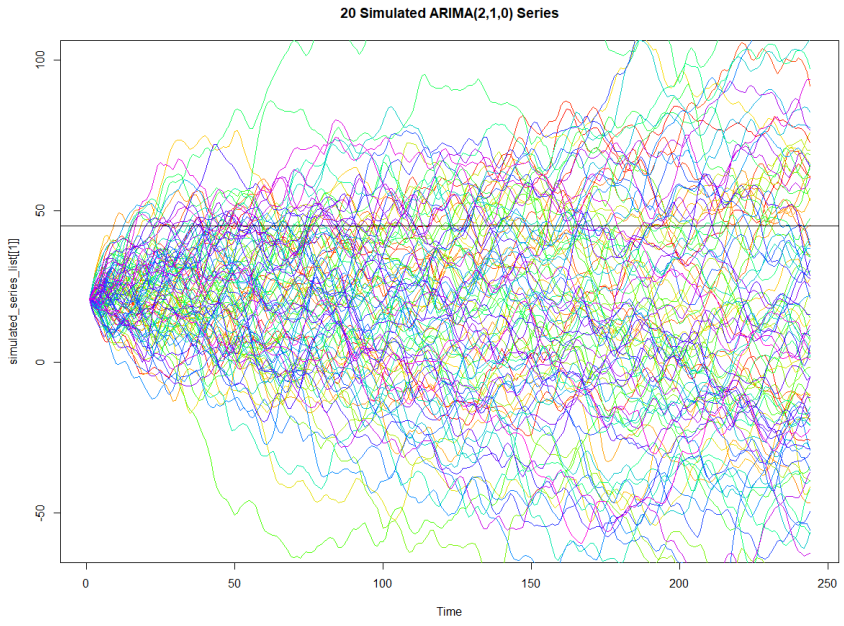

Figure 19 ARIMA(2,1,0) simulations, n=100

```
#=======================================Simulations=============================================``
set.seed(2024)
for (i in 1:n) {
  simulated_series <- arima.sim(n = 243, model=list(order=c(2,1,0), ar=ar_params))
  simulated_series <- simulated_series + 20.59 - simulated_series[1]  # Shift the series to start from 20.59
  simulated_series_list[[i]] <- simulated_series
}

#=======================================creating dataframe=======================================``
sequence <- 1:244
dataframe_list <- list()
for (i in 1:length(simulated_series_list)) {
  df <- data.frame(Column1 = sequence, Column2 = unlist(simulated_series_list[[i]]))
  dataframe_list[[i]] <- df
}
##=========================================making boolean lists=========================================
bool_list <- list()
for (i in 1:length(dataframe_list)) {
  filtered_df <- subset(dataframe_list[[i]], Column2 > 45)
  sequence <- rle(diff(filtered_df$Column1) == 1)$lengths
  bool_list[[i]] <- any(sequence >= 72)
}
bool_df <- data.frame('BoolValues' = unlist(bool_list))

#=========================================calculating probs=========================================
true_count <- sum(bool_df$BoolValues)
probability <- true_count / n
print(probability)
```

Figure 20 Code for calculating the probability