LINEAR TIME SERIES

Assignment

ARIMA MODELLING OF THE INDUSTRIAL PRODUCTION INDEX OF REFRACTORY PRODUCT MANUFACTURING

Ahmed Ayan, Ahmed Rayyan

May 23, 2024

Contents

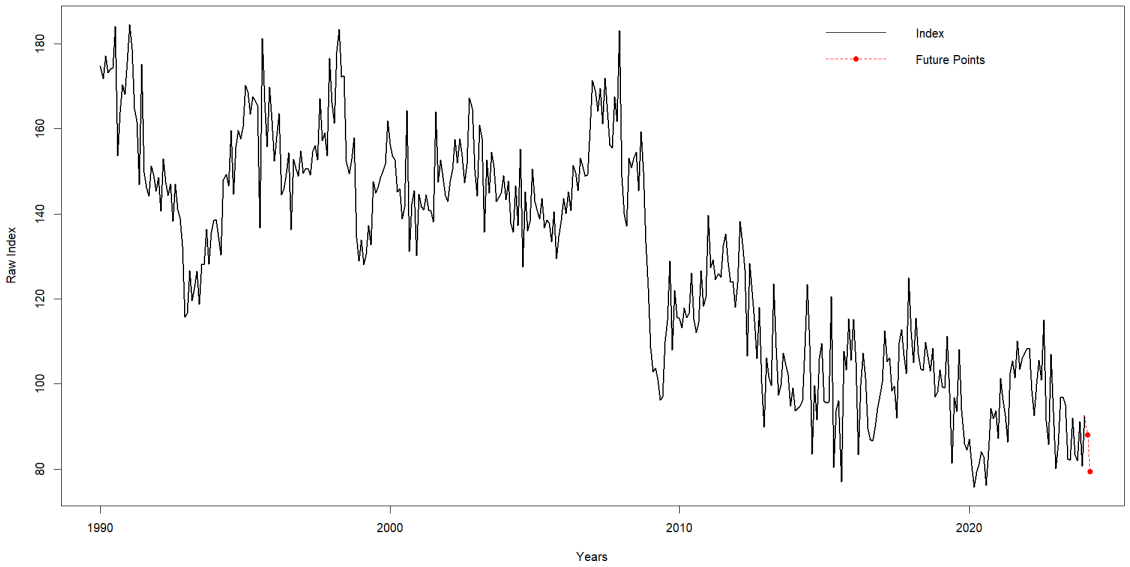
1 Data

1.1 Data Presentation

We have used the dataset utilizing the Industrial Production Index(base 100 in 2021) of Refractory Product Manufacturing(NAF rev. 2, class level, item 23.20). This dataset is CVS-CJO, i.e., it is corrected from seasonal variations and working days with a monthly frequency. The Industrial Production Index (IPI) of Refractory Product Manufacturing measures the real output of the manufacturing sector that produces refractory products. These products are materials that are resistant to high temperatures and are used in various high-temperature industrial processes, such as in furnaces, kilns, incinerators, and reactors. The IPI is typically reported on a monthly basis and is used to track changes in production levels over time, providing insights into the economic health and trends within the refractory product manufacturing industry. This index is an important indicator for economists, industry analysts, and policymakers to assess the performance and growth of this specific sector within the broader manufacturing industry. The Industrial Production Index (IPI) of Refractory Product Manufacturing offers several other key benefits:

- **Economic Indicator**: It serves as a vital economic indicator, providing insights into the production trends and overall health of the refractory product manufacturing sector. This helps in understanding the sector's contribution to the economy.
- **Trend Analysis**: Analysts can use the IPI to identify short-term and long-term trends in the production of refractory products, enabling businesses and investors to make informed decisions.
- **Supply Chain Management**: Manufacturers and suppliers in related industries can use this data to anticipate demand fluctuations, optimize inventory levels, and manage supply chains more effectively.
- **Investment Decisions**: Investors can analyze the IPI to assess the performance and potential growth of companies within the refractory product manufacturing sector, aiding in making informed investment choices.
- **Policy Making**: Policymakers can utilize the IPI to gauge the effectiveness of industrial policies, design better economic strategies, and support the growth of the manufacturing sector.
- **Benchmarking**: Companies within the industry can benchmark their performance against the overall production trends indicated by the IPI, helping them to identify areas for improvement and growth opportunities.

The series that we have taken covers monthly data from January 1990 to March 2024. Let us first look at the time series.



The graph displays the Industrial Production Index (IPI) for Refractory Production Manufacturing from around 1990 to 2024. The y-axis represents the Raw Index values, while the x-axis represents the years. The black line represents the historical index values, and the red dots represent future points.

- **Overall Trend:** The general trend over the entire period is downward, indicating a long-term decline in refractory production manufacturing. This trend might reflect broader industrial and economic shifts, such as changes in demand, advancements in technology, outsourcing, or other structural changes in the industry.
- **Volatility:** There is significant volatility throughout the entire period, with frequent and sharp fluctuations. Peaks and troughs are evident, especially in the 1990s and early 2000s, suggesting cyclical patterns or responses to economic cycles and external shocks.
- **Major Peaks and Troughs:** A prominent peak is observed in the early 1990s, with another notable high around the mid-2000s. Significant declines are seen, particularly around the late 2000s (likely related to the global financial crisis) and post-2010, indicating periods of downturn or reduced industrial activity.
- **Recent Years:** From 2010 onwards, the index shows increased fluctuations but with a more pronounced downward trend. The red dots representing future points show a continuation of this declining trend, suggesting expectations of further decreases in the near future.
- **Future Points:** The future points (in red) indicate a predicted continuation of the downward trend, possibly projecting lower production levels. The proximity of these points to the historical data suggests that recent patterns are influencing these predictions.
- **Potential Influences:** Factors influencing these trends could include changes in the steel and construction industries (major consumers of refractory materials), shifts towards alternative materials or technologies, environmental regulations, and global economic conditions.
- **Comparative Analysis:** Comparing the highest peak (around 1990) and the lowest trough (around 2020) shows a substantial decline in the index value, emphasizing the long-term reduction in production.

1.2 Data Processing

From the above observations, we conclude that our series is not stationary. We perform the Augmented Dickey Fuller(ADF) Test to validate this fact. Before performing the unit root tests, we need to check if there is an intercept and / or a non null linear trend. The graph representation of the series showed that the trend probably isn't linear, but if we need to pick one, it would be negative. Let's regress the series on its dates to check.

```
> summary(lm(index1~date1))

Call:
lm(formula = index1 ~ date1)

Residuals:
    Min      1Q  Median      3Q     Max
-45.404 -10.269   0.911   9.315  55.667

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4651.95685  156.65857   29.70   <2e-16 ***
date1         -2.25335    0.07806  -28.87   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.53 on 407 degrees of freedom
Multiple R-squared:  0.6719,    Adjusted R-squared:  0.6711
F-statistic: 833.4 on 1 and 407 DF,  p-value: < 2.2e-16
```

The coefficient associated with the linear trend (dates) is indeed negative, and may be statistically significant (which cannot be confirmed because the test is not valid where there are possibly autocorrelated residuals). We need to study the case of unit root tests with intercept and possibly non-zero trends.

The augmented Dickey-Fuller test (ADF) in the intercept and trend case consists of the following regression, with a given X variable:

$$\Delta X_t = c + bt + \beta X_{t-1} + \sum_{l=1_{k>0}}^{k} \phi_l \Delta X_{t-l} + \varepsilon_t \tag{1}$$

where $\beta + 1$ is the autocorrelation of order 1 of X and k is the number of lags needed to render our residuals non-autocorrelated.

The null hypothesis of unit root $H_0 : \beta = 0$ is tested by the test statistic $\frac{\hat{\beta}}{\hat{\sigma}(\hat{\beta})}$ which follows a Dickey-Fuller distribution depending on the number of observations and the case of the test we are studying. After using the script(see Appendix), we conclude that the minimum number of lags needed to erase residual autocorrelation is 14. And finally, we conclude by saying that the series is not stationary since the p-value is greater than 0.05 and the unit root is not rejected at the 95 percent- level for the series in levels, the series is thus at least I(1).

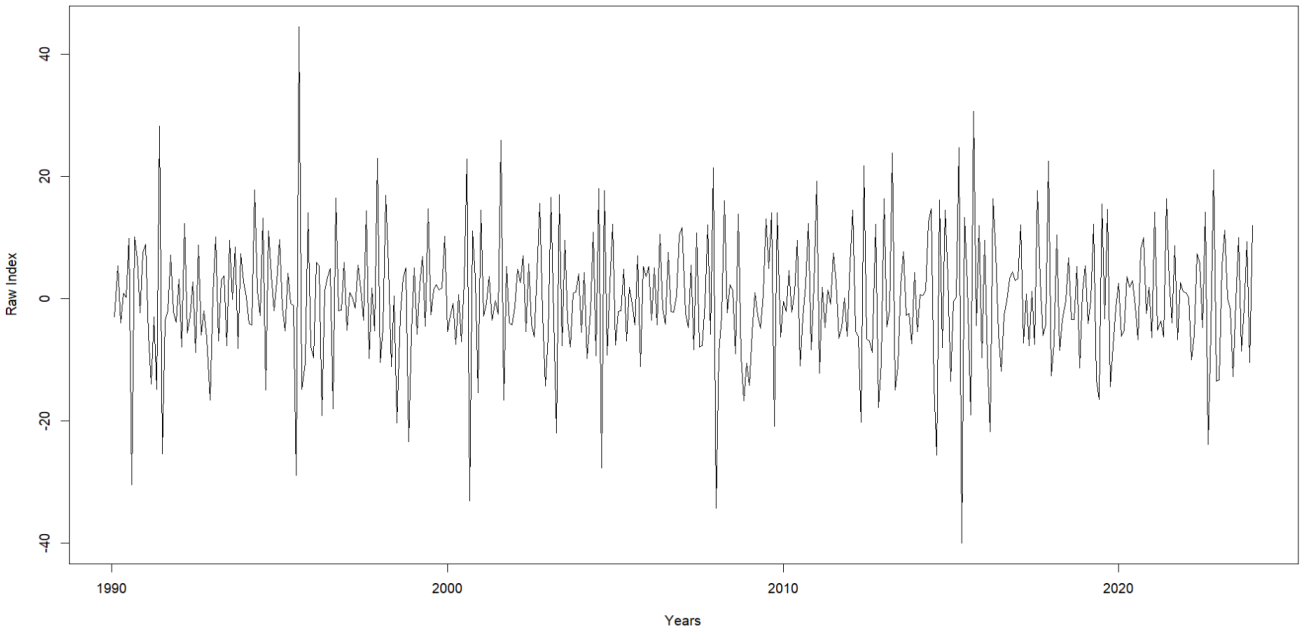
1.3 Transformation of the Series

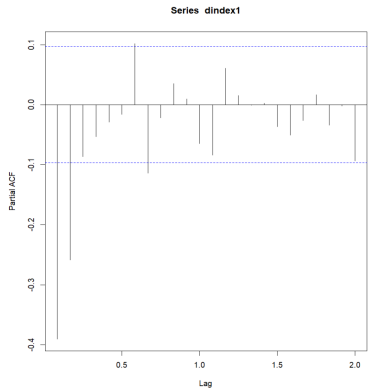
We thus difference the series once using $X_t = \Delta Y_t = Y_t - Y_{t-1}$. We observe that the series appears to be stationary with almost 0 mean and variance over time implying that the trends have been removed. But we need to verify that this new series is indeed stationary. Here we apply three tests namely ADF, PP and KPSS whose observations are given below.

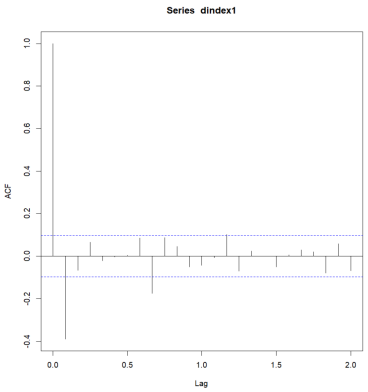
Test	**Test Statistic**	**Lag Order**	**p-value**
ADF	-16.0498	2	< 0.05
PP	-489.44	5	< 0.05
KPSS	0.01875	5	> 0.05

Table 1: Test Statistics

The first two tests provided us with a p-value less than 0.05, thus allowing us to reject the null hypothesis at 5 percent significance while the KPSS test which has the null hypothesis as the stationarity series had a p-value greater than 0.05, thus we couldn't find enough evidence to reject our null hypothesis. Thus, these tests showed that our series is indeed stationary and we do not need to difference our series further. The difference series when plotted looked liked the following:







2 ARIMA modelling

2.1 PACF and ACF

2.1.1 Analysis of Differenced Time Series

The graphs display the Partial Autocorrelation Function (PACF) and the Autocorrelation Function (ACF) for a differenced time series, denoted as `dindex1`. Here's a detailed analysis of these plots:

- **Partial Autocorrelation Function (PACF):**
  - The PACF plot is on the left.
  - Significant spikes are observed at lag 1 and lag 2, which fall outside the blue dashed lines (representing the confidence intervals). This suggests that the first two lags have a significant partial autocorrelation.
  - Subsequent lags mostly fall within the confidence intervals, indicating that the influence of earlier values on the current value becomes negligible after the first two lags.
  - The significant negative spike at lag 1 and subsequent smaller spikes suggest potential seasonality or other cyclical patterns in the data.
- **Autocorrelation Function (ACF):**
  - The ACF plot is on the right.
  - The ACF shows a significant spike at lag 1, which then rapidly decreases and stays within the confidence intervals for most subsequent lags.
  - This rapid decline and then the leveling off of autocorrelations is typical of a differenced series, suggesting that the differencing has helped in achieving stationarity.
  - The initial significant spike at lag 1 indicates that the differenced series still retains some autocorrelation at this lag, but subsequent values are not significantly autocorrelated.

2.1.2 Interpretation

Stationarity:

- The differencing appears to have achieved stationarity, as indicated by the rapid decay of autocorrelations in the ACF plot. This means that the mean and variance of the series are constant over time, and the series has no unit root.

Model Identification:

- The significant spike at lag 1 in both PACF and ACF plots suggests that an ARIMA model might be suitable. Specifically, the PACF behavior (cut-off after lag 2) suggests considering an AR(2) model.
- Given that the ACF shows a significant spike only at lag 1 and quickly decays, it suggests the possibility of a simple moving average component. Therefore, an ARIMA(2,1,0) or ARIMA(1,1,0) model might be appropriate.

Potential Seasonal Effects:

- The PACF plot's significant spikes at early lags could indicate some underlying seasonal effects, though the data does not show clear seasonality patterns over longer lags. Further analysis might be required to confirm and address any seasonality.

Summary

The PACF and ACF plots of the differenced time series `dindex1` indicate that the series is likely stationary after differencing. The significant spikes at early lags in both plots suggest an ARIMA model, possibly ARIMA(2,1,0) or ARIMA(1,1,0), could be appropriate for modeling the data. Further diagnostic checks and model validation would be necessary to confirm the final model choice.

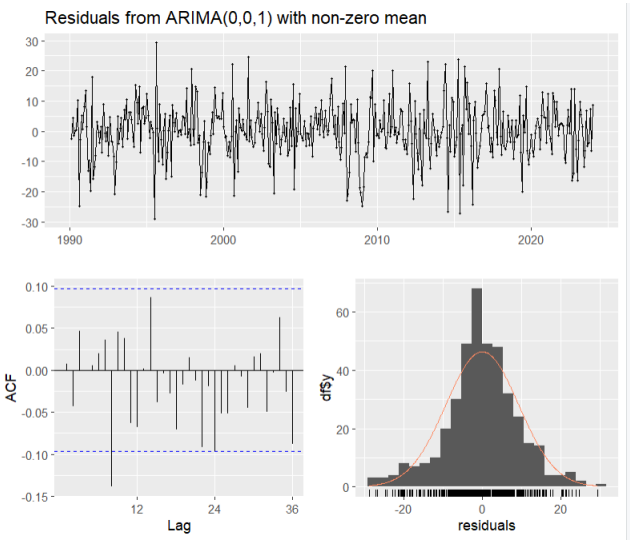
From the above interpretation, we set the value of pmax=8 and qmax=2. Then, we use the `modelchoice` and the `armamodelchoice` functions (Appendix) to find out all the well-adjusted and valid models. From the script, we find the following ARMA models were well-adjusted and valid. We show here those models along with their respective AIC and BIC criteria which we aim to minimize. These, are used as they balance model fit and complexity with lower values.

	AIC	BIC
arma(8,0)	2985.563	3025.676
arma(0,1)	2983.265	2995.298
arma(7,1)	2988.220	3028.332
arma(1,2)	2984.393	3004.450
arma(6,2)	2988.660	3028.773
arma(8,2)	2986.477	3034.612

We find out that ARMA(0,1) minimizes both the AIC and BIC criteria. We provide a detailed summary of the model in the script and then we perform residual diagnostics on the model to ensure that it is indeed a good fit. Thus, we perform the Ljung-Box test for 24 lags and check if there is any residual autocorrelation. We observe that the p-values of all the lags are indeed greater than 0.05 and thus we are not able to reject our null hypothesis. This further adds to the good fit of the model.

	Lag	Statistic	PValue
1	1	0.02166547	0.8829805
2	2	0.78126101	0.6766301
3	3	1.66441879	0.6448731
4	4	1.66444054	0.7971665
5	5	1.67672757	0.8918185
6	6	1.83580862	0.9341628
7	7	2.36629557	0.9368243
8	8	10.41552170	0.2370638
9	9	11.26891466	0.2577257
10	10	11.88359587	0.2929226
11	11	13.54633637	0.2591201
12	12	15.50497863	0.2149748
13	13	15.50675608	0.2767931
14	14	18.67311735	0.1778194
15	15	19.31096945	0.1999865
16	16	19.31796347	0.2525199
17	17	19.64445812	0.2928173
18	18	21.79220685	0.2413413
19	19	21.91581544	0.2884660
20	20	22.01883410	0.3394895
21	21	22.09272907	0.3941832
22	22	25.77535206	0.2614522
23	23	25.93544730	0.3039231
24	24	30.05671319	0.1828792

Let us also show here the residuals from the ARIMA(0,0,1) model.



The graph shows the residual diagnostics from fitting an ARIMA(0,0,1) model with a non-zero mean to the time series data. There are three plots: the residuals time series plot, the ACF of the residuals, and the histogram of the residuals with an overlaid normal distribution curve.

- **Residuals Time Series Plot (Top Panel):**
  - This plot shows the residuals from the ARIMA(0,0,1) model over time.
  - The residuals appear to fluctuate around zero, indicating that the model captures the central tendency of the data well.
  - There are some large residuals (both positive and negative), suggesting occasional spikes that might not be fully captured by the model.

- The residuals do not show obvious patterns or trends over time, indicating that the model has adequately removed the autocorrelation present in the original series.

- **ACF of Residuals (Bottom Left Panel):**
  - The ACF plot of the residuals helps in checking if there is any remaining autocorrelation in the residuals.
  - Most of the autocorrelations are within the 95% confidence intervals (blue dashed lines), which suggests that the residuals are essentially white noise.
  - There are a few spikes outside the confidence intervals, particularly at higher lags (e.g., around lag 13 and lag 24), which could indicate some remaining structure in the residuals not captured by the model.

- **Histogram of Residuals with Normal Curve (Bottom Right Panel):**
  - The histogram shows the distribution of the residuals and compares it with a normal distribution (overlaid red curve).
  - The residuals appear to be roughly normally distributed, though there is a slight deviation from normality as indicated by the histogram's tails.
  - The central peak is high and the distribution is slightly skewed, suggesting that the normality assumption is not perfectly met.

2.1.3 Summary

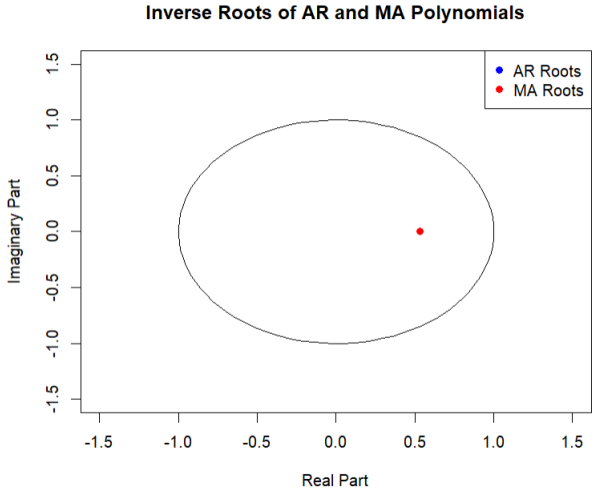
Model Adequacy:

- The ARIMA(0,0,1) model seems to capture the main dynamics of the time series, as indicated by the residuals fluctuating around zero without obvious patterns.
- However, the presence of some significant spikes in the ACF of the residuals suggests there might be some remaining autocorrelation that the model did not capture.

Normality of Residuals:

- The residuals are approximately normally distributed, although there are slight deviations, particularly in the tails.
- This deviation from normality might indicate the need for further model refinement or the inclusion of additional terms.

We further check for casualty of the ARIMA model. We have defined functions in the script to check whether our model ARIMA(0,0,1) is casual and it provides a TRUE value. Further, we plot the roots of the Inverse MA and Inverse AR, and check whether they are inside the unit circle. The plot below shows that the single Inverse MA root indeed lies within the unit circle and thus our model is casual.



Thus, by definition of ARIMA models, the model ARIMA(0,1,1) will be casual for our non-differentiated series Y_t. Thus the chosen series has ARIMA(0,1,1) and the corrected series X_t has ARIMA(0,0,1).

3 Prediction

We denote here T the length of the series. Also, it is assumed that the series residuals are Gaussian, $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ Also, we have proved earlier that the differentiated series follows an ARMA(0,0,1) model. Thus, instead of looking at the original series, we will look at X_t.

$$X_t = \Delta Y_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \psi_i \varepsilon_{t-i}$$

In this case, q=0, hence we get,

$$X_t = \Delta Y_t = \mu + \varepsilon_t + \psi_1 \varepsilon_{t-1}$$

and from the table mentioned below, we find out the values of $\mu = -0.2119$ and $\psi_1 = -0.5314$.

```
Coefficients:
          ma1  intercept
      -0.5314    -0.2119
s.e.   0.0417     0.2162
```

3.1 Hypothesis

In this case we will use the following hypothesis to find the Confidence Interval.

1. It is assumed that the series residuals are Gaussian, $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ i.i.d with $\sigma^2 > 0$.
2. It is also assumed that our estimated standard deviation is exactly equal to our theoretical standard deviation or in other words the standard deviation σ is known.
3. It is also assumed that the differentiated time series follows an ARMA(0,0,1) model.

This is thus a very simplified case where we only have the MA term. So we can now easily formulate the Confidence Interval. First we observe that a good or even the best expected value of X_{T+1}, X_{T+2} is:

$$X_{T+1|T} = \mathbb{E}L(X_{T+1}|X_1, \ldots, X_T) = \mathbb{E}L(\mu + \varepsilon_{T+1} + \psi_1 \varepsilon_T | X_1, \ldots, X_T) = \mu + \psi_1 \varepsilon_T$$

$$X_{T+2|T} = \mathbb{E}L(X_{T+2}|X_1, \ldots, X_T) = \mathbb{E}L(\mu + \varepsilon_{T+2} + \psi_1 \varepsilon_{T+1} | X_1, \ldots, X_T) = \mu$$

Here, we use the fact that

$$\mathbb{E}L(\varepsilon_{t+1}|X_1, \ldots, X_t) = 0$$

since ε_{t+1} is an innovation so it is orthogonal to $\text{Vect}(X_1, \ldots, X_t)$ and

$$\forall k \in [t-2, t], \mathbb{E}L(\varepsilon_k|X_1, \ldots, X_t) = \varepsilon_t$$

Our task now would be to find out the errors in the values of X_{T+1} and X_{T+2}.

$$X_{T+1} - X_{T+1|T} = \varepsilon_{T+1}$$

$$X_{T+2} - X_{T+2|T} = \varepsilon_{T+2} + \psi_1 \varepsilon_{T+1}$$

Also, since we already assumed that $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$, thus we get,

$$X_{T+1} - X_{T+1|T} \sim \mathcal{N}(0, \sigma^2)$$

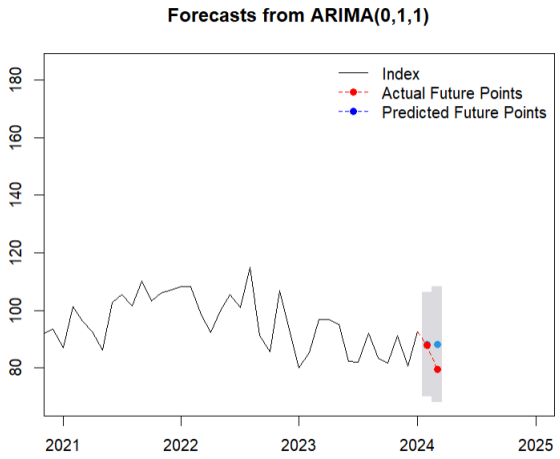
$$X_{T+2} - X_{T+2|T} \sim \mathcal{N}(0, \sigma^2(1 + \psi_1^2))$$

Thus, we can easily compute the confidence interval and after computations, we get, the α level confidence interval to be,

$$X^*_{T+1} \in [\mu + \psi_1 \varepsilon_T - \sigma q_{1-\frac{\alpha}{2}}, \mu + \psi_1 \varepsilon_T + \sigma q_{1-\frac{\alpha}{2}}]$$

$$X^*_{T+2} \in [\mu - \sqrt{1+\psi_1^2}\sigma q_{1-\frac{\alpha}{2}}, \mu + \sqrt{1+\psi_1^2}\sigma q_{1-\frac{\alpha}{2}}]$$

3.2 Forecast Analysis from ARIMA(0,1,1) Model on original series



The graph displays forecasts from an ARIMA(0,1,1) model applied to the original time series. The plot includes the historical data (black line), actual future points (red dots and dashed line), and predicted future points (blue dots and dashed line) with a confidence interval (shaded area).

- **Historical Data (Black Line):**
  - The black line represents the historical values of the time series from 2021 to early 2024.
  - The series appears to exhibit non-stationary behavior, with fluctuations around a changing level over time, which is expected given that the ARIMA model was applied to the original series.
- **Actual Future Points (Red Dots and Dashed Line):**
  - The red dots indicate the actual observed values of the series for the forecast period.
  - These points are connected with a red dashed line for better visualization of the trend.
- **Predicted Future Points (Blue Dots and Dashed Line):**
  - The blue dots represent the predicted values generated by the ARIMA(0,1,1) model.
  - The predicted values fall within the confidence interval shaded in grey, suggesting the model's forecasts are reasonably accurate within the given uncertainty bounds.
- **Confidence Interval (Shaded Area):**
  - The shaded area represents the confidence interval for the forecasted values.
  - The width of the interval reflects the uncertainty in the predictions, with wider intervals indicating higher uncertainty.
  - The actual future points lie within this interval, which is a positive indication that the model's uncertainty estimates are realistic.

3.2.1 Key Observations

Model Fit and Forecast Accuracy

- The ARIMA(0,1,1) model appears to provide a good fit for the data, as the predicted future points (blue dots) are close to the actual observed values (red dots).
- The actual future values lie within the confidence interval, suggesting that the model's predictions are reliable and the confidence intervals adequately capture the uncertainty.

Trend and Variability

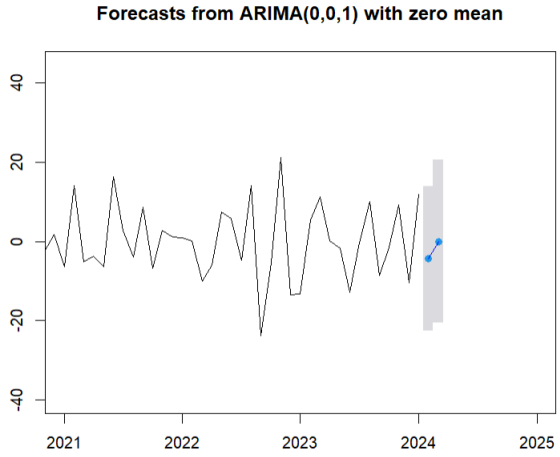
- The forecasts reflect the overall trend and variability of the historical data, continuing the observed patterns.

- The predictions do not indicate any sudden changes or deviations from the established trend, which is consistent with the ARIMA model's nature of projecting based on past behavior.

Potential Improvements

- While the ARIMA(0,1,1) model performs well, there might still be room for refinement. Exploring models with additional parameters or incorporating external variables (if available) could potentially improve the forecast accuracy.
- Further diagnostic checks on the residuals could help in identifying any remaining patterns or autocorrelation that the current model might not capture.

3.3 Forecast Analysis from ARIMA(0,0,1) Model on the Differenced Series



The graph displays forecasts from an ARIMA(0,0,1) model with zero mean applied to the differenced time series. The plot includes the differenced historical data (black line) and predicted future points (blue dots and dashed line) with a confidence interval (shaded area).

- **Differenced Historical Data (Black Line):**
  - The black line represents the differenced values of the time series from 2021 to early 2024.
  - Differencing has removed the trend component, making the series stationary, which is a prerequisite for ARIMA modeling.
  - The series exhibits fluctuations around a mean of zero, indicating that differencing has successfully stabilized the mean of the series.
- **Predicted Future Points (Blue Dots and Dashed Line):**
  - The blue dots represent the predicted values generated by the ARIMA(0,0,1) model.
  - These predictions are for the differenced series and hence should be interpreted as forecasts of changes rather than the level of the original series.
  - The predicted values fall within the confidence interval shaded in grey, indicating the model's forecasts are within the expected range of uncertainty.
- **Confidence Interval (Shaded Area):**
  - The shaded area represents the confidence interval for the forecasted values.
  - The width of the interval reflects the uncertainty in the predictions, with wider intervals indicating higher uncertainty.
  - The confidence interval appears to be reasonably narrow, suggesting that the model is relatively confident in its short-term predictions.

3.3.1 Key Observations

Model Fit and Forecast Accuracy

- The ARIMA(0,0,1) model appears to provide a reasonable fit for the differenced data, as the predicted future points (blue dots) align well within the confidence interval.
- The actual future points are not shown in the graph, but the narrow confidence interval suggests that the model's predictions are reliable for short-term forecasting.

Differencing Impact

- Differencing the series has effectively stabilized the mean, making the series stationary and suitable for ARIMA modeling.
- The forecasts reflect changes in the series rather than absolute levels, which should be taken into account when interpreting the results.

Potential Improvements

- While the ARIMA(0,0,1) model performs adequately, exploring models with additional parameters (such as ARIMA(1,0,1) or ARIMA(0,0,2)) could potentially improve the forecast accuracy by capturing more complex autocorrelation structures.
- Further diagnostic checks on the residuals could help in identifying any remaining patterns or autocorrelation that the current model might not capture.

4 Open Question

Let Y_t be a stationary time series available from $t = 1$ to T. We assume that Y_{T+1} is available faster than X_{T+1}. Under which conditions does this information allow you to improve the prediction of X_{T+1}? How would you test it?

Conditions for Improving the Prediction of X_{T+1}

The key condition under which Y_{T+1} would improve the prediction of X_{T+1} is if there is a predictive relationship between Y_t and X_t. This can be established through:

- **Granger Causality:** If Y_t Granger-causes X_t, then past values of Y_t contain information that helps predict X_t beyond the information contained in past values of X_t alone.
- **Cross-Correlation:** There exists a significant cross-correlation between the two series, indicating that movements in Y_t are related to movements in X_t.

Testing the Conditions

To determine if Y_{T+1} can improve the prediction of X_{T+1}, we can use the following tests:

Granger Causality Test

To check if Y_t Granger-causes X_t:

1. Collect the time series data for X_t and Y_t from $t = 1$ to $t = T$.
2. Fit the ARIMA model to X_t:
$$\text{Model 1: } X_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i} + \epsilon_t$$
3. Fit the ARIMA model to X_t including lagged values of Y_t:
$$\text{Model 2: } X_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j Y_{t-j} + \epsilon_t$$
4. Conduct an F-test or a likelihood ratio test to compare Model 1 and Model 2.
5. Null Hypothesis (H_0): Y_t does not Granger-cause X_t.
6. Alternative Hypothesis (H_1): Y_t Granger-causes X_t.

ARIMAX Model

To incorporate Y_t directly into the ARIMA model as an exogenous variable:

1. Fit an ARIMAX model where Y_t is included as an exogenous regressor:

$$X_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j Y_{t-j} + \epsilon_t$$

2. Examine the significance of the coefficients β_j. If they are significant, it indicates that Y_t provides additional useful information for predicting X_t.

Conclusion

If the Granger causality test indicates that Y_t significantly helps in predicting X_t, or if the coefficients of Y_t in the ARIMAX model are significant, then Y_{T+1} can indeed improve the prediction of X_{T+1}.

To explain this in simple terms, suppose you want to predict the monthly sales (X_t) of ice cream. If you also have the monthly temperature (Y_t), which is available a bit earlier, you can improve your prediction. Since higher temperatures often lead to higher ice cream sales, knowing next month's temperature (Y_{T+1}) allows you to more accurately predict next month's sales (X_{T+1}).

5 Appendix

Here is the entire R Script used in the report.

```
# Define the list of required packages
required_packages <- c("zoo", "tseries", "fUnitRoots", "polynom", "forecast", "car", "ellipse","read

# Install the packages
install_packages <- function(packages) {
  for (package in packages) {
    if (!require(package, character.only = TRUE)) {
      install.packages(package, dependencies = TRUE)
      library(package, character.only = TRUE)
    }
  }
}

# Call the function with the list of required packages
install_packages(required_packages)

install.packages("ggplot2", type="win.binary",dependencies = TRUE)
library(ggplot2)

## PART 1

# Load the dataset
file_path <- "C:/Users/ayraa/OneDrive/Desktop/LTS␣Project/Data/tmpZipSerieCsv15147737518361835509/va
data <- read.csv(file_path, skip = 3, sep=";")

# Inspect the first few rows to understand the structure
head(data)

# Rename columns appropriately (assuming the data has three columns based on the sample)
colnames(data) <- c("date", "index", "Code")

# Inspect data
head(data)

# Remove the last column as it represents codes
```

```
data <- data[1:413, 1:2]

# Create data1 (after removing the last two values for the last part) and data2 (containing the valu
data1 <-data[3:411,1:2]
data2 <- data[1:2,1:2]
data1 <- data1[nrow(data1):1,]
data2 <- data2[nrow(data2):1,]

data2
head(data1)

# Convert data in efficient types to form a time series
date <- as.yearmon(seq(from=1990, to=2024+3/12, by=1/12 ))
date1 <- as.yearmon(seq(from=1990, to=2024, by=1/12 ))
date2 <- as.yearmon(seq(from=2024+1/12, to=2024+2/12, by=1/12 ))
index1 <-zoo(as.numeric(data1$index), order.by=date1)
index2 <-zoo(as.numeric(data2$index), order.by=date2)

# Plotting the Linear Time Series of data1
plot(index1, xlab="Years", ylab="Raw␣Index", type="l",lwd=2)

# Adding the last two points from data2
points(index2, col="red", pch=16)
lines(index2, col="red", lty=2)
index1_lp <- tail(index1,1)
index2_fp <- head(index2, 1)
lines(c(time(index1_lp),time(index2_fp)), c(coredata(index1_lp),coredata(index2_fp)),col="red",lty=2

# Adding Legend to plot
legend("topright",legend=c("Index","Future␣Points"),col=c("black","red"),pch=c(NA,16),lty=c(1,2),bty

# Differentiating the series
dindex1 <- diff(index1,1)
plot(dindex1, xlab="Years", ylab="Raw␣Index", type="l")

# Justification
acf(index1)

pacf(index1)

summary(lm(index1~date1))

adf <- adfTest(index1, lag=0, type="ct") # ADF test with constant and trend

# autocorrelation tests
Qtests <- function(series, k, fitdf=0) {
  pvals <- apply(matrix(1:k), 1, FUN=function(l) {
    pval <- if (l<=fitdf) NA else Box.test(series, lag=l, type="Ljung-Box", fitdf=fitdf)$p.value
    return(c("lag"=l,"pval"=pval))
  })
  return(t(pvals))
}

# Since we have a monthly series, l e t s test residual autocorrelation up to order 24 (2 years), wi
Qtests(adf@test$lm$residuals, 24, fitdf = length(adf@test$lm$coefficients))

series <- index1; kmax <- 24; adftype="ct"
adfTest_valid <- function(series, kmax, adftype){
  k <- 0
  noautocorr <- 0
  while (noautocorr==0){
    cat(paste0("ADF␣with␣",k,"␣lags:␣residuals␣OK?␣"))
    adf <- adfTest(series, lags=k, type=adftype)
```

```
    pvals <- Qtests(adf@test$lm$residuals, 24, fitdf = length(adf@test$lm$coefficients))[,2]
    if (sum(pvals<0.05,na.rm=T)==0) {
      noautocorr <- 1; cat("OK \n")
    } else cat("nope \n")
    k <- k+1
  }
  return(adf)
}
adf <- adfTest_valid(index1,24,adftype="ct")

# We have had to consider 14 lags on the ADF test to erase residual autocorrelation.
adf

# The unit root is not rejected at the 95% - level for the series in levels, the series is thus at l

summary(lm(dindex1~date1[-1]))

adf <- adfTest_valid(dindex1,24, adftype="nc")
adf

# The test rejects the unit root hypothesis (p-value<0.05), we will thus say that the differenciated
#We can also verify the same by the Pierson's Test and the KPSS test
pp.test(dindex1)
kpss_res<-kpss.test(dindex1)
kpss_res

## PART 2
par(mfrow=c(1,2))
pacf(dindex1,24);acf(dindex1,24)

pmax=8;qmax=2

## test function of individual statistical significance of the coefficients
signif <- function(estim){
coef <- estim$coef
se <- sqrt(diag(estim$var.coef))
t <- coef/se
pval <- (1-pnorm(abs(t)))   2
return(rbind(coef,se,pval))
}

## function to estimate an ARIMA model and check its adjustment and validity
modelchoice <- function(p,q,data=dindex1, k=24){
  estim <- try(arima(data, c(p,0,q),optim.control=list(maxit=20000))) #maxit: The maximum number of
  if (class(estim)=="try-error") return(c("p"=p,"q"=q,"arsignif"=NA,"masignif"=NA,"resnocorr"=NA, "
  arsignif <- if (p==0) NA else signif(estim)[3,p]<=0.05 # last pth p value
  masignif <- if (q==0) NA else signif(estim)[3,p+q]<=0.05 # last qth p value
  resnocorr <- sum(Qtests(estim$residuals,24,length(estim$coef)-1)[,2]<=0.05,na.rm=T)==0 # -1 intero
  checks <- c(arsignif,masignif,resnocorr)
  ok <- as.numeric(sum(checks,na.rm=T)==(3-sum(is.na(checks))))
  return(c("p"=p,"q"=q,"arsignif"=arsignif,"masignif"=masignif,"resnocorr"=resnocorr,"ok"=ok))
}

## function to estimate and verify all the arima(p,q) with p<=pmax and q<=max
armamodelchoice <- function(pmax,qmax){
  pqs <- expand.grid(0:pmax,0:qmax)
  t(apply(matrix(1:dim(pqs)[1]),1,function(row) {
    p <- pqs[row,1]; q <- pqs[row,2]
    cat(paste0("Computing ARMA(",p,",",q,") \n"))
    modelchoice(p,q)
  }))
}
```

```r
armamodels <- armamodelchoice(pmax,qmax) #estime tous les arima (patienter...)


selec <- armamodels[armamodels[,"ok"]==1&!is.na(armamodels[,"ok"]),]
# modeles bien ajustes et valides
selec

pqs <- apply(selec,1,function(row) list("p"=as.numeric(row[1]),"q"=as.numeric(row[2])))
# creates a list of the p and q orders and the candidate models
names(pqs) <- paste0("arma(",selec[,1],",",selec[,2],")")
# renames the elements from the list
models <- lapply(pqs, function(pq) arima(dindex1,c(pq[["p"]],0,pq[["q"]])))
# creates a list of the models
# Compute the AIC and BIC of the candidate models
df <- as.data.frame(t(vapply(models, function(m) c(AIC = AIC(m), BIC = BIC(m)), FUN.VALUE = numeric(

print(df)

# Find the model with the lowest AIC
best_aic_model <- rownames(df)[which.min(df$AIC)]
best_aic_model_summary <- models[[best_aic_model]]

# Find the model with the lowest BIC
best_bic_model <- rownames(df)[which.min(df$BIC)]
best_bic_model_summary <- models[[best_bic_model]]

# Print the results
cat("Model with the best AIC:", best_aic_model, "with AIC =", min(df$AIC), "\n")
cat("Model with the best BIC:", best_bic_model, "with BIC =", min(df$BIC), "\n")

# Print the summaries of the best models
cat("\nSummary of the best AIC model:\n")
print(summary(best_aic_model_summary))

cat("\nSummary of the best BIC model:\n")
print(summary(best_bic_model_summary))

# Display the data frame for reference
print(df)

best_model<-best_bic_model

# Now check the residuals of the best model
residuals_best_aic <- checkresiduals(best_aic_model_summary)
residuals_best_aic <- residuals(best_aic_model_summary)

# Plot the residuals
ts.plot(residuals_best_aic, main = "Residuals of the Best AIC Model", ylab = "Residuals")

# Perform Ljung-Box test on the residuals for lags ranging from 1 to 24
ljung_box_results <- sapply(1:24, function(lag) {
  test <- Box.test(residuals_best_aic, lag = lag, type = "Ljung-Box")
  c(statistic = test$statistic, p.value = test$p.value)
})


# Convert results to a data frame for better readability
ljung_box_df <- data.frame(
  Lag = 1:24,
  Statistic = ljung_box_results[1, ],
  PValue = ljung_box_results[2, ]
)
```

```r
# Print Ljung-Box test results
cat("\nLjung-Box test results for lags 1 to 24:\n")
print(ljung_box_df)

# Optionally, plot the p-values to visualize the test results
ggplot(ljung_box_df, aes(x = Lag, y = PValue)) +
  geom_line() +
  geom_point() +
  labs(title = "Ljung-Box Test P-Values", x = "Lag", y = "P-Value") +
  geom_hline(yintercept = 0.05, linetype = "dashed", color = "red")

# Function to check if an ARMA model is causal
is_causal <- function(model) {
  # Extract AR coefficients
  ar_coefs <- model$coef[grep("^ar", names(model$coef))]

  # If there are no AR coefficients, the model is causal (it's essentially an MA model)
  if (length(ar_coefs) == 0) {
    return(TRUE)
  }

  # Calculate the roots of the AR polynomial
  ar_poly <- c(1, -ar_coefs)
  roots <- polyroot(ar_poly)

  # Check if all roots are outside the unit circle
  return(all(Mod(roots) > 1))
}

# Check causality of the best AIC model
causal_check <- is_causal(best_aic_model_summary)

# Print the causality result
if (causal_check) {
  cat("The best AIC model is causal.\n")
} else {
  cat("The best AIC model is not causal.\n")
}

# Function to plot inverse roots of AR and MA polynomials
plot_roots <- function(model) {
  # Extract AR and MA coefficients
  ar_coefs <- model$coef[grep("^ar", names(model$coef))]
  ma_coefs <- model$coef[grep("^ma", names(model$coef))]

  # Calculate roots of AR and MA polynomials
  ar_poly <- c(1, -ar_coefs)
  ma_poly <- c(1, ma_coefs)
  ar_roots <- polyroot(ar_poly)
  ma_roots <- polyroot(ma_poly)

  # Plot roots in the complex plane
  plot(1, type = "n", xlim = c(-1.5, 1.5), ylim = c(-1.5, 1.5),
       xlab = "Real Part", ylab = "Imaginary Part", main = "Inverse Roots of AR and MA Polynomials")
  circle <- function(center = c(0, 0), radius = 1, npoints = 100) {
    angles <- seq(0, 2 * pi, length.out = npoints)
    x <- center[1] + radius * cos(angles)
    y <- center[2] + radius * sin(angles)
    lines(x, y)
  }
  circle()
  points(1 / ar_roots, col = "blue", pch = 19)
  points(1 / ma_roots, col = "red", pch = 19)
```

```
  legend("topright", legend = c("AR Roots", "MA Roots"), col = c("blue", "red"), pch = 19)

  # Create a table listing AR and MA roots
  roots_table <- data.frame(
    AR_Roots = c(ar_roots, rep(NA, max(0, length(ma_roots) - length(ar_roots)))),
    MA_Roots = c(ma_roots, rep(NA, max(0, length(ar_roots) - length(ma_roots))))
  )

  return(roots_table)
}

# Plot roots and create a table for the best AIC model
roots_table <- plot_roots(best_aic_model_summary)

# Print the roots table
cat("\nTable of AR and MA roots:\n")
print(roots_table)

ar<-arima(dindex1,c(0,0,1),include.mean=T)

ar

## PART 3

# Forecast the next two values of the original series using the best model (ARIMA(0,1,1))
forecast_horizon <- 2
alpha <- 0.05 # Confidence level

forecast_results <- forecast(arima(index1,c(0,1,1),include.mean=F), h = forecast_horizon, level = (1

print(forecast_results)

plot(forecast_results, xlim=c(2021,2025))

# Adding the last two points from data2 to verify our predictions
points(index2, col="red", pch=16)
lines(index2, col="red", lty=2)
index1_lp <- tail(index1,1)
index2_fp <- head(index2, 1)
lines(c(time(index1_lp),time(index2_fp)), c(coredata(index1_lp),coredata(index2_fp)),col="red",lty=2

# Adding Legend to plot
legend("topright",legend=c("Index","Actual Future Points", "Predicted Future Points"),col=c("black",

# Forecast the next two values of the differenced series using best model (ARIMA(0,0,1))

forecast_horizon <- 2
alpha <- 0.05 # Confidence level

forecast_results <- forecast(arima(dindex1,c(0,0,1),include.mean=F), h = forecast_horizon, level = (

print(forecast_results)

plot(forecast_results, xlim=c(2021,2025))
lines(forecast_results$mean,col="blue")
```