LINEAR TIME SERIES ASSIGNMENT

ARIMA Modelling and Prediction

Eva Mukherjee and Maria Joseph

Professor:

Jean-David Fermanian

TD Manager :

Daniel Nkameni

1 PART I: The Data

1.1 Representation of Series

We are choosing to study the **Industrial Production Index (IPI) for Cocoa, Chocolate, and Confectionery Products (CVS-CJO)** in France. The data spans from January 1990 to March 2025. This index reflects the volume of production in this specific sector of the French industry, adjusted for seasonal variations and working days. The data is available on a monthly frequency and comprises 423 observations, satisfying the requirement of having at least 100 observations.

The series shows noticeable fluctuations and a long-term non-linear trend (can be seen in appendix 2). During the 1990s and early 2000s, production levels remained relatively stable with moderate upward movement. From around 2005 to 2014, the index showed a slight downward trend, which could reflect changes in consumer preferences, production constraints, or broader industry changes.

A significant recovery begins in the mid-2010s, with production steadily increasing through to the end of the observed period. This recent growth may be related to the increase in demand for high-quality chocolate, increased exports, or advances in manufacturing processes. The index also reflects the effects of the COVID-19 pandemic, with a sharp drop in early 2020 coinciding with national lockdowns. However, this decline was short-lived, followed by a rapid recovery that led the index to reach record levels in 2022–2023, surpassing 110.

Visual inspection of the series suggests potential issues with stationarity, as both the mean and variance appear to evolve over time. A formal stationarity analysis will be conducted in later sections.

1.2 Stationarity of Series

Time series stationarity is crucial for applying ARIMA models. A stationary series has a constant mean, variance, and autocorrelation structure over time.As mentioned earlier, the series we are working with has already been adjusted for seasonal variations and working days. However, as a precaution, we verify that it indeed does not exhibit any seasonal effects. To confirm this, we look that the month of the year does not influence the distribution (boxplots) of the series (in appendix 2), confirming the absence of seasonality.

Additionally, the moving averages shown in the original series plot(appendix 2) appear to vary over time, leading us to suspect non-stationarity in the series. We therefore proceed with statistical tests to verify this. These tests will also help us identify the type of non-stationarity, particularly whether it is stochastic (Difference-Stationary) or deterministic. Indeed, knowing the type of non-stationarity is crucial for selecting the appropriate method to make the series stationary.

The stationarity of the original series was assessed using the following statistical tests:

We interpret the results for each of the stationarity tests provided for the original series data. A common significance level (α) of 0.05 will be used for decision making.

KPSS Test for Trend Stationarity

```
            KPSS Test for Trend Stationarity

data:  base$Index_CVS_CJO
KPSS Trend = 0.31826, Truncation lag parameter = 5, p-value = 0.01
```

- **Null Hypothesis (H_0):** The series is stationary around a deterministic trend (i.e., trend stationary).
- **Alternative Hypothesis (H_1):** The series has a unit root (i.e., is not trend stationary, implying a stochastic trend).

- **Result:** KPSS Trend = 0.31826, Truncation lag parameter = 5, p-value = 0.01
- **Interpretation:** Since the p-value (0.01) is less than the significance level of 0.05, we **reject the null hypothesis**. This suggests that the series is **not trend stationary** and likely contains a **stochastic trend (unit root)**.

KPSS Test for Level Stationarity

```
        KPSS Test for Level Stationarity

data:  base$Index_CVS_CJO
KPSS Level = 1.0963, Truncation lag parameter = 5, p-value = 0.01
```

- **Null Hypothesis (H_0):** The series is level stationary (i.e., stationary around a constant mean, with no trend at all).
- **Alternative Hypothesis (H_1):** The series has a unit root (i.e., is not level stationary).
- **Result:** KPSS Level = 1.4345, Truncation lag parameter = 5, p-value = 0.01
- **Interpretation:** Similar to the trend stationarity test, the p-value (0.01) is less than 0.05, leading us to **reject the null hypothesis**. This indicates that the series is **not level stationary** and also likely contains a **stochastic trend**.

Augmented Dickey-Fuller Test (ADF)

```
       Augmented Dickey-Fuller Test

data:  base$Index_CVS_CJO
Dickey-Fuller = -3.7008, Lag order = 7, p-value = 0.02411
alternative hypothesis: stationary
```

- **Null Hypothesis (H_0):** The series has a unit root (i.e., is non-stationary).
- **Alternative Hypothesis (H_1):** The series is stationary or trend-stationary.
- **Result:** Dickey-Fuller = -3.7008, Lag order = 7, p-value = 0.02411
- **Interpretation:** With a p-value (0.02411) less than 0.05, we **reject the null hypothesis**. This suggests that the series is **stationary**. The alternative hypothesis for the ADF test encompasses both strict stationarity and trend stationarity.

Phillips-Perron Unit Root Test (PP)

```
        Phillips-Perron Unit Root Test

data:  base$Index_CVS_CJO
Dickey-Fuller Z(alpha) = -229.26, Truncation lag parameter = 5, p-value = 0.01
alternative hypothesis: stationary
```

- **Null Hypothesis (H_0):** The series has a unit root (i.e., is non-stationary).
- **Alternative Hypothesis (H_1):** The series is stationary.
- **Result:** Dickey-Fuller Z(alpha) = -229.26, Truncation lag parameter = 5, p-value = 0.01
- **Interpretation:** The p-value (0.01) is less than 0.05, leading us to **reject the null hypothesis**. This also suggests that the series is **stationary**. The Phillips-Perron test is known for its robustness to general forms of heteroskedasticity and serial correlation in the error term.

Combined Conclusion

The individual test results present a seemingly contradictory picture:

- Both **KPSS tests** (Trend and Level) reject their null hypotheses of stationarity, strongly indicating the presence of a **stochastic trend (unit root)**.
- Both **ADF and Phillips-Perron tests** reject their null hypotheses of a unit root, suggesting that the series is **stationary** (either strictly stationary or trend-stationary).

When unit root tests (ADF, PP) suggest stationarity, but stationarity tests (KPSS) suggest non-stationarity, the most common and plausible interpretation is that the series is **trend-stationary**.

Therefore, the most likely conclusion for the original series is that it has a **deterministic trend**. This means that while the series exhibits a clear upward or downward movement over time, this movement is predictable and can be modeled. Once this deterministic trend is taken into account and removed, the remaining series is stationary. The implications are that any shocks to the series will have only temporary effects, and the series will eventually revert to its underlying deterministic trend path.

To try and address the deterministic trend we detrend the series by the method of linear detrending via OLS regression.

Let's define the components of the time series and the linear trend model:

- X_t = Original time series, representing the value at time t.
- Y_t = Time index, which is a numeric sequence (e.g., $1, 2, \ldots, T$), where T is the total number of observations.

The linear trend model is formulated as a simple linear regression equation:

$$X_t = \beta_0 + \beta_1 Y_t + \epsilon_t$$

where:

- β_0 = Intercept, representing the estimated starting level of the series when the time index Y_t is zero.
- β_1 = Slope, representing the linear trend coefficient, which indicates the average change in X_t for a one-unit increase in the time index Y_t.
- ϵ_t = Residuals, which represent the portion of the original series that is not explained by the linear trend. This is precisely the detrended series.

The detrended series (ϵ_t) is computed by subtracting the estimated linear trend component ($\hat{\beta}_0 + \hat{\beta}_1 Y_t$) from the original series (X_t):

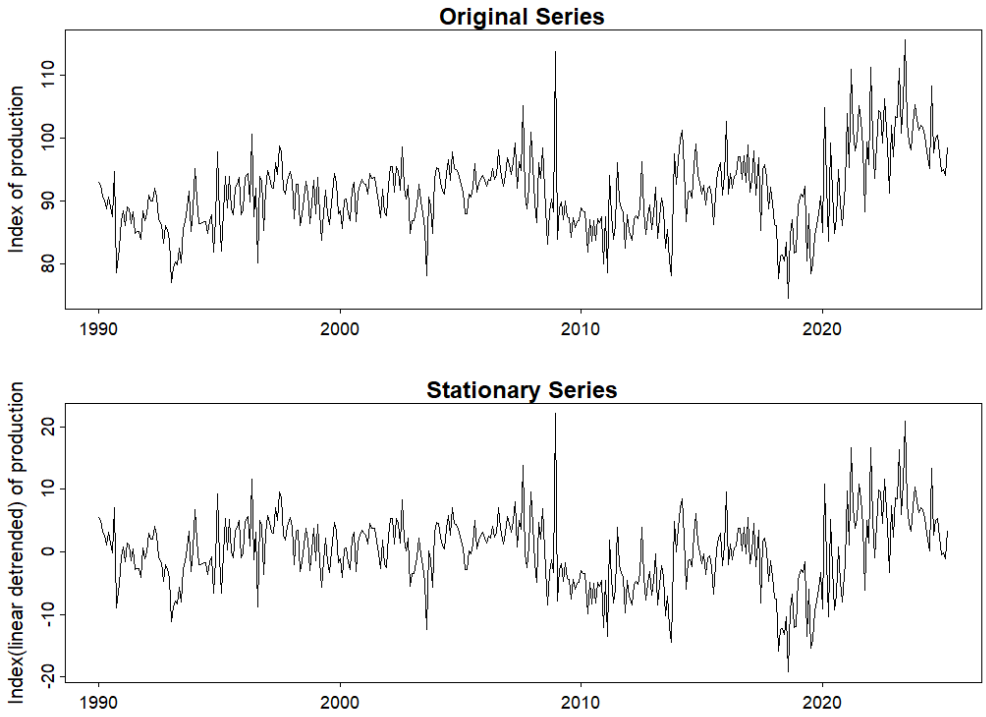
$$\epsilon_t = X_t - (\hat{\beta}_0 + \hat{\beta}_1 Y_t)$$

Here, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated coefficients from the linear regression.

On re-running the stationarity tests on the residuals of the detrended series. We can see that the residuals are found to be level stationary.

```
    Augmented Dickey-Fuller Test

data:  detrended
Dickey-Fuller = -3.7008, Lag order = 7, p-value = 0.02411
alternative hypothesis: stationary

    KPSS Test for Level Stationarity

data:  detrended
KPSS Level = 0.31826, Truncation lag parameter = 5, p-value = 0.1
```

1.3 Graphical representation of series



2 PART II: ARMA Models

2.1 Selection and Justification of the ARMA Model

In order to select the appropriate ARMA (p,q) model for the series, we will observe the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots (seen in appendix 2).

- **ACF of Detrended Series:**

  The ACF plot shows significant positive spikes at lags 1, 2, 3, and 4. There are also noticeable significant spikes around lag 12 and its multiples, suggesting a seasonal pattern. The ACF does not decline after lag 1; it slowly converges to zero with notable positive values for several initial lags and clear seasonal autocorrelation.

- **PACF of Detrended Series:**

  The PACF plot shows significant positive spikes at lags 1, 2, 3, 4, 5, and 6. It does not decline after lag 1.

Based on these patterns, indicating complex AR and MA components and a potential seasonal influence, we explored a range of ARIMA(p,0,q) models where p ranges from 0 to 3 and q ranges from 0 to 2. It is important to note that these models were fitted directly on the original series (base$Index_CVS_CJO) with d=0, despite the non-stationarity indicated by the ADF and KPSS tests. The `sarima` function from the `astsa` package was used for model estimation.

Model Selection using AIC and BIC: After fitting multiple ARIMA(p,0,q) models, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used to select the best model.

These criteria balance model fit with complexity, penalizing models with more parameters. Lower AIC and BIC values indicate a more concise and better-fitting model.

The AIC and BIC values for the tested models are presented in Table 1 in Appendix 2.

As shown in Table 1 , `model8`, which is an **ARIMA(1,0,1)** model, yielded the lowest AIC (2484.747) and BIC (2500.936) values. This indicates that ARIMA(1,0,1) is the most suitable model among those tested.

Model Validity Check (Ljung-Box Test and Residual Analysis): The validity of the selected model was checked by analyzing the residuals. For a valid ARIMA model, the residuals should ideally be white noise (i.e., uncorrelated, zero mean, and constant variance). The Ljung-Box test is used to check for autocorrelation in the residuals.

The Ljung-Box test results (p-values for various lags) and `checkresiduals` output for `model8` (ARIMA(1,0,1)) are detailed in Table 2 in Appendix 2.

- All p-values are well above the significance level of 0.05, indicating that we fail to reject the null hypothesis of no autocorrelation in the residuals. This suggests that the residuals are indeed white noise.
- `checkresiduals` **Output:**

```
Ljung-Box test

data:  Residuals from ARIMA(1,0,1) with zero mean
Q* = 11.625, df = 8, p-value = 0.1687

Model df: 2,   Total lags used: 10
```

  This confirms the Ljung-Box test result, with a p-value of 0.1687 for 10 lags (df=8 after accounting for the AR(1) and MA(1) parameters). The residuals also show no obvious patterns in their ACF/PACF plots and they appear to be centered around zero.

Conclusion or Model Selection: Based on the lowest AIC and BIC values and the successful validation of residuals (no significant autocorrelation), the ARIMA(1,0,1) model is chosen as the most appropriate model for the industrial production index series.

2.2 Expression of ARIMA(p,d,q) model for the chosen series.

Based on the AIC and BIC values, the chosen model for the industrial production index series is ARIMA(1,0,1).

The coefficients for this model are as follows:

- AR(1) coefficient (ϕ_1): 0.9406
- MA(1) coefficient (θ_1): -0.6774
- Intercept (xmean): 0.1667
- Estimated variance of the residuals (σ^2): 20.26

The general form of an ARIMA(p,d,q) model is given by:

$$(1 - \sum_{i=1}^{p} \phi_i B^i)(1 - B)^d X_t = \delta + (1 + \sum_{j=1}^{q} \theta_j B^j)\epsilon_t$$

where X_t is the time series, B is the backshift operator, ϕ_i are the autoregressive coefficients, θ_j are the moving average coefficients, δ is the constant term (intercept), d is the order of differencing, and ϵ_t is a white

noise error term with mean 0 and variance σ^2.

For the chosen ARIMA(1,0,1) model, we have $p = 1$, $d = 0$, and $q = 1$. Substituting the estimated coefficients, the model equation is:

$$(1 - 0.9406B)X_t = 0.1667 + (1 - 0.6774B)\epsilon_t$$

Expanding the equation, we get:

$$X_t - 0.9564X_{t-1} = 91.6074 + \epsilon_t - 0.6774\epsilon_{t-1}$$

Rearranging the terms, the final equation for the ARIMA(1,0,1) model is:

$$X_t = 91.6074 + 0.9406X_{t-1} + \epsilon_t - 0.6774\epsilon_{t-1}$$

Here, ϵ_t represents the white noise error term, which is assumed to be independently and identically distributed with a mean of 0 and an estimated variance $\hat{\sigma}^2 = 20.39054$.

3 PART III

Denote T the length of the series. Assume the series' residuals are Gaussian.

3.1 Confidence Region for Future Values

Given the linear trend model $X_t = \beta_0 + \beta_1 Y_t + \epsilon_t$, we are interested in forecasting future values X_{T+1} and X_{T+2}.

Equation for Prediction Intervals (Confidence Region for Individual Future Values)

For any two-step-ahead forecast $\big(X_{T+1}, X_{T+2}\big)$, the $100(1-\alpha)\%$ joint confidence region is the set of all $\big(x_{T+1}, x_{T+2}\big)$ satisfying the quadratic form:

$$\big(\mathbf{x} - \boldsymbol{\mu}\big)^\top \Sigma^{-1} \big(\mathbf{x} - \boldsymbol{\mu}\big) \leq \chi^2_{2,1-\alpha},$$

where

$$\mathbf{x} = \begin{pmatrix} x_{T+1} \\ x_{T+2} \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} \hat{X}_{T+1|T} \\ \hat{X}_{T+2|T} \end{pmatrix},$$

and $\chi^2_{2,1-\alpha}$ is the $1-\alpha$ quantile of a χ^2–distribution with 2 degrees of freedom.
with parameter estimates:

$$\delta = 0.1667, \quad \phi_1 = 0.9406, \quad \theta_1 = -0.6774, \quad \sigma^2 = 20.26.$$

Forecast Means

The one-step and two-step ahead forecasts, setting future shocks to zero, are given by:

$$\hat{X}_{T+1|T} = \delta + \phi_1 X_T,$$

$$\hat{X}_{T+2|T} = \delta + \phi_1 \hat{X}_{T+1|T}.$$

Let

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} \hat{X}_{T+1|T} \\ \hat{X}_{T+2|T} \end{pmatrix}.$$

Forecast Error Covariance Matrix

Define the impulse response function:

$$\psi(B) = \frac{1+\theta_1 B}{1-\phi_1 B} = \sum_{j=0}^{\infty} \psi_j B^j,$$

with

$$\psi_0 = 1, \quad \psi_1 = \phi_1 + \theta_1 = 0.9406 + (-0.6774) = 0.2632.$$

The forecast error vector is:

$$\begin{pmatrix} X_{T+1} - \hat{X}_{T+1|T} \\ X_{T+2} - \hat{X}_{T+2|T} \end{pmatrix} = \begin{pmatrix} \psi_0 \varepsilon_{T+1} \\ \psi_0 \varepsilon_{T+2} + \psi_1 \varepsilon_{T+1} \end{pmatrix},$$

so the covariance matrix is:

$$\Sigma = \sigma^2 \begin{pmatrix} \psi_0^2 & \psi_0\psi_1 \\ \psi_0\psi_1 & \psi_0^2 + \psi_1^2 \end{pmatrix} = 20.26 \begin{pmatrix} 1 & 0.2632 \\ 0.2632 & 1 + 0.2632^2 \end{pmatrix} \approx \begin{pmatrix} 20.2600 & 5.3340 \\ 5.3340 & 21.6601 \end{pmatrix}.$$

Joint Confidence Region

The $100(1-\alpha)\%$ joint confidence region for (X_{T+1}, X_{T+2}) is the set:

$$\mathcal{E}_\alpha = \left\{ \mathbf{x} \in \mathbb{R}^2 \ : \ (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \le \chi^2_{2,1-\alpha} \right\},$$

where $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ and $\chi^2_{2,1-\alpha}$ is the $1-\alpha$ quantile of the chi-squared distribution with 2 degrees of freedom.

With $\alpha = 0.05$, we use:

$$\chi^2_{2,0.95} = 5.991,$$

so the 95% confidence ellipse is:

$$\mathcal{E}_{0.05} = \left\{ (x_1, x_2) \in \mathbb{R}^2 \ : \ \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^\top \begin{pmatrix} 20.2600 & 5.3340 \\ 5.3340 & 21.6601 \end{pmatrix}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \le 5.991 \right\}.$$

This defines the confidence ellipse for the joint forecast (X_{T+1}, X_{T+2}).

3.2 Hypotheses used to get the above region

To obtain the joint confidence region for the two-step ahead forecasts (X_{T+1}, X_{T+2}) of your ARIMA(1,0,1) model, several key hypotheses are made:

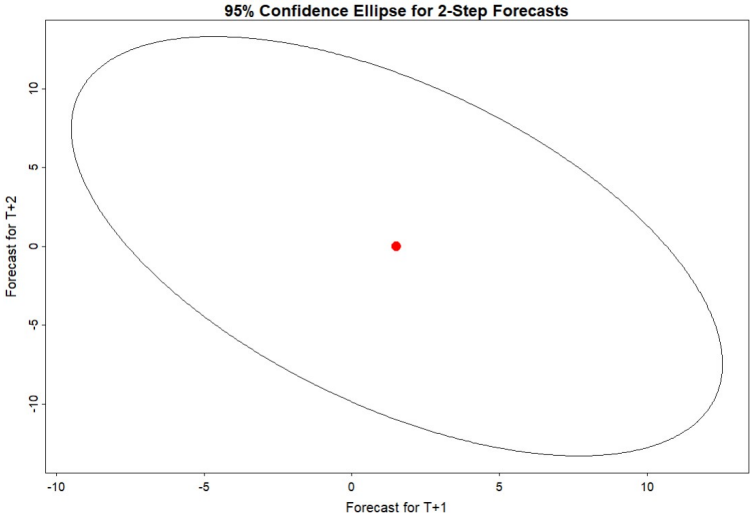
- **Correct Model Specification:** It is assumed that the chosen ARIMA(1,0,1) model accurately represents the underlying data-generating process of the time series X_t. This means the true relationships between X_t, X_{t-1}, ε_t, and ε_{t-1} are correctly captured by the model structure.
- **Stationarity and Invertibility of the Model:**
  - The AR coefficient (ϕ_1) must satisfy the **stationarity condition** (i.e., $|\phi_1| < 1$). This ensures that the series X_t (or the detrended series if d was non-zero) has a constant mean, variance, and autocorrelation structure over time.
  - The MA coefficient (θ_1) must satisfy the **invertibility condition** (i.e., $|\theta_1| < 1$). This ensures that the MA process can be represented as an infinite AR process, which is important for unique parameter estimation and for the properties of the impulse response function.
- **Residuals are Independent and Identically Distributed (i.i.d.) White Noise:** The error terms (ε_t) are assumed to be uncorrelated across time, have a constant mean of zero, and a constant variance (σ^2). This implies that all the predictable information in the series has been captured by the ARIMA model.

- **Normality of Residuals:** This is the most crucial assumption for constructing a confidence region using the χ^2 distribution. It is assumed that the error terms (ε_t) follow a normal (Gaussian) distribution, i.e., $\varepsilon_t \sim N(0, \sigma^2)$. If the errors are not normal, the chi-squared distribution for the quadratic form of the forecast errors may not be appropriate, and the confidence region's coverage might be inaccurate.

- **Parameter Estimates Are Assumed to Be True (or large sample approximation):** When constructing the confidence region, the estimated parameters (ϕ_1, θ_1, δ, σ^2) are treated as if they are the true population parameters. In practice, these are estimates, and for very small sample sizes, the uncertainty in these estimates could affect the true coverage of the confidence region. However, for sufficiently large sample sizes, the impact of parameter estimation error on forecast error variance is often considered negligible, or more complex methods are used to account for it.

- **Positive Definiteness of Covariance Matrix:** The forecast error covariance matrix (Σ) is assumed to be positive definite, which is guaranteed if the estimated residual variance (σ^2) is strictly positive. This ensures that the inverse Σ^{-1} exists and that the quadratic form defines a valid ellipse.

3.3 Graphical represenation of $\alpha = 95\%$



The plot above displays the 95% joint confidence region for the two-step-ahead forecasts, X_{T+1} and X_{T+2}, as an ellipse. This ellipse represents the set of all probable pairs of future values, given the fitted ARIMA(1,0,1) model and the assumption of Gaussian residuals.

- The center of the ellipse corresponds to the point $(\hat{X}_{T+1|T}, \hat{X}_{T+2|T})$, which are the point forecasts for the next two periods. These are the most likely values for X_{T+1} and X_{T+2} based on the model.

- The elliptical shape of the confidence region is due to the correlation between the forecast errors for X_{T+1} and X_{T+2}, as captured by the off-diagonal elements of the covariance matrix Σ. The covariance is $\psi_0\psi_1\sigma^2 = 5.3340$. Since this value is positive, the ellipse is tilted, indicating a positive correlation between the forecast errors. This means that if the forecast for X_{T+1} is higher than expected, the forecast for X_{T+2} is also likely to be higher than expected, and vice versa.

- The size of the ellipse reflects the uncertainty associated with the joint forecasts. A larger ellipse implies greater uncertainty in the predictions, while a smaller ellipse indicates more precise forecasts. The spread along the axes of the ellipse is related to the variances of the individual forecast errors, while the tilt and eccentricity are related to the covariance.

- The 95% confidence level means that if we were to repeatedly sample new data and construct such confidence regions, approximately 95% of these regions would contain the true, unobserved future values (X_{T+1}, X_{T+2}).

3.4 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}.

Conditions for Improving the Prediction of X_{T+1}:
The information from Y_{T+1} can improve the prediction of X_{T+1} under the condition that Y_t has predictive power for X_t beyond what is already captured by the past values of X_t itself. This concept is formally known as Granger causality.

Specifically, for Y_{T+1} to improve the prediction of X_{T+1}, the following conditions should ideally be met:

1. **Granger Causality:** Y_t must Granger-cause X_t. This means that lagged values of Y_t provide statistically significant information about future values of X_t that is not already contained in the lagged values of X_t.
2. **Timeliness of Y_{T+1}:** As stated, Y_{T+1} must be available faster than X_{T+1}. This is crucial because if Y_{T+1} is only known at the same time or later than X_{T+1}, it cannot be used for forecast of X_{T+1}.
3. **Stationarity of Y_t:** The problem statement specifies that Y_t is stationary. If X_t is also stationary, then standard regression or multivariate time series models can be applied directly. If one or both were non-stationary, cointegration might be a relevant concept, but since Y_t is stationary, Granger causality is the primary condition.
4. **No Redundancy:** The predictive power of Y_t for X_t should not be entirely redundant with the information already present in X_t. If X_t is already highly predictable by its own past, the marginal improvement from Y_t might be small.

How to Test These Conditions:
To test whether Y_t can improve the prediction of X_t, we can follow these steps:

1. **Granger Causality Test:**
   - **Formulate Hypotheses:**
     - Null Hypothesis (H_0): Y_t does not Granger-cause X_t. (Lagged values of Y_t do not help predict X_t given lagged X_t).
     - Alternative Hypothesis (H_1): Y_t Granger-causes X_t. (Lagged values of Y_t do help predict X_t given lagged X_t).
   - **Procedure:**
     (a) Estimate a restricted regression model for X_t using only its own lagged values:
     $$X_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i} + \epsilon_t$$
     where p is the optimal lag length for X_t.
     (b) Estimate an unrestricted regression model for X_t using its own lagged values and lagged values of Y_t:
     $$X_t = \beta_0 + \sum_{i=1}^{p} \beta_i X_{t-i} + \sum_{j=1}^{q} \gamma_j Y_{t-j} + u_t$$
     where q is the optimal lag length for Y_t.
     (c) Perform an F-test to check if the coefficients of the lagged Y_t terms (γ_j) are jointly significantly different from zero. A small p-value (typically less than 0.05) would lead to the rejection of the null hypothesis, suggesting that Y_t Granger-causes X_t.

- **Considerations for Lag Lengths (p and q):** The choice of lag lengths is crucial. Information criteria like AIC or BIC can be used to select optimal lag lengths for both the restricted and unrestricted models.

2. **Incorporating Information for Forecasting:** If Granger causality is established, there are several ways to incorporate Y_t into the forecasting model for X_t:

   - **ARIMAX Model:** Extend the ARIMA(1,0,1) model to an ARIMAX model, which includes exogenous variables.

     $$(1 - \phi_1 B)X_t = \delta + (1 - \theta_1 B)\epsilon_t + \sum_{j=0}^{k} \omega_j B^j Y_{t-m}$$

     Here, ω_j are coefficients for the exogenous variable Y_t, and m represents the lag at which Y_t impacts X_t. Since Y_{T+1} is available, we could potentially include Y_{T+1} directly if $m = 0$, or lagged values if $m > 0$.

   - **Vector Autoregression (VAR) Model:** If X_t also has predictive power for Y_t (i.e., a feedback loop), a VAR model would be more appropriate. A VAR model treats both X_t and Y_t as endogenous variables:

     $$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \mathbf{c} + \sum_{i=1}^{p} \mathbf{\Phi}_i \begin{pmatrix} X_{t-i} \\ Y_{t-i} \end{pmatrix} + \mathbf{e}_t$$

     where $\mathbf{\Phi}_i$ are matrices of coefficients. This approach allows for simultaneous modeling of the dynamic relationship between the two series.

   - **Transfer Function Model:** If the relationship is unidirectional (Y influences X, but X does not influence Y), a transfer function model (a more general form of ARIMAX) could be used to explicitly model the dynamic impact of Y_t on X_t.

By performing the Granger causality test, we can statistically assess whether Y_t adds significant predictive value to X_t. If it does, incorporating it into an appropriate model like ARIMAX or VAR will likely improve the accuracy of your forecasts for X_{T+1}.

4 APPENDIX

4.1 Appendix 1

Listing 1: Project Code

```
rm(list=ls(all=TRUE))
library(ggplot2)
library(zoo)
library(readr)
library(lubridate)
library(forecast)
library(tsoutliers) # Load the tsoutliers package
library(tseries)
library(astsa)
library(portes)
library(dplyr)

# Load and prepare data
valeurs_mensuelles <- read_delim("valeurs_mensuelles.csv", delim = ";")
View(valeurs_mensuelles)
base <- valeurs_mensuelles[-c(1:3),c(1:2)]
colnames(base) <- c("Period","Index_CVS_CJO")
View(base)
base$Index_CVS_CJO <- as.numeric(base$Index_CVS_CJO)
base$Period <- as.Date(paste0(base$Period, "-01"))

qplot(y=base$Index_CVS_CJO, x =as.yearmon(base[[1]]), geom = c("point","
    smooth","line"), xlab = "Period", ylab = "Index of production of cocoa,
    chocolate and confectionary products")
boxplot(base$Index_CVS_CJO ~ month(base$Period),
        col = "lightblue", # Color of the boxes
        pch = 20,          # Plotting character for outliers
        cex = 0.5,         # Size of outliers
        main = "Monthly Boxplots of Production Index", # Title
        ylab = "Index (Base 100 in 2021)",            # Y-axis label
        xlab = "Month")

# Convert to time series object
start_year <- year(base$Period[1])
start_month <- month(base$Period[1])
ts_data <- ts(base$Index_CVS_CJO, start = c(start_year, start_month), frequency =
    12)

# Decompose time series
ts_data_dec <- decompose(ts_data, type = "multiplicative")
autoplot(ts_data_dec, main = "Decomposition of the CVS-CJO Series", xlab = "Year")

# Question 2

# Stationarity tests
adf_test_result <- adf.test(base$Index_CVS_CJO)
print(adf_test_result)
kpss_test_result <- kpss.test(base$Index_CVS_CJO)
print(kpss_test_result)

# Removing deterministic trend based on ADF and KPSS test results
time_index <- time(ts_data)
trend_model <- lm(ts_data ~ time_index)
```

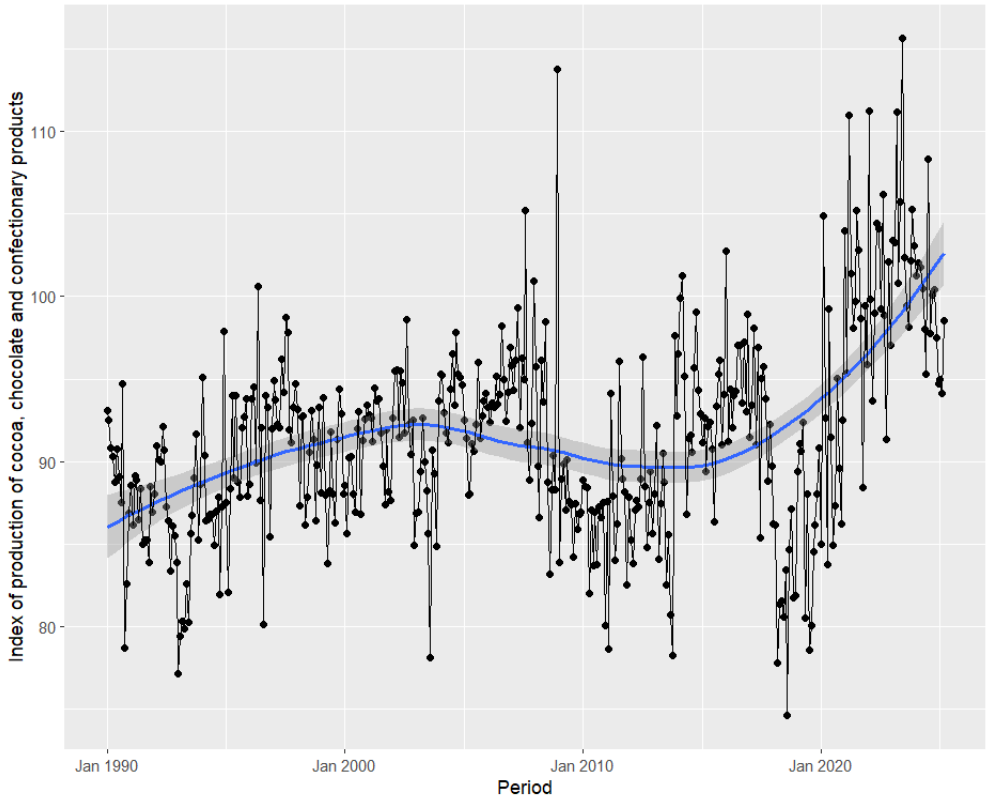
```r
detrended_numeric <- residuals(trend_model) # Store as numeric first

# Convert detrended_numeric back to a time series object
detrended_ts_data <- ts(detrended_numeric, start = start(ts_data), frequency =
    frequency(ts_data))

adf.test(detrended_ts_data)  # Expect p < 0.05 (stationary)
kpss.test(detrended_ts_data, null = "Level")  # Expect p > 0.05 (stationary)
kpss.test(detrended_ts_data, null = "Trend")

#Question 3

par(mfrow=c(2,1))

plot(y=base$Index_CVS_CJO, x =as.yearmon(base[[1]]),type = "line", main="Original
    Series", xlab = "", ylab = "Index of production")

plot(y=detrended_ts_data, x =as.yearmon(base[[1]]),type = "line", main="Stationary
     Series", xlab = "", ylab = "Index(linear detrended) of production")

par(mfrow=c(1,1))

##PART II

# Visualize ACF and PACF of the transformed series to help identify ARMA orders (
    useful for Part 2)
# Visualize ACF and PACF of the transformed (detrended) series to help identify
    ARMA orders (useful for Part 2)
acf(detrended_ts_data, main = "ACF of Detrended Series", lag.max = 36)
pacf(detrended_ts_data, main = "PACF of Detrended Series", lag.max = 36)

# ACF and PACF remind us of an AR(3) and an MA(2). We will test all ARMA models
    such that
#p<=3 and q<=2
model1=sarima(base$Index_CVS_CJO, 3, 0, 0)
model2=sarima(base$Index_CVS_CJO, 3, 0, 1)
model3=sarima(base$Index_CVS_CJO, 3, 0, 2)
model4=sarima(base$Index_CVS_CJO, 2, 0, 0)
model5=sarima(base$Index_CVS_CJO, 2, 0, 1)
model6=sarima(base$Index_CVS_CJO, 2, 0, 2)
model7=sarima(base$Index_CVS_CJO, 1, 0, 0)
model8=sarima(base$Index_CVS_CJO, 1, 0, 1)
model9=sarima(base$Index_CVS_CJO, 1, 0, 2)
model10=sarima(base$Index_CVS_CJO, 0, 0, 1)
model11=sarima(base$Index_CVS_CJO, 0, 0, 2)

# Estimated coefficients

model1$ttable
model2$ttable
model3$ttable
model4$ttable
model5$ttable
model6$ttable
model7$ttable
model8$ttable
model9$ttable
model10$ttable
model11$ttable
```

```r
# Ljung-Box test (non-autocorrelation of residuals)
#firgure out how many lags to do
LjungBox(model1$fit)
LjungBox(model2$fit)
LjungBox(model3$fit)
LjungBox(model4$fit)
LjungBox(model5$fit)
LjungBox(model6$fit)
LjungBox(model7$fit)
LjungBox(model8$fit)
LjungBox(model9$fit)
LjungBox(model10$fit)
LjungBox(model11$fit)



# Tests joints et visualisation des r sidus
checkresiduals(model1$fit)
checkresiduals(model2$fit)
checkresiduals(model3$fit)
checkresiduals(model4$fit)
checkresiduals(model5$fit)
checkresiduals(model6$fit)
checkresiduals(model7$fit)
checkresiduals(model8$fit)
checkresiduals(model9$fit)
checkresiduals(model10$fit)
checkresiduals(model11$fit)

# Selecting the best model based on the information criteria
# AIC
aic <- AIC(model1$fit, model2$fit, model3$fit, model4$fit, model5$fit,
           model6$fit, model7$fit, model8$fit, model9$fit, model10$fit, model11$
               fit)
aic
which.min(aic$AIC)

#BIC
bic <- BIC(model1$fit, model2$fit, model3$fit, model4$fit, model5$fit,
           model6$fit, model7$fit, model8$fit, model9$fit, model10$fit, model11$
               fit)
bic
which.min(bic$BIC)

#Model 8 which is ARIMA(1,0,1) shows the best results for AIC and BIC criterion

#Question 5

# Fit selected ARIMA model
arima_finalmodel <- arima(detrended_ts_data, order = c(1, 0, 1))

final_model <- arima_finalmodel

# To find final model coefficients
summary(final_model)

#PART III
```

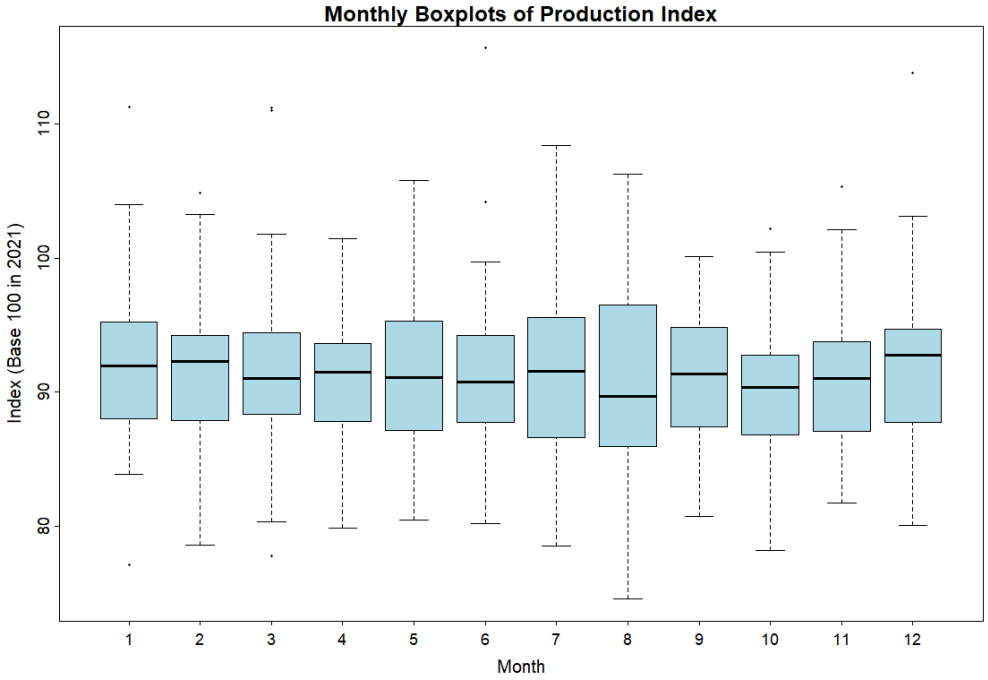
```r
#Question 6 and 7
library(ellipse)

# Step 1: Forecast 2 steps ahead (using detrended series and d=0)
final_model_clean <- Arima(detrended_ts_data, order = c(0, 0, 1))

# Now forecast 2 steps ahead safely
Pred <- predict(final_model_clean, n.ahead = 2, se.fit = TRUE)

# Step 2: Compute variance-covariance matrix of (X_{T+1}, X_{T+2})
sigma2 <- final_model$sigma2
theta1 <- ifelse("ma1" %in% names(final_model$coef), final_model$coef["ma1"], 0)

# Variance of 1-step and 2-step ahead forecasts
sigma_g1 <- sqrt(sigma2)
sigma_g2 <- sqrt(sigma2 * (1 + theta1^2))

# Covariance between X_{T+1} and X_{T+2}
rho <- sigma2 * theta1
Sigma <- matrix(c(sigma_g1^2, rho, rho, sigma_g2^2), nrow = 2)
print("Variance-Covariance Matrix:")
print(Sigma)


# Step 3: Confidence ellipse
ell <- ellipse(Sigma, centre = Pred$pred, level = 0.95, npoints = 1000)

# Step 5: Plotting
plot(ell,xlab = "Forecast for T+1",
     ylab = "Forecast for T+2",
     main = "95% Confidence Ellipse for 2-Step Forecasts",type = "l")
points(x = Pred$pred[1], y = Pred$pred[2], pch = 19, col = "red", cex = 1.5)
```

4.2 Appendix 2

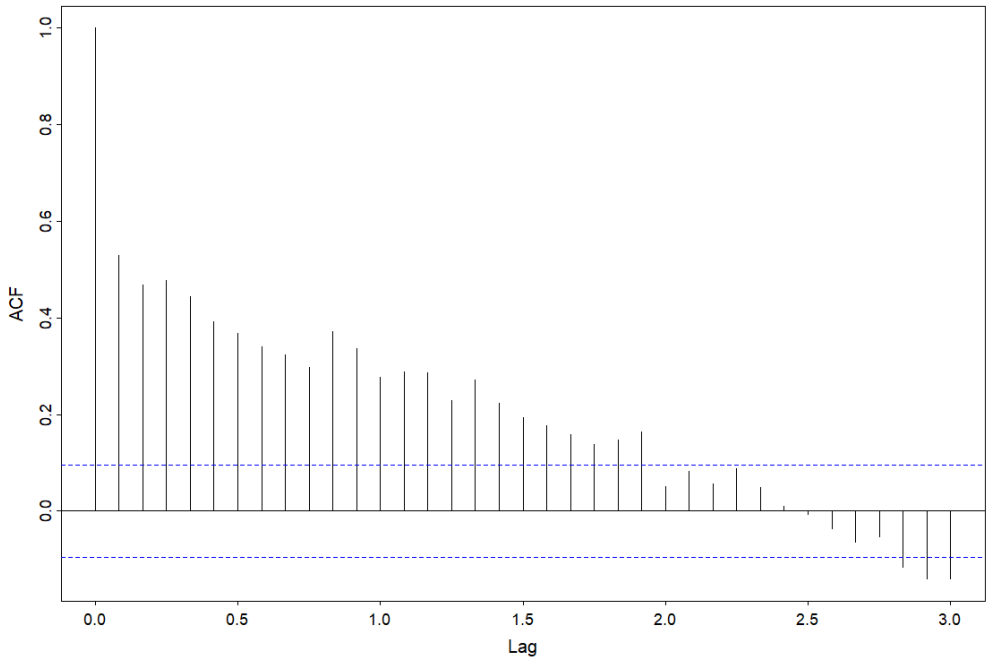
4.2.1 Original Time Series Plot



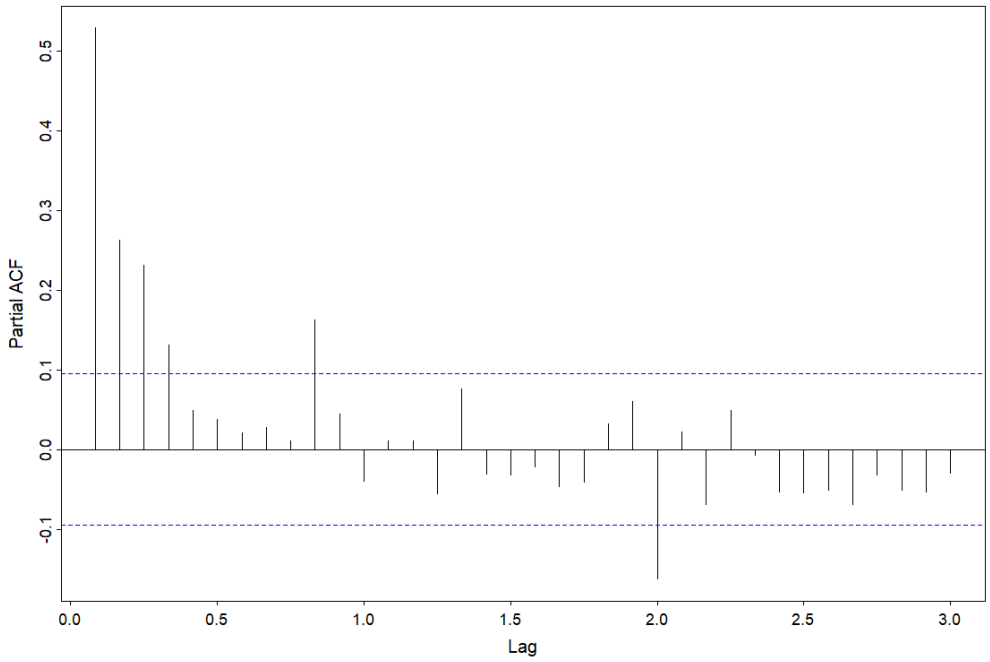
4.2.2 Box Plot



4.2.3 ACF Plot



4.2.4 PACF Plot



4.2.5 AIC and BIC test

Table 1: AIC and BIC Values for Tested ARIMA(p,0,q) Models

Model	df	AIC	BIC
model1 (3,0,0)	5	2498.599	2518.836
model2 (3,0,1)	6	2487.699	2511.983
model3 (3,0,2)	7	2489.497	2517.828
model4 (2,0,0)	4	2523.431	2539.621
model5 (2,0,1)	5	2486.162	2506.398
model6 (2,0,2)	6	2487.674	2511.958
model7 (1,0,0)	3	2558.964	2571.106
model8 (1,0,1)	4	**2484.747**	**2500.936**
model9 (1,0,2)	5	2486.099	2506.335
model10 (0,0,1)	3	2631.586	2643.728
model11 (0,0,2)	4	2597.726	2613.915

4.2.6 Ljung box test results

Table 2: Ljung-Box Test Results from `portes::LjungBox`

lags	statistic	df	p-value
5	2.088778	3	0.5541864
10	11.624792	8	0.1687461
15	15.801076	13	0.2600387
20	19.354358	18	0.3703143
25	29.051688	23	0.1786023
30	32.676528	28	0.2478599