Series Temporales

Proyecto

Jakub Maciążek

Project description

Search for a dataset (with temporary information) related to one of the great problems in the world today, which has been published in 2021 or 2022.

Conduct data analysis, extract useful knowledge and draw conclusions. Extract trends, conclusions, etc. with any technique seen in class.

Explain everything.

Data source

As one of the great problems of current times are health problems related to air pollution, I have decided to analyze information about PM 2.5 levels in my university home city, as among air pollution, this one is the most detrimental for human health.

Historic data was sourced from governmental agency "Main inspectorate of environmental protection" at following address: https://powietrze.gios.gov.pl/pjp/archives.

```
data_daily <- read.csv("S:/0_Universidad_de_Malaga/MI_Ingenieria_y_ciencia_de_datos/Estatistica_avanzada
head(data_daily)
##         date value
## 1 2019-01-01  16.5
## 2 2019-01-02   8.2
## 3 2019-01-03   9.7
## 4 2019-01-04  13.0
## 5 2019-01-05   9.4
## 6 2019-01-06  11.0
```

Basic data overview

As following paper focuses on time series analysis, xts object will be used. (In order to do that, dates in string format were first converted to date class.)

```
data_daily$date <- as.Date(data_daily$date)
xts_daily <- as.xts(data_daily)
head(xts_daily)
##            value
## 2019-01-01  16.5
## 2019-01-02   8.2
```

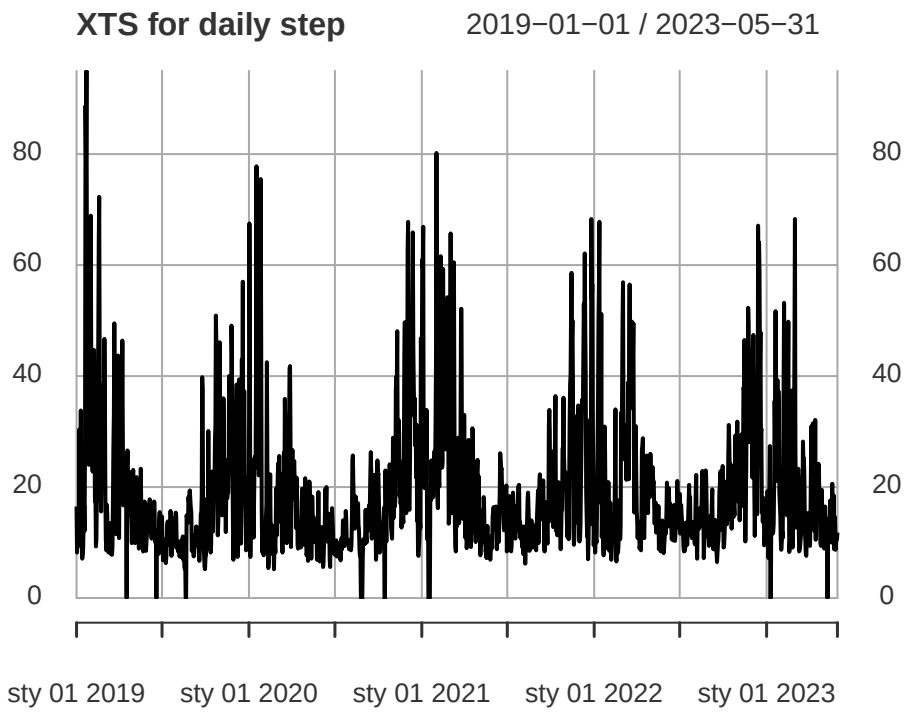
```
## 2019-01-03   9.7
## 2019-01-04  13.0
## 2019-01-05   9.4
## 2019-01-06  11.0
```

Timeseries contains information for days from 2019-01-01 to 2023-05-31.

```
periodicity(xts_daily)
## Daily periodicity from 2019-01-01 to 2023-05-31
```

As can be seen on below graph, daily time series is very volatile. Therefore analysis will be conducted on time series based on monthly, quarterly or yearly periods.

```
plot(xts_daily, main="XTS for daily step")
```



Following transformations resulted in 3 new time series. Monthly one, based on 53 months, quarterly one, based on 18 quarters, and yearly one, based on 5 years.

Yearly time series contains only 5 values, therefore will be used only for very general overview. More information will be derived from other two.

```
xts_monthly <- apply.monthly(xts_daily, mean)
nmonths(xts_monthly)
## [1] 53

xts_quarterly <- apply.quarterly(xts_daily, mean)
nquarters(xts_quarterly)
```

```
## [1] 18

xts_yearly <- apply.yearly(xts_daily, mean)
nyears(xts_yearly)
## [1] 5
```

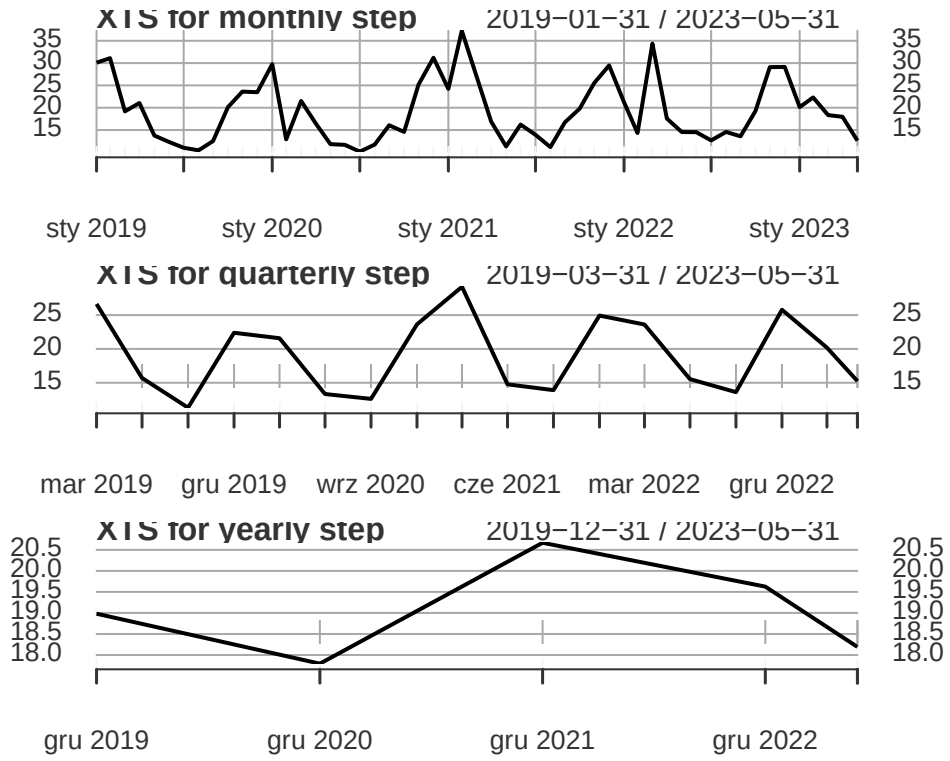
Exploratory analysis

Initial plotting

Next step was initial plotting of new 3 transformations of TS. From them, following information may be seen:

- starting with last one, in yearly plot no clear tendency can be seen. This however may be caused by partial data for the last year, and covid lock downs in year 20220 that may have decreased pollution (Those anomalies are also seen in daily plot).
- Quarterly plot shows, that there may be a slight positive tendency. It also shows that seasonality of data may exists.
- Monthly plot further shows seasonality of the data. It also shows that besides pollution being higher in the winter and lower during summer, for 4 consecutive years at the turn of the year (all visible ones), pollution declines in the middle of the winter. This may be caused by Christmas break and holidays, that result in less working activity, and lack of students in the city.

```
par(mfrow=c(3,1))
plot(xts_monthly, main="XTS for monthly step")
plot(xts_quarterly, main="XTS for quarterly step")
plot(xts_yearly, main="XTS for yearly step")
```



Data split

As later in the analysis we would like to conduct some predictions, dataset will be now split. This way remaining data will serve as comparison for predictions. Moreover, for a split, year 2023 will be used as reminding data, which will also exclude problems related to its incomplete data (For only up to the end of May).

This split will be conducted for monthly time series, on which analysis will focus, as it is already stable and provides more exact information than quarterly one.

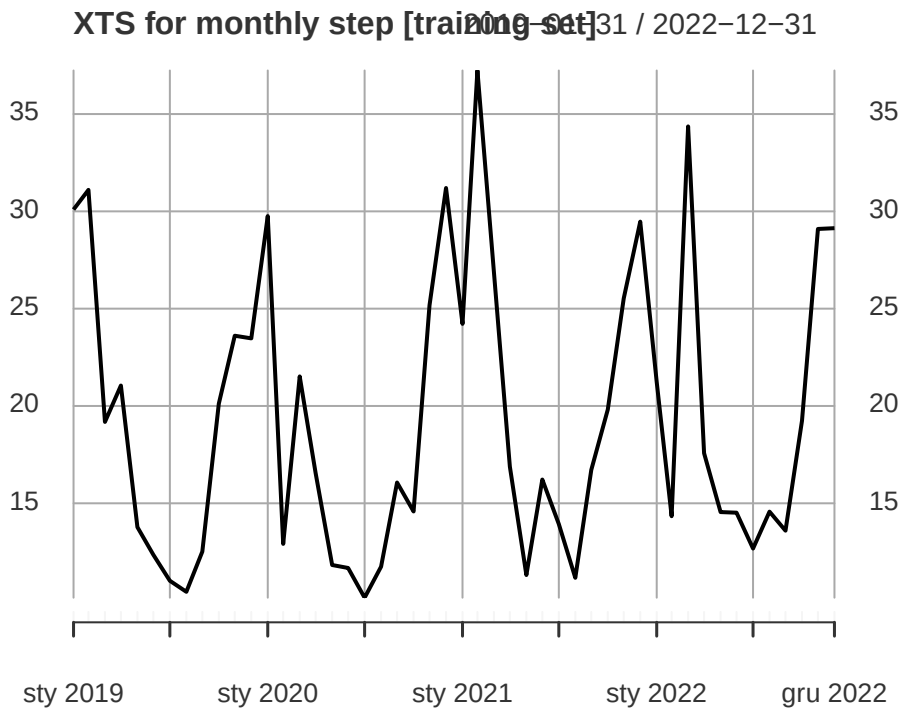
```
xts_monthly_training <- xts_monthly["/2022"]
# Training TS end
end(xts_monthly_training)
## [1] "2022-12-31"

# Testing TS start
xts_monthly_testing <- xts_monthly["2023"]
start(xts_monthly_testing)
## [1] "2023-01-31"
```

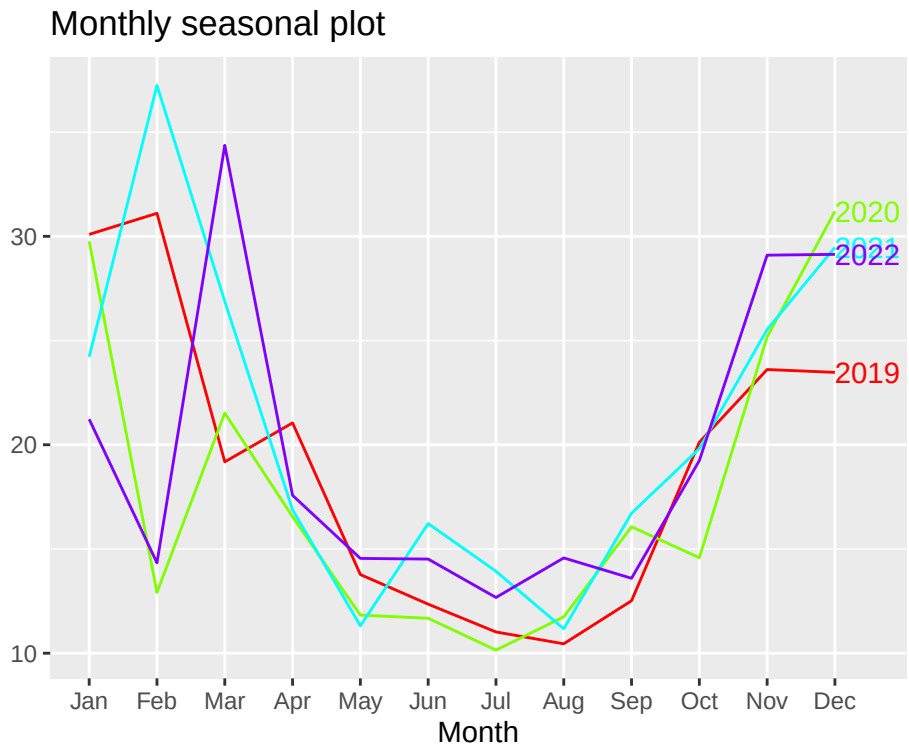
Basic properties analysis

In following section trend, seasonality, heteroskedasticity and stationarity willl be discussed. However, that trend is not visible in seasonal plot.

```
plot(xts_monthly_training, main="XTS for monthly step [training set]")
```




```
ts_monthly_training <- ts(xts_monthly_training$value, frequency = 12, start = c(2019, 1))
ggseasonplot(ts_monthly_training, col = rainbow(4), year.labels = TRUE, continuous = TRUE, main="Monthly
```



Trend

From quarterly plot it was said that data may have **slight positive trend**, and long-term increase. However, it is so small, that we can assume there is none.

Seasonality

There is a yearly seasonality of data. This pattern is **clearly visible** on seasonal plot.

Heteroskedasticity

Data seems to have similar variability, therefore **does not satisfy this property**.

Stationarity

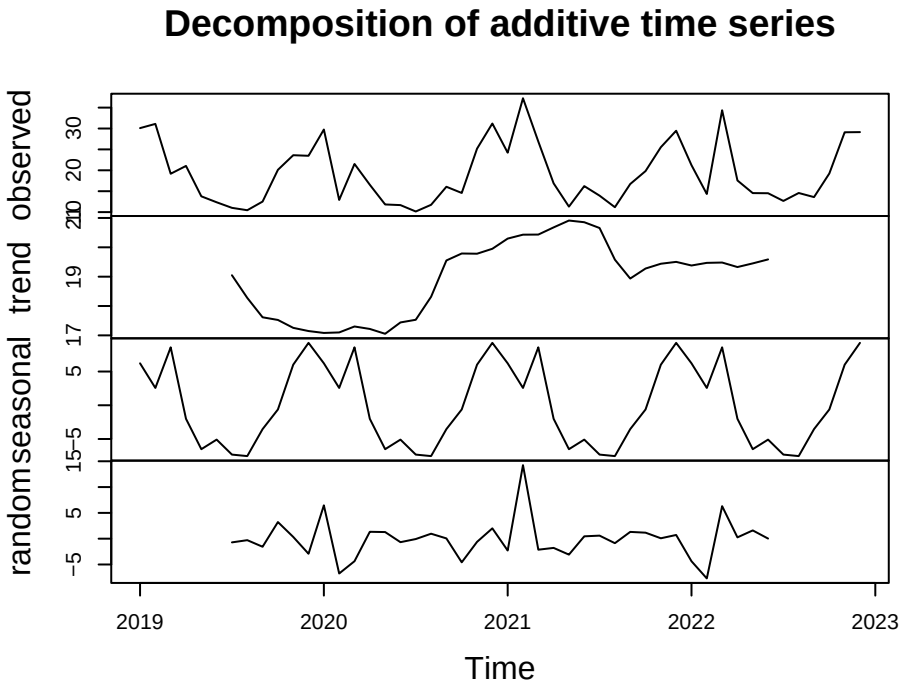
As both variance and mean seem constant, it **satisfies this property**. Moreover, further *Augmented Dickey-Fuller Test* results in very small p-value of 0.01, which confirms stationarity and that time series does not have a unit root.

```
adf.test(ts_monthly_training)
## Warning in adf.test(ts_monthly_training): p-value smaller than printed p-value
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_monthly_training
## Dickey-Fuller = -4.3616, Lag order = 3, p-value = 0.01
## alternative hypothesis: stationary
```

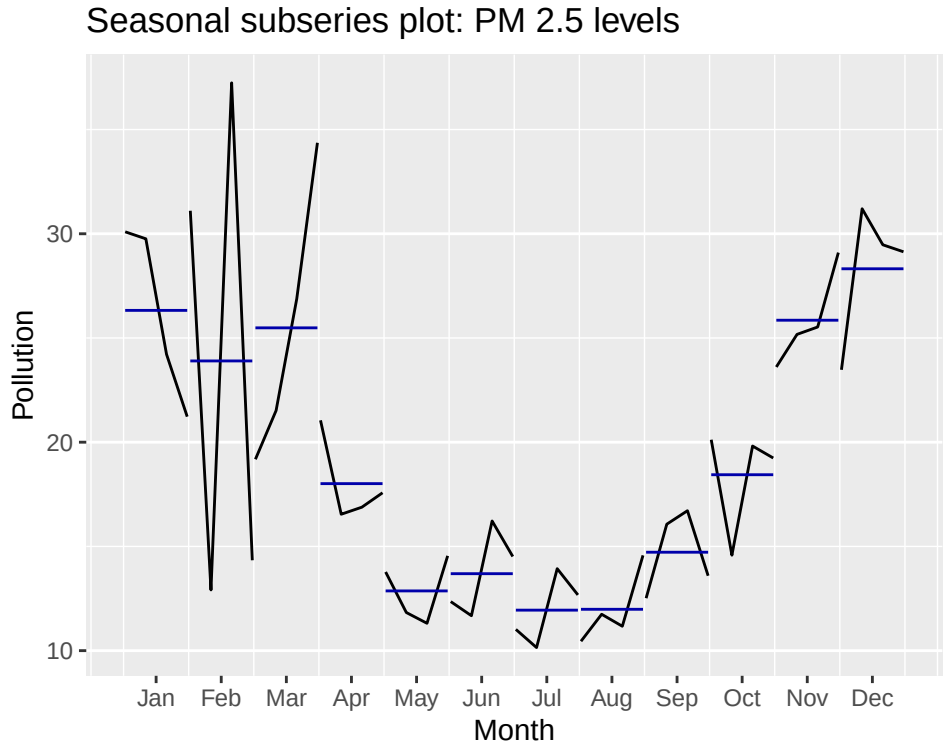
Decomposition

There is no trend.

```
plot(decompose(ts_monthly_training))
```




```
ggsubseriesplot(ts_monthly_training) +
  ylab("Pollution") +
  ggtitle("Seasonal subseries plot: PM 2.5 levels")
```

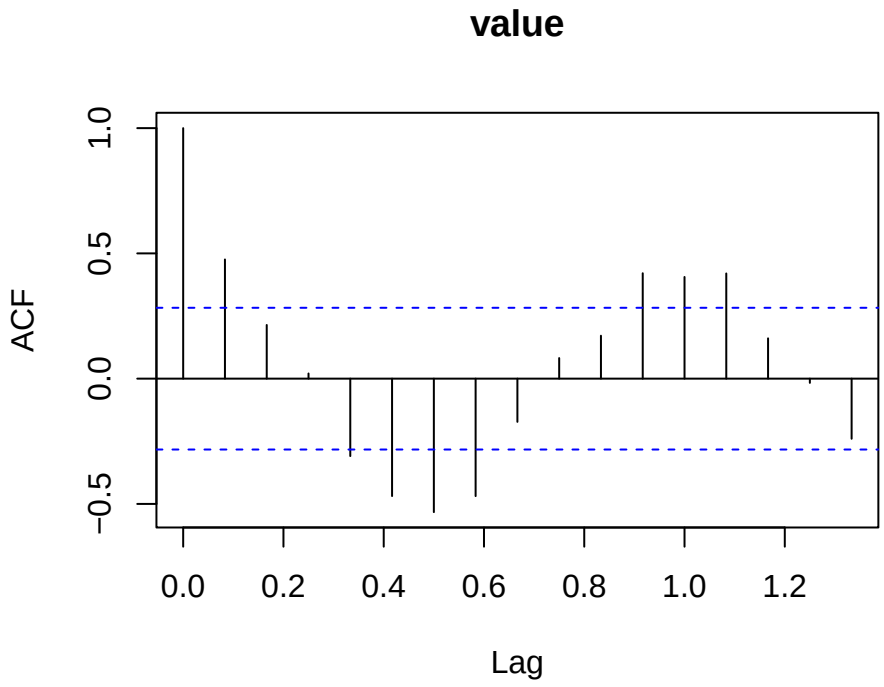


ACF and PACF - autocorrelation diagnosis

Autocorrelation is tested, as it is necessary for Auto-regressive model, therefore model selection depends on it.

As ACF cuts of multiple lags (1, 4-7, 11-12) as those are above 95% confidence interval and therefore statistically non-zero. It means that it is highly auto correlated and AR can be used. AM(12) can be used.

```
acf(ts_monthly_training)
```



In PACF lag 3 shows strong correlation, therefore AR(3) can be used.

```
pacf(ts_monthly_training)
```

Model estimation and forecasting

As it was previously proven that time series is stationary and has rather no trend, ARIMA model (which is widely used) can be constructed. However, it is not based on seasonality, so it should be removed (or specified by parameter). Moreover autocorrelation analysis proofed existence of auto correlation, therefore this model should work.

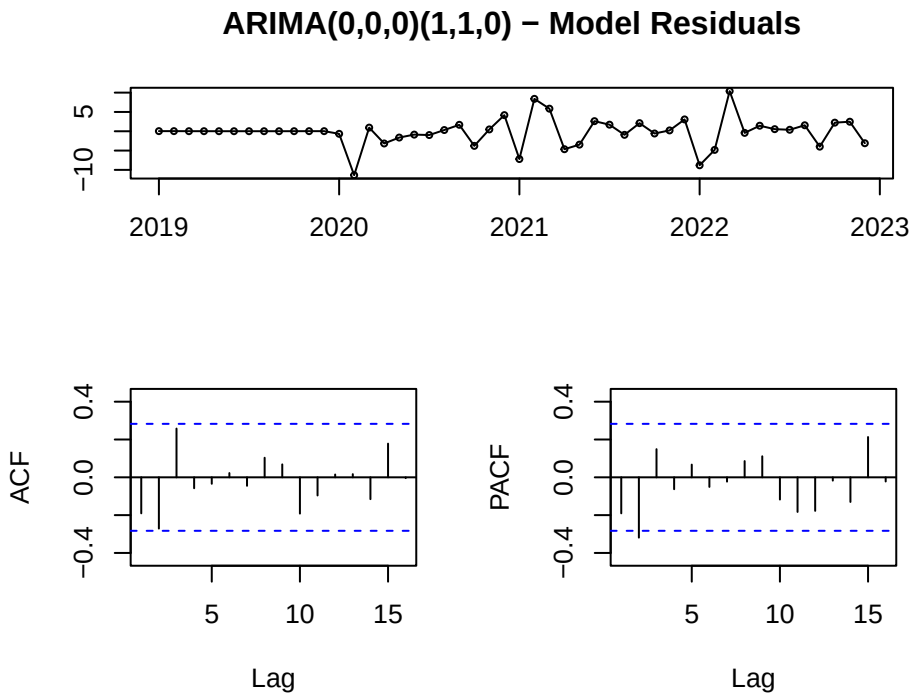
Model construction

Model was constructed using automated method, and resulted in **ARIMA(0,0,0)(1,1,0)**.

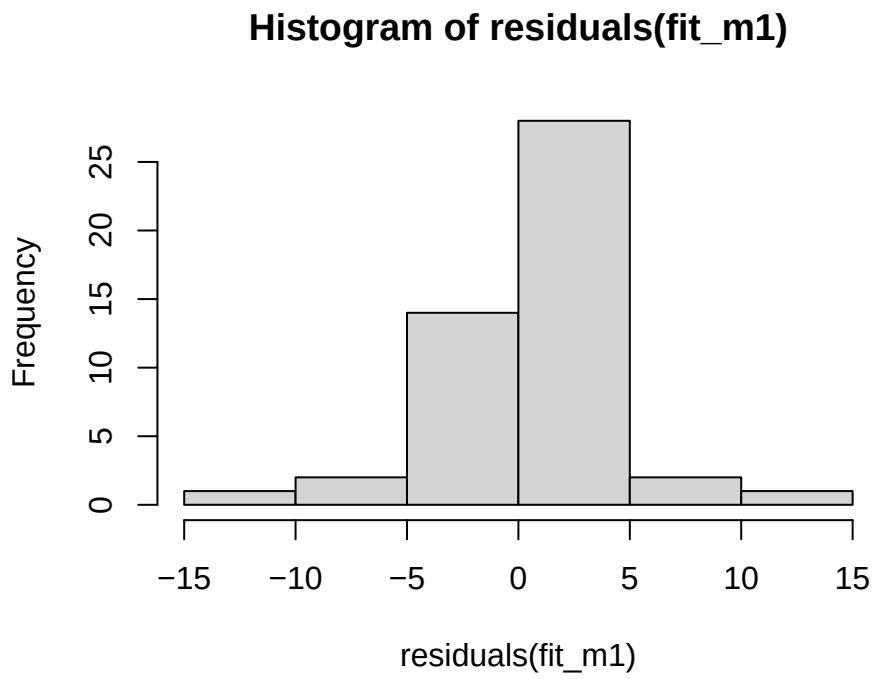
```
fit_m1 <- auto.arima(ts_monthly_training, seasonal = TRUE)
fit_m1
## Series: ts_monthly_training
## ARIMA(0,0,0)(1,1,0)[12] with drift
##
## Coefficients:
##          sar1   drift
##       -0.8000  0.0642
## s.e.   0.0823  0.0392
##
## sigma^2 = 18.98:  log likelihood = -109.16
## AIC=224.32   AICc=225.07   BIC=229.07
```

Residuals of the model, not perfectly, but seem to resemble normal distribution, therefore forecasting can be conducted.

```
par(mfrow=c(2,1))
tsdisplay(residuals(fit_m1), main = "ARIMA(0,0,0)(1,1,0) - Model Residuals")
```




```
hist(residuals(fit_m1))
```

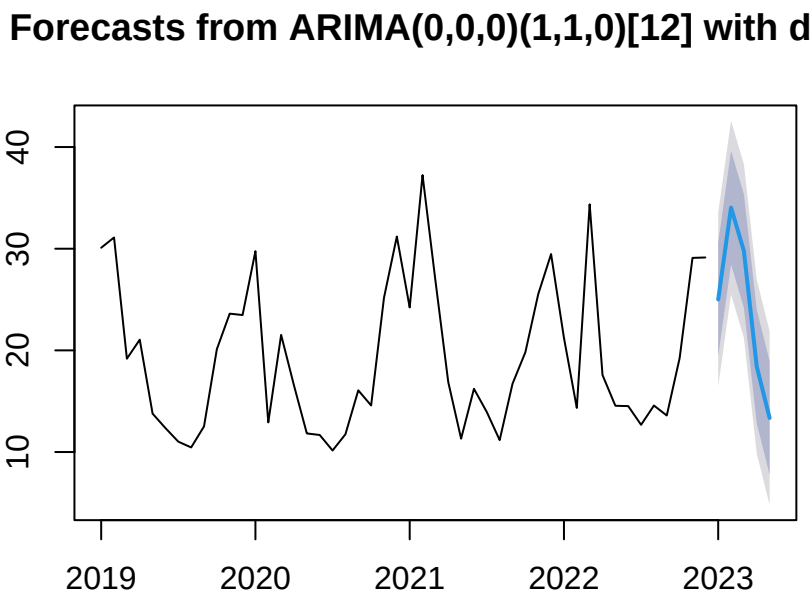


Forecasting

Using generated model, forecast for year 2023 is conducted. Prediction for the first 5 months seems to follow real values. **Accurate model was generated.**

```
fcast <- forecast(fit_m1, h = 5)

#par(mfrow=c(2,1))
plot(fcast)
```




```
plot(xts_monthly, main = "TS reminder")
```