**Time Series – Report**

**À faire :**

1. **Checker les méthodologies des tests (cf. TD 4 et 5 je crois, d’informatique) notamment pour l’interaction de certains tests (ex ADF et portmanteau / Ljung-Bix) et vérifier les les lags utilisés pour calculer les stats des tests.**
2. **Essayer d’utiliser plusieurs modèles d’ARMA ou ARIMA (avec différentes valeurs de p et q) et exploiter AIC ou BIC (inforamtion criteria) (le meilleur modèle, celui qui sera choisi, est celui** **qui minimise l’un des critères).**
3. **Code latex**

Part I:

Question 1:

This series represents the seasonally and working-day adjusted industrial production index for the pharmaceutical industry (NAF rev. 2, division 21) in Metropolitan France, with base 100 in 2021. The index measures monthly production, adjusted to remove seasonal and calendar effects. It’s well-suited for time series modeling, but we must check for stationarity.

Part II:

* **Interpretation of ARIMA Model Results:**

Based on the output from summary(auto\_model) and checkresiduals(auto\_model), here's the breakdown of the results and the steps you can take to interpret and validate the chosen ARIMA model.

* **ARIMA Model Summary:**
* **Model Chosen**: ARIMA(2,0,1)(0,0,2)[12] with non-zero mean

This indicates the following components of the model:

* + **ARIMA(2,0,1)**:
    - **AR(2)**: The model includes two autoregressive (AR) terms, meaning that the current value depends on the previous two observations.
    - **I(0)**: The series is already stationary, so no differencing is needed (d = 0).
    - **MA(1)**: The model includes one moving average (MA) term, which means that the current value is influenced by the previous error term.
  + **Seasonal Component (0,0,2)[12]**:
    - **Seasonal AR(0)** and **Seasonal MA(2)**: The model also accounts for seasonal effects, where it includes two seasonal moving average terms over a 12-period seasonal cycle.
    - **No seasonal differencing** (d = 0 for the seasonal part).
* **Coefficients**: The coefficients for the AR and MA terms are as follows:
  + **AR1**: -0.0636
  + **AR2**: -0.0211
  + **MA1**: -0.6452
  + **SMA1**: -0.1829 (Seasonal MA term)
  + **SMA2**: -0.1119 (Seasonal MA term)
  + **Mean**: 0.0034 (Indicates a slight non-zero mean in the data)

These coefficients suggest the impact of past observations (AR terms) and past errors (MA and seasonal MA terms) on the current value.

* **Model Diagnostics:**

1. **Sigma-squared**: 0.001792  
   This represents the variance of the residuals. A smaller value typically suggests a better fit.
2. **Log-likelihood**: 736.17  
   This value is used to compare models. A higher log-likelihood indicates a better model fit.
3. **AIC/BIC**:
   * **AIC**: -1458.35
   * **BIC**: -1430.05
   * Lower values of AIC and BIC suggest a better-fitting model. The AIC and BIC values here are quite low, indicating that this model fits the data well relative to other potential models.
4. **Training Set Error Measures**:
   * **ME** (Mean Error): 0.0002686 (close to zero, good sign)
   * **RMSE** (Root Mean Squared Error): 0.0420 (a low value indicates good model fit)
   * **MAE** (Mean Absolute Error): 0.0305
   * **MAPE** (Mean Absolute Percentage Error): 143.36% (This value is high, possibly due to high variance in the series, but could also be seasonal effects.)
   * **ACF1**: 0.00098 (close to zero, indicating that there are no significant autocorrelations in the residuals at lag 1)

* **Ljung-Box Test for Residuals:**
* **Ljung-Box test**: Q\* = 34.747, df = 19, p-value = 0.01498
  + The **Ljung-Box test** checks for autocorrelation in the residuals. The null hypothesis of the test is that the residuals are white noise (no autocorrelation).
  + The **p-value** = 0.01498, which is **less than 0.05**, suggesting that **the residuals may still exhibit autocorrelation**, meaning that the model might not fully capture all dependencies in the data.
* **What to Do Next:**

**1. Model Selection:**

* **ARIMA(2,0,1)(0,0,2)[12]** is the model selected by auto.arima(), and while it appears to fit well based on AIC/BIC, there is some residual autocorrelation (as indicated by the Ljung-Box test). This could mean that the model may not have captured all the patterns, particularly seasonal autocorrelation.

**Recommendation**: You might consider re-fitting the model with more seasonal components or further adjusting for potential seasonality or residual autocorrelation.

**2. Model Improvement:**

* **Increase p or q values**: Try testing higher AR or MA terms to capture more dependencies in the residuals, especially after checking ACF/PACF plots again.
* **Seasonal adjustments**: If there’s significant seasonal autocorrelation, you might increase the seasonal AR and MA terms.

**3. Forecasting and Validation:**

* Use this model to make forecasts and validate them with out-of-sample data.
* Perform further diagnostic checks (like residual analysis, out-of-sample validation) to assess the model's validity over time.
* **ARIMA Model Representation:**

The ARIMA model chosen is:

Xt=−0.0636Xt−1−0.0211Xt−2+ϵt−0.6452ϵt−1−0.1829ϵt−12−0.1119ϵt−13+0.0034X\_t = -0.0636 X\_{t-1} - 0.0211 X\_{t-2} + \epsilon\_t - 0.6452 \epsilon\_{t-1} - 0.1829 \epsilon\_{t-12} - 0.1119 \epsilon\_{t-13} + 0.0034

Where:

* XtX\_t is the differenced series,
* ϵt\epsilon\_t is the error term at time tt,
* Seasonal terms are included for 12-period seasonality.

Part III : Prediction

Question 6: Equation for the confidence region

Assuming the residuals are Gaussian (normal distribution), the **(1−α)% prediction interval** for the forecasted values XT+hX\_{T+h}XT+h​ is given by:

X^T+h±z1−α/2⋅SEh\hat{X}\_{T+h} \pm z\_{1 - \alpha/2} \cdot \text{SE}\_hX^T+h​±z1−α/2​⋅SEh​

* X^T+h\hat{X}\_{T+h}X^T+h​: point forecast for step h
* z1−α/2z\_{1 - \alpha/2}z1−α/2​: standard normal quantile (e.g., 1.96 for 95%)
* SE\_h ​: standard error of the forecast at step h

Question 7: Hypotheses for the confidence region

To derive this region, we assume:

1. **Gaussian residuals**: Residuals εt∼N(0,σ2)\varepsilon\_t \sim \mathcal{N}(0, \sigma^2)εt​∼N(0,σ2)
2. **Linearity**: The series is modeled by a linear ARIMA process
3. **Stationarity**: The differenced series is stationary
4. **Correct model specification**: The ARIMA model captures the data dynamics properly

Question 8: Plot of the confidence region

Une image contenant capture d’écran

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, Tracé, ligne

Le contenu généré par l’IA peut être incorrect.

Une image contenant diagramme, ligne, Tracé

Le contenu généré par l’IA peut être incorrect.

Une image contenant diagramme, texte, ligne, Parallèle

Le contenu généré par l’IA peut être incorrect.

Une image contenant capture d’écran, diagramme, texte

Le contenu généré par l’IA peut être incorrect.

Une image contenant capture d’écran

Le contenu généré par l’IA peut être incorrect.

Question 9 :

### Understanding the Problem

We have two time series:

1. \*\*Yₜ\*\*: A stationary time series observed from t = 1 to T.

2. \*\*Xₜ\*\*: Another time series where the value at T+1 (X\_{T+1}) is not immediately available, but Y\_{T+1} is available sooner.

The core question is: \*\*Under what conditions can the early availability of Y\_{T+1} improve the prediction of X\_{T+1}?\*\* Additionally, how can we test these conditions?

### Breaking Down the Components

1. \*\*Stationarity of Yₜ\*\*: Since Yₜ is stationary, its statistical properties (mean, variance, autocorrelation) are constant over time. This is a useful property for modeling and prediction.

2. \*\*Predicting X\_{T+1}\*\*: Typically, to predict X\_{T+1}, we might use its own past values (autoregressive models), past values of other series (like Yₜ, in a vector autoregressive framework), or a combination.

3. \*\*Early Availability of Y\_{T+1}\*\*: If Y\_{T+1} is available before X\_{T+1}, and if Yₜ has a predictive relationship with Xₜ, then incorporating Y\_{T+1} could improve the forecast of X\_{T+1}.

### Conditions for Improvement

For Y\_{T+1} to be useful in predicting X\_{T+1}, the following conditions must hold:

1. \*\*Granger Causality\*\*: Yₜ must Granger-cause Xₜ. This means that past values of Yₜ contain information that helps predict Xₜ beyond what is contained in past values of Xₜ alone. If Y\_{T+1} is correlated with X\_{T+1} conditional on past values of Xₜ and Yₜ, then its early availability can improve the forecast.

2. \*\*Contemporaneous Relationship\*\*: There might be a contemporaneous relationship between Xₜ and Yₜ, meaning that Y\_{T+1} and X\_{T+1} are correlated at the same time point. If we can model this relationship (e.g., Xₜ = αYₜ + error), then knowing Y\_{T+1} gives a direct estimate of X\_{T+1}.

3. \*\*Lead-Lag Relationship\*\*: Yₜ might lead Xₜ, meaning changes in Yₜ precede changes in Xₜ. If Y\_{T+1} leads X\_{T+1}, then it can be a useful predictor.

4. \*\*No Instantaneous Feedback\*\*: If Xₜ also affects Yₜ instantaneously, then the relationship might be more complex, and the improvement might not be straightforward unless the feedback is modeled.

### Testing the Conditions

To test whether Y\_{T+1} can improve the prediction of X\_{T+1}, we can perform the following statistical tests and analyses:

1. \*\*Granger Causality Test\*\*:

- Estimate two models:

- Model 1: Xₜ = f(past Xₜ)

- Model 2: Xₜ = f(past Xₜ, past Yₜ)

- Perform an F-test to see if the inclusion of past Yₜ significantly improves the prediction of Xₜ. If it does, Yₜ Granger-causes Xₜ, suggesting that Y\_{T+1} could be useful.

2. \*\*Cross-Correlation Analysis\*\*:

- Compute the cross-correlation function (CCF) between Xₜ and Yₜ at different lags.

- If the correlation is significant at lag -1 (i.e., Y at t+1 is correlated with X at t), this suggests that Y\_{T+1} might be useful in predicting X\_{T+1}.

3. \*\*Contemporaneous Correlation Test\*\*:

- Check the correlation between Xₜ and Yₜ at the same time points. If high, a model like Xₜ = αYₜ + εₜ could be fit, and Y\_{T+1} would directly predict X\_{T+1}.

4. \*\*Out-of-Sample Forecasting Evaluation\*\*:

- Split the data into training and test sets.

- Train two models:

- One using only past Xₜ to predict X\_{T+1}.

- Another using past Xₜ and Y\_{T+1} (where available) to predict X\_{T+1}.

- Compare the forecast accuracy (e.g., using MSE, MAE) to see if the inclusion of Y\_{T+1} improves predictions.

5. \*\*Vector Autoregression (VAR) Models\*\*:

- Fit a VAR model including both Xₜ and Yₜ.

- Check if the coefficient on Y\_{T+1} in the Xₜ equation is statistically significant.

### Potential Models to Utilize

1. \*\*Autoregressive (AR) Model with External Input\*\*:

- Xₜ = α₁Xₜ₋₁ + ... + αₚXₜ₋ₚ + βY\_{T+1} + εₜ

- Here, Y\_{T+1} is included as an exogenous variable.

2. \*\*Dynamic Regression Models\*\*:

- Models that allow for both autoregressive terms and exogenous variables can directly incorporate Y\_{T+1}.

3. \*\*State-Space Models\*\*:

- These can model more complex dependencies and incorporate the latest information flexibly.

### Practical Example

Suppose:

- Xₜ is the daily sales of a product.

- Yₜ is the daily temperature, which is recorded earlier in the day than sales.

If higher temperatures lead to higher sales (e.g., ice cream), then knowing today's temperature early can help predict today's sales before sales data is available.

### Testing:

1. \*\*Granger Causality\*\*: Test if past temperatures help predict sales beyond past sales.

2. \*\*Cross-Correlation\*\*: Check if today's temperature is correlated with today's sales.

3. \*\*Model Comparison\*\*: Compare a sales AR model with and without temperature.

### Conclusion

\*\*Conditions for Improvement:\*\*

- Yₜ must have a predictive relationship with Xₜ (Granger causality, contemporaneous correlation, or lead-lag).

- The relationship must be such that Y\_{T+1} provides additional information not contained in the past values of Xₜ alone.

\*\*Testing Methods:\*\*

1. Granger causality tests.

2. Cross-correlation analysis.

3. Contemporaneous correlation tests.

4. Out-of-sample forecast comparisons.

5. VAR or other multivariate time series models.

If these tests indicate a significant relationship, then incorporating Y\_{T+1} can improve the prediction of X\_{T+1}.