Countries Consumer Price Indexes (CPIs) Analysis and Forecasting

Filipa Capela
University of Coimbra
Faculty of Science and Technology
Coimbra, Portugal
filipacapela@outlook.pt

Tiago Conceição
University of Coimbra
Faculty of Science and Technology
Coimbra, Portugal
ecaldeira@dei.uc.pt

***Abstract*—The Consumer Price Index (CPI) is a time series that may vary according to time and other vital factors. With this paper, we are going to predict the way that the data is going to flow. The approach we will use falls back on some models lectured in Analysis of Temporal Series. However, these models do not offer a reliable future prediction of the CPI, and it is possible to understand how the data will trend.**

***Index Terms*—Consumer Price Index, Trend, Seasonality, Stationary, Erratic Component, Time Series**

I. Introduction

The Countries Consumer Price Indexes (CPI) is an important data source that retrieves information on how inflation tends accordingly over the years. Unfortunately, as we all know, the CPI is increasing without any signal of stopping soon, meaning the cost of living will rise. The analysis and forecasting of this time series can help in some decisions like:

- Will this affect the cost of living?
- How does the inflation will trend?

The decision question analysis mentioned before can help an average person with a lack of knowledge understand the tendency of the CPI and how that will affect society. The official Countries Consumer Price Index was first created in 1919 and aimed to measure the average change in the prices paid by consumers for goods and services. Data like this is always hard to predict the trend; with the help of some methods, we can fetch some conclusive information, although it needs to be more confident in accuracy. This paper will follow up with the given topics: in the section II, the main topic discussed is the dataset's source and how the information about the data is structured. The subsection III is going to reference the first analysis of the time series in order to understand some essential tendencies and behaviors, in the section methods, there is a brief theoretical review of how the methods are applied for work and what is being applied in the section results. Next, the final subsection V is an analysis of our dataset comparing all the results, applying the basic descriptive techniques and conclusions obtained.

II. Data

Firstly, making this paper required a dataset with all the countries' consumer price indexes so that exploration could be made. It is possible to do this by fetching the data from a dataset called "CPITimeSeries.csv." In order to make an analysis and forecasting about CPI, was first needed the selection of two countries, with the intention of making an in-depth study about both performances and comparing them. Then the selected countries were Switzerland and Luxembourg. Those options had some elements in common, but the main character was the richness of both regions, which was the breaker for that decision. Making the first approach to the dataset, the selected indicator for further analysis was "Consumer Price Index, All items," with the indicator code "PCPI-IX" after those columns, there is a vast amount of columns that indicate the year and the correspondent month as well as the value matching the CPI. With the retrieved data, it is possible to see the value change across the years.

III. Methods

A. Time Series Analysis

Regarding time series analysis, it implies that the data contains a particular starting timestamp and an ending one. Between those periods, it is possible to get the CPI value accordingly to the time. So then, knowing that the first approach was collecting the data, from 1955 to 2022, Switzerland had several CPI numerical values.

As seen in figure 1, the CPI of Switzerland in the early years was below 30, and to the current year, 2022, the number is above 100 and not showing signs of slowing down. With time series, we can see that the tendency is to increase, and with further analysis with other methods.

B. Decomposition Models

Times Series can be decomposed into three components: trend, seasonality, and erratic/irregular. Our goal is to obtain the erratic component to apply the forecasting models.

Therefore, to decompose the times series, there are three models: The additive Model, the Multiplicative Model, and the pseudo-additive Model.

For the additive model (1), we consider that the trend, seasonality, and erratic component are independent, that is, the times series is the sum of the three components.

For the multiplicative model (2), we consider the times series to be the product between the trend, the seasonality, and the erratic component.

$$x(n) = tr(n) + sn(n) + e(n) \tag{1}$$

$$x(n) = tr(n)sn(n)e(n) \tag{2}$$

To analyze the results we apply the models described above to the polynomial functions

C. Trend

One of the most important methods when analyzing a time series problem is to estimate the trend. A trend can give an increasing slope or a decreasing slope or stay constant, also known as stationary.
This approach allows us to understand the pattern of the inflation movement across the time given by the data. In this project were applied three types of polynomial functions and those are:

- Linear Fitting;
- Quadratic Fitting;
- Cubic Fitting;

These polynomial functions can vary the equation according to the degree applied, so the equation is given by:

$$t(n) = a0+a1(nTs)+a2(nTs)^2+a3(nTs)^3+...+ap(nTs)^P \tag{3}$$

Another way to estimate the trend is by applying some smoothing methods like the moving average, also known as MA. Smoothing is a technique that is applied to the time series in order to remove the fine-grained variations. This allows the data to be cleaner and without noise.
The moving average has the following equation:

$$t(n) = \sum_{k=n-\frac{M+1}{2}}^{n+\frac{M+1}{2}} w_k x(k) \tag{4}$$

Looking at the formula, the result obtained through t(n) gets the estimated trend, where M is the span of the filter and the number of samples used to calculate the mean. For the weights, we consider that: [

$$\sum_{j=0}^{M} w_j = 1$$

, and thus implement Equally Weighted Moving Average (EWMA). For this method, we varied M between {5,13,25,51}. We used odd numbers for the filter to be centered on the samples.

Finally, we will implement LOcally WEIghted (LOWESS/LOESS), which is a smoothing method of non-parametric fitting the regression model one window. In the analysis part, to have a comparison term with the EWMA, we will also consider M between [5,13,25,51].

D. Seasonality

Seasonality occurs when a time series is a repetition in a certain time interval, and between that time, some events take place. In order to make a good analysis of seasonality, there are some methods that allow that like:

- Seasonality Assessment by filtering - Fourier series;
- Removal by differencing;
- Auto correlation Sequence;
- Confidence Bounds.

By filtering with Fourier transform, the seasonality assessment assumes that with a given time series like the one seen in the figure 1, this can be represented by a complex sum of sin and cosin functions. A supplemental method is a removal by differencing. This technique is not explicit like the ones seen before. It is considered an implicit way to do it. We can obtain the seasonality by removing the differencing with the following calculation:

$$\Delta x(n) = x(n) - x(n-1) = y(n) \tag{5}$$

If the result of the function 5 given by y(n) is without trend, and if there is no seasonality present, it can be considered stationary. With the equation obtained before, if a trend is still visible, it is required to apply a high-order differencing. Usually, first-order or second-order differences.

$$\Delta^2 x(n) = \Delta x(n)-\Delta x(n-1) = x(n)-2x(n-1)+x(n-2) = z(no) \tag{6}$$

Another method is the sample autocorrelation sequence is used to find repeating patterns. It provides the correlation between the time series and a delayed version of itself over some intervals. The formula is presented by :

$$r_{xy} = \frac{\sum_{n=0}^{N-T-1} (x(n) - x)(x(n+T) - x}{\sum_{n=0}^{N-1} (x(n) - x)^2} \tag{7}$$

If the result of the equation 7 is positive, that will mean that the time series changes in the same way as x(n) and x(n+T). Otherwise, if the result of the equation 7 is negative, that results in a change of the time series in a different direction in the same variables. One more outcome can be expected, and that is if the result of the equation 7 is zero, that will end up with a non-correlated time series.

E. Stationary Assessment

In this subsection, to make a stationary analysis and understand if the time series is stationary, the method Dickey-Fuller unit root test is used. This procedure is a statistical test that assesses the existence of the unit root. The test null and alternative hypotheses are: H0 with the unit is present in a time series sample (non-stationary). H1 with the unit root is not present in a time series sample (stationary). The model

assumed by the dickey-fuller test is an autoregressive (AR) model of order 1. If we fail to reject the null hypothesis, the time series is non-stationary.

F. SARIMA

SARIMA stands for a seasonal autoregressive integrated moving average. This model is used if the time series is non-stationary. In 1970, the ARIMA was generalized to deal with seasonality, and the SARIMA model was created. The model is formed by two ARIMA models, one for short-term dependencies and the other for seasonality. The formula that makes it is:

$$\phi_p(B)\Phi_P(B^S)\nabla^d\nabla_S^D X(n) = \theta_q(B)\Theta_Q(B^S)Z(n) \quad (8)$$

As possible, see in the 8, the values that B takes refer to the backward shift operator. Also, the function (p,d,q) refers to the short term and (P,D,Q) to the seasonal.

G. Exponential Smoothing

Exponential Smoothing, better known by the abbreviation ES. This method is a proper extrapolation when the time series is formed by samples that summarize periodic data, like a yearly period. There are three distinct types of exponential smoothing:

- Simple Exponential Smoothing;
- Double Exponential Smoothing;
- Triple Exponential Smoothing;

Each one of the different types is applied to different types of time series, for example, if the TS does not have trend and seasonality, the most adequate is the simple exponential smoothing. On the other hand, if the time series is a trend but no seasonality, the most appropriate is the DES (Double Exponential Smoothing). Also, if a trend and seasonality are used, the TES (Triple Exponential Smoothing) is used.

As seen, all these different types work for different purposes in the context of this project is used the TES. There are three parameters to be defined and three equations, one for model level and the other for trend.

The equation is given by:

$$x(N+h) = (L(N)+hT(N))I(N-S+mod(h-1,S)) \quad (9)$$

IV. Multivariate Classical

In most cases, in the real world, the analysis of a single time series needs to be supplemented with other sources of data, in particular other time series. When several TS co-exist, one changes from the univariate TS analysis and forecasting domain to multivariate (multi-ts) analysis (MVA).

This chapter has several steps, with the cross-correlation being the first. Cross-correlation is an extension of auto-correlation.

The second step is applying the SARIMAX (Seasonal Auto-regressive integrated moving average model with exogenous inputs). ARMAX (Auto-regressive moving average model with exogenous inputs) with integrated versions ARMA is known as ARIMAX(Auto-Regressive Integrated Moving Average model with exogenous inputs) and SARIMAX.

In the context of the project, SARIMAX is applied, and the equation is:

$$\phi_p(B)\Phi_P(B^S)\nabla^d\nabla_S^D Y(n) = \sum_{i=1}^{M}\xi_i X_i + \theta_q(B)\Theta_Q(B^S)Z(n) \quad (10)$$

The next step goes through the multiple equation models, and an analysis is required. Once the single equation models only assume that the exogenous/predictor time series are not influenced by the single target TS. This presents a limitation because there are several reasons one time series can influence others, and multi-equations solve this. In the multiple equation models, the one applied is - VARMAX which considers the exogenous inputs.

A. Multivariate Machine Learning

Multivariate Machine Learning is a data-driven approach that is considered a non-classical approach. Apply a time series analysis successfully, and it is required preparation, following a pipeline. The first step is the data preparation part, and this project's next steps are.

1) Preparation of the timestamps;
2) Removing the nan values;
3) Split of data;
4) Standardization of the data;
5) Applied Lag Data;

After following those steps, it was time to apply a feed-forward model, generate data, train the MLP, evaluate the residuals, take conclusions regarding the forecast some steps ahead, and analyze the performance.

V. Results and Discussion

1) Decomposition Models: When we consider additive models, since the components of TS are independent, we expect the difference between TS and the polynomial fit to be close to 0. However, when we look at figure 2, this is not true since there are values that are discrepant from zero, especially when considering the dates from 2013. When considering multiplicative models, we expect the division and the ts with the polynomial fit to be around 1. When we look at the figure 3, we see that the ts without cubic trend is around, thus having a better behavior.

2) Trend: For the polynomial functions, the results obtained are shown in figure 4. By observing the figure, we can conclude that the cubic polynomial is better fitting than the others. This can be corroborated with the table I, since MSE is smaller.

For the Equally Weighted Moving Average (EWMA) model, we use M= {5,13,25,51}. The results are presented in figure 4. To obtain a more detailed analysis, we chose the window

Polynomial Fitting	MSE
Linear Fitting	44,47
Quadratic Fitting	23,98
Cubic Fitting	4,43

TABLE I: Mean Square Error in polynomial fitting

	TS Original	TS without trend	TS erratic component
ADF statistic	-0.933	-3.559	-10.109
p-value	0.776	0.006	0.000
Critical Values (1%)	-3.439	-3.439	-3.439
Critical Values (5%)	-2.865	-2.865	-2.865
Critical Values (10%)	-2.569	-2.569	-2.569

TABLE II: ADF test

2000 until 2022 (figure 6), since it is in this period that more seasonality patterns are observed. With the analysis of the figure 6, we see that the values of M = 13 and M = 25 are the most appropriate since the trend can be seen clearly, without seasonality patterns being observed.

For the LOcally WEIghted (LOWESS/LOESS), we used the same parameters in the EWMA model to have a reliable comparison between the two models. In figure 7, we can see in more detail the application of the model between the years 2000 to 2022. Like in the EWMA model, the M = 13 and M = 25 values allow a better trend observation without interfering with the seasonality.

3) Seasonality: As stated in the section III-D, we use filtering to remove seasonality patterns. For that, it is necessary to apply the Fourier discrete transform to find out which is the best model (EWMA or Lowee and the respective M) for trend extraction in order to be able to remove the seasonality more accurately. As said in the previous topic V-2, for each model, the best values of M = 13 and M = 25. Therefore, with the EWMA AND LOWESS spectra analysis with M=13 and M=25, we conclude that the spectrum with the best results in trend extraction is when the EWMA model is used and when M=13 (figure 8) since it is smoothed. We can see the cycles: there is a seasonal pattern that repeats every 12 months and one that repeats every six months. From now on, for the removal of the seasonality, we will use the EWMA model with M=13 for trend removal.

When filtering, we get the figures 9 and 10. When we apply to difference, we get the results in figures 11 and 12.

For the Sample Autocorrelation Sequence analysis, we obtained the figures 13, 14 and 15. Figure 13 shows the autocorrelation of the original TS, where we can see the trend (the points are decreasing so that you can see the trend). Figure 14 shows the autocorrelation of the TS without the trend (which was removed by differentiation). However, the seasonality is still present, i.e., repeated patterns are observed every 12 and 6 months (visible towards the end of the TS). Figure 15 shows the autocorrelation of the erratic component, which is stationary, given that the trend and seasonality were removed. We observe in the figure that when T=1, there is a significant correlation between the current and the previous sample.

4) Stationality: We apply the ADF test on the original TS, TS without the trend (differencing), and TS with the erratic component (differencing). The results are shown in TABLE II. Analyzing the table, we conclude that the TS without the trend and the erratic component are stationary since the p-values are less than 0.05 or 0.1. However, the smaller the values of the statistical test, the better the results. We conclude that the erratic component of the TS has a more stationary behavior than the TS without the trend.

5) SARIMA: When we analyze the figure 16, we see that the TS of Switzerland is nonstationary, as the autocorrelation shows a clear trend. Therefore, the next step is to remove the trend. In figure 17, we see that there is no longer a trend (no need for further differencing). However, we can observe some seasonal patterns. To remove the seasonality, we used the S = 12, as seen in figure 18.

For the application of SARIMA, we need the values of p, q, P, Q. With the analysis of the figure 18, we observe that ACS starts to have non-significant values when q>1 and that p (analyzing PACS) starts to have non-significant values when p>1. To obtain the values of P and Q, we need to observe ACS and PACS at lags [0,12,24,36,48,60,72,84,96]. Looking at the figure 19, we see that Q starts to have non-significant values when Q>1 and P starts to have non-significant values when P>7.However, we applied a grid search to obtain the best hyperparameters to have a more reliable result. Therefore, the best values obtained were: p=2, q=1, P=3, and Q=1, with AIC=-289.35.

6) TES: To apply the TES with the multiplicative model, we separated the TS into two sets: the training set from 1955 to 2019 and the validation set from 2020-01 to 2022-07. To obtain the best hyperparameters of the TES, we applied grid search using a multiplicative and additive model. The hyperparameters =[alpha, gamma, delta] obtained for the multiplicative model were = [0.5, 0.8, 0.1] and for the additive model, were: [0.6, 0.8, 0.1].

The model that obtained the lowest error when performing the prediction was the additive model with RMSE=0.97, although there is not much discrepancy between this RMSE and the RMSE of the multiplicative model with 1.02 error. In figure 20, we observe the TS broken down into training and validation set that is used for forecasting and the comparison between predicted and actual values when applying the TES with an additive model.

7) Cross-correlation: From now on, we will use the Luxembourg TS in the time interval from 2000 to 2022 and the TS of Switzerland from 2000 to 2022. In figures 21 and 22, we observe the original TS, the TS without the trend, and the TS without the seasonality, the erratic component.

When we applied cross-correlation to the two original TS, we found that the results greatly exceeded the confidence interval. Therefore, for cross-correlation, we need to put the TS on the same scale using standardization. In the case of autocorrelation, we also applied differencing since when applying autocorrelation in the original data, the results were way

out of the confidence interval, as seen in switzerland's TS. In figure 23, we present the results obtained from autocorrelation and cross-validation from the two TS. When we analyze the scatter plot and cross correlatiom that are present in figures 23 and 24, we see that Luxembourg and Switzerland have some correlation, but not much.

8) Simple Equation Models: In this subsection, with implement SARIMAX model. To do this, we first obtained the best hyperparameters using a grid search. The best results obtained were: p=1, q=1, P=0, Q=2.

We separated the dataset into two sets: a training set from 2000 to 2020 and a test set from 2021 to 2022. We trained the model with the test set and obtained the autocorrelation and distribution as in figure 25. With the analysis of this figure, we denote that results are within the confidence interval with larger lags, so there are good results. We see that the model follows a mostly Gaussian distribution. However, it has exceptions, as the ACS has some results that exceed the Gaussian distribution. In figure 26, we observe the model's prediction when we tested on the test set, obtaining an RMSE of 0.086, consequently being a good prediction.

9) Multiplicative Equation Model: In this section of the multiplicative equation model, we applied VARMAX. After applying this model, we obtained the figure 27 and concluded that the model captures some dynamic processes, but there are exceptions. We also obtained the following values:

$$\phi = \begin{bmatrix} 0.03 & 0.10 \\ 0.05 & 0.05 \end{bmatrix}$$

With the analysis of this matrix, we verify that there is a weak relation between Switzerland and Luxembourg and a better relation between Luxembourg and Switzerland compared to the previous one. In figure 28, we present the forecast of the test set. As concluded earlier, since there is not much correlation between the countries, the forecast did not have good results. Herefore, the model chosen for the forecast could have been more appropriate.

A. Multivariate Machine Learning

Finally and after analyzing the simple equation models, it is time to apply the machine learning multivariate models. In order to start this approach, we decided to split it into three different time series, one for training data, one for validation, and the other for testing data. Most of the requirements necessary for applying multivariate machine learning were already made. So that left us using the already applied sarima to define the MLP. For the delays applied, the decided values were [1,2,3,12,13,14,15] and then used for training and testing. Once the variables with the delays were defined, it was time to apply the MLP, with the maximum iterations set to 5000 and the hidden layer sizes defined for 50,500. As a result, we obtained the capability to evaluate the residuals once again, but this time for the MLP solution.

Next, an evaluation of fitting quality by inspecting residuals was made, and it is possible to see that the bar chart follows a gaussian distribution. However, there are some bars outside of the bound defined like in the figure 30.

After the evaluation of the residuals, a forecast on the testing set with one step ahead was made, obtaining a relatively high RMSE with value 1.10, but that might be because of how small the testing set is, as can be seen in the picture 31.

Although the RMSE for one step was high, for the several steps ahead, it was high getting values close to 1.22 too because of the data of the figure 32.

We were not satisfied with the significant error given by the RMSE, so we decided to increase the value of the test set, and indeed the values of error decreased for the single step taken, and for several steps matching RMSE of 0,79. Like it is possible to see in the figure 33 and 34.

One of the required pre-processing data is either the normalization or the standardization for the data to be coherent. So, we applied the to both Luxembourg and Switzerland the normalization to be in the same interval of values. By applying the normalization to each set interval, we can see that they follow the same trend in the train, but the validation and test are quite different like in the picture 35. Moreover, it is possible to see in the last two pictures (36 and 37) that both countries have the same behavior and can help determine each other for a better prediction. After the analysis of this data, to obtain an evaluation of the validation data and the forecast, it was required to define the delay combination values, and the best one achieved from the variety decided, the one selected was [0,1,12],[1,2,12].

VI. Conclusion

In conclusion, making predictions can be real hard, and all the themes that surround prediction methods, SARIMA, and the algorithms of machine learning can be accurate, but still far from reality. There are way too many factors that impact the real trend of a country's CPI, on one day can be up because of inflation and the next can be down. We understand now how the times series works and how it behaves over time.And with this analysis it's possible to gather some clues how the time series will trend, although it's impossible to predict the real behavior.

References

[1] Time Series Analysis and Forecasting by César A. D. Teixeira

Appendix

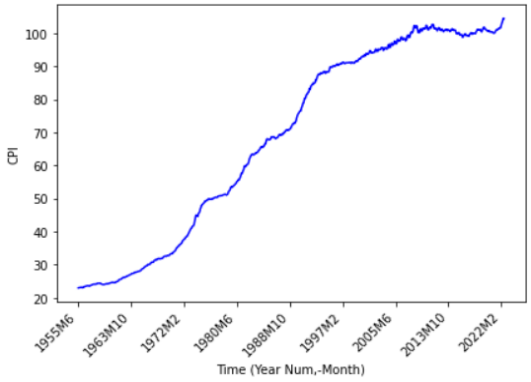

Fig. 1: Switzerland CPI since 1955

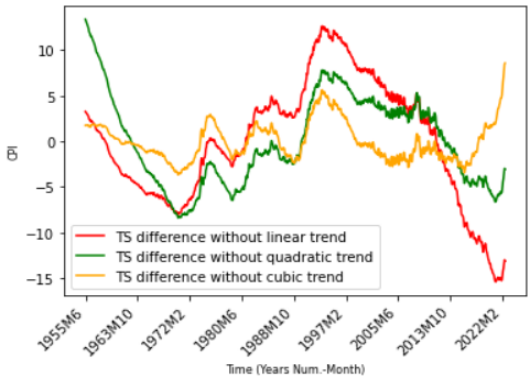

Fig. 2: Polynomial functions with additive Model

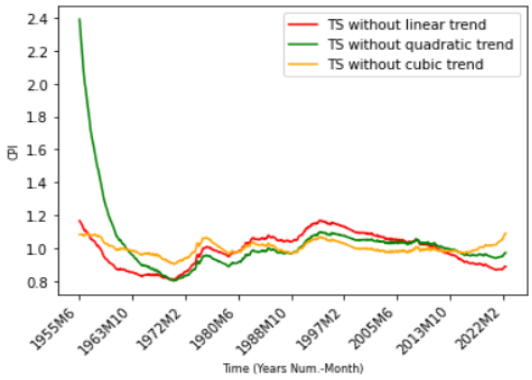

Fig. 3: Polynomial functions with Multiplicative model

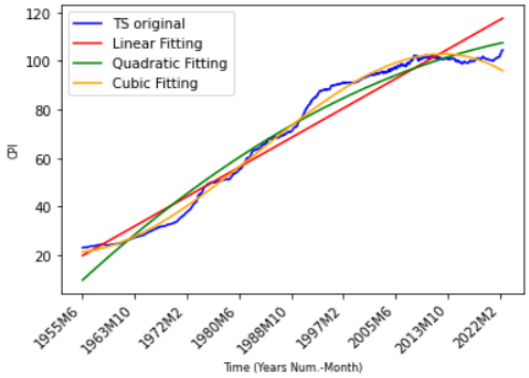

Fig. 4: Fitting Polynomial functions

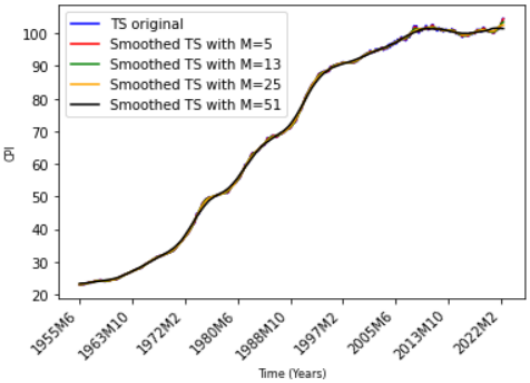

Fig. 5: EWMA

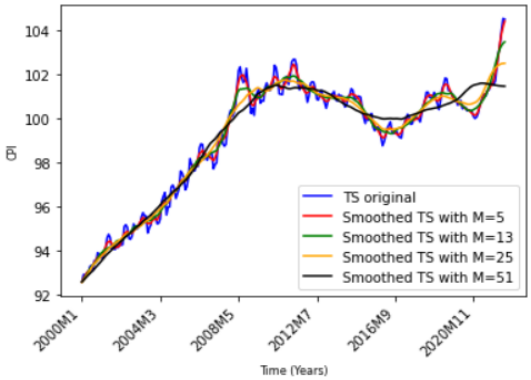

Fig. 6: EWMA more detail

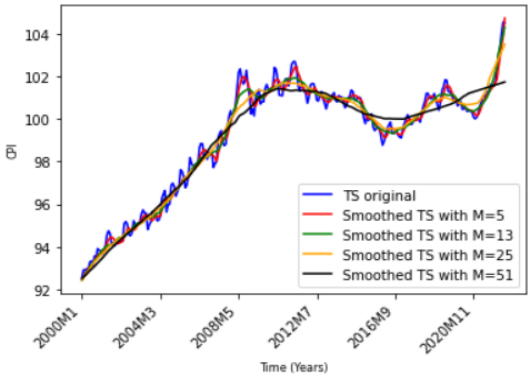

Fig. 7: LOWESS/LOESS more detaild

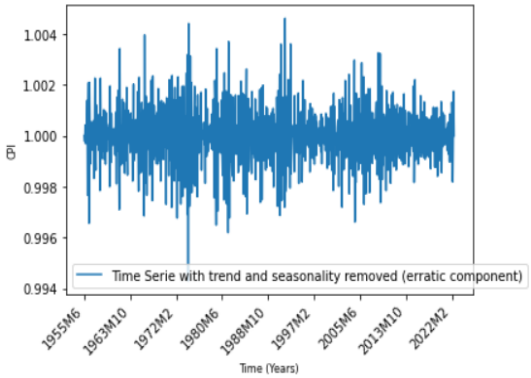

Fig. 10: Erratic Component with Filtering

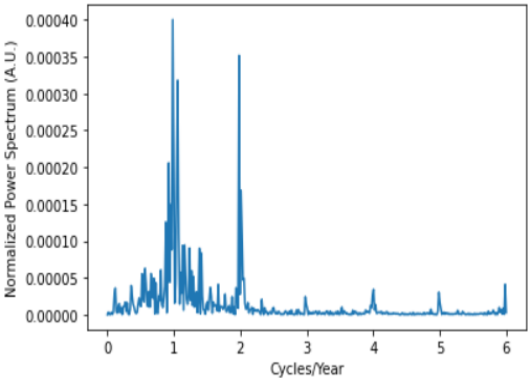

Fig. 8: EWMA M=13

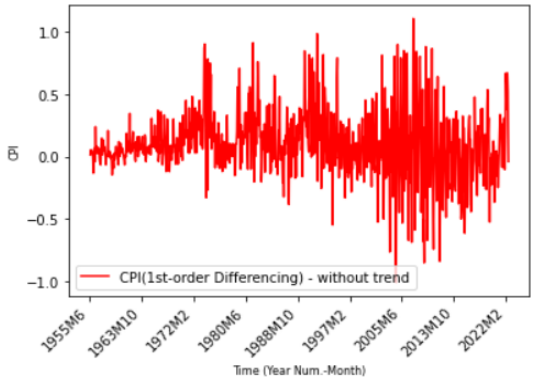

Fig. 11: Differecing 1st order

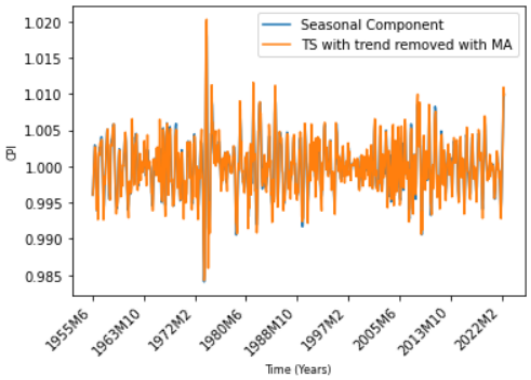

Fig. 9: Components Filtered

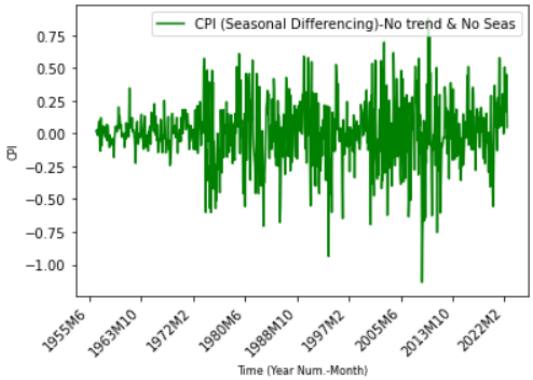

Fig. 12: Erratic Component with Differecing

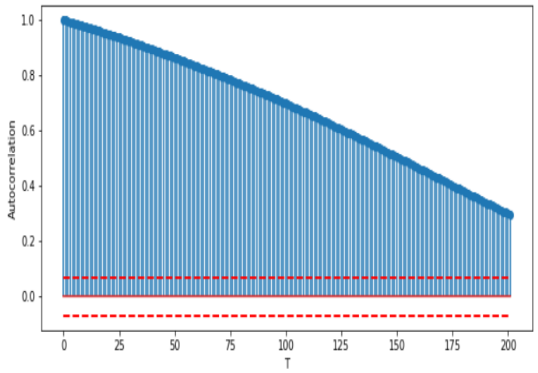

Fig. 13: TS Original Autocorrelation

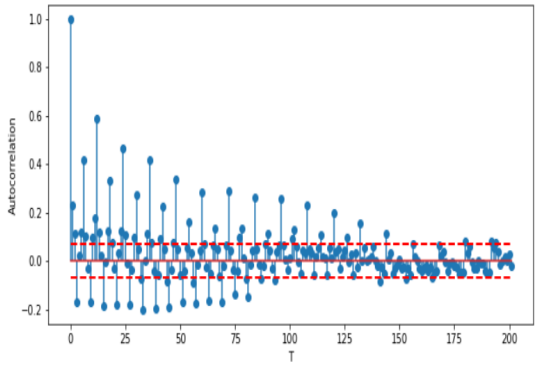

Fig. 14: Differecing 1st order Autocorrelation

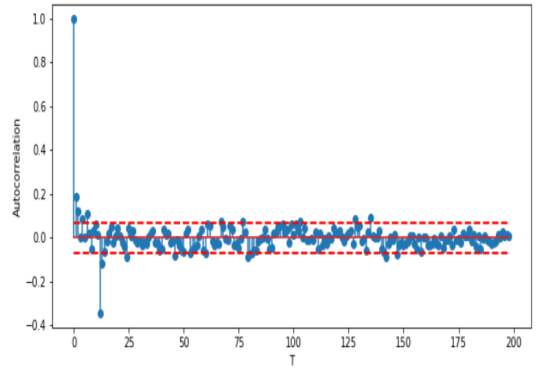

Fig. 15: Erratic Component Autocorrelation

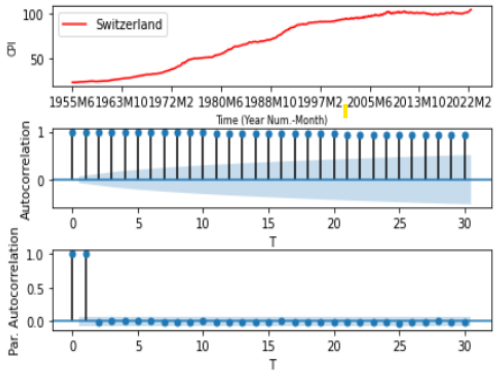

Fig. 16: Autocorrelation and Parcial Autocorrelation in Switzerland TS

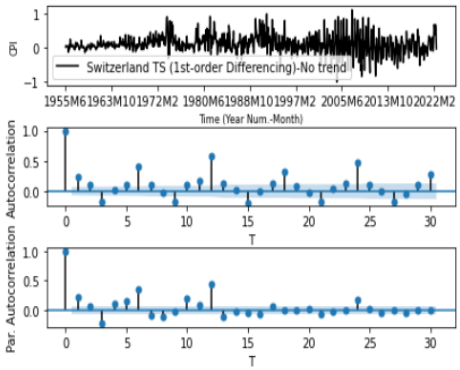

Fig. 17: Autocorrelation and Parcial Autocorrelation in Switzerland TS with trend removal

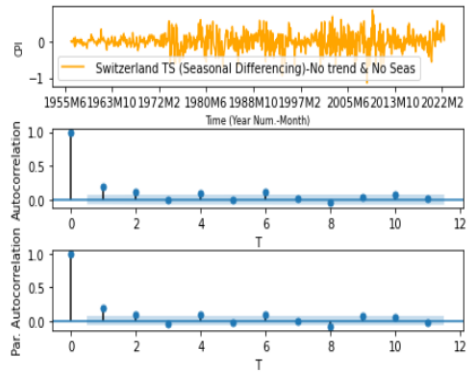

Fig. 18: Autocorrelation and Parcial Autocorrelation in Switzerland TS with seasonal differencing

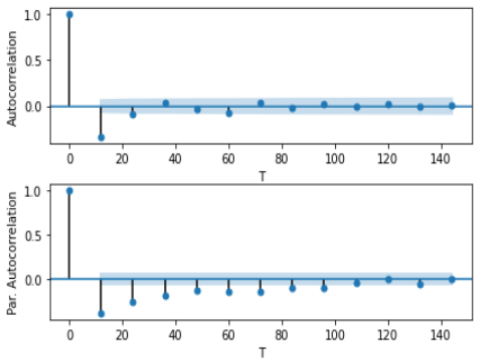

Fig. 19: Autocorrelation and Parcial Correlation in in Switzerland TS with seasonal differencing with different lags

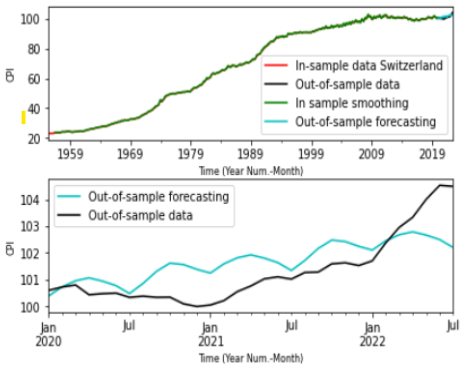

Fig. 20: TES forecasting with additive model

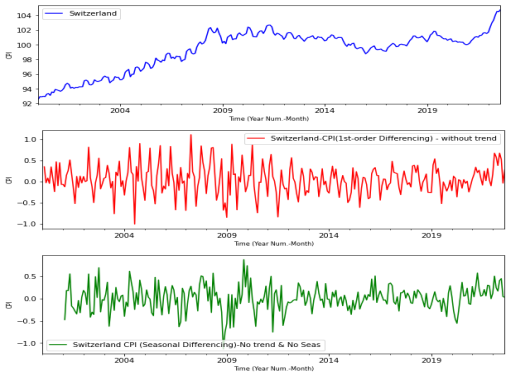

Fig. 21: Switzerland with differences

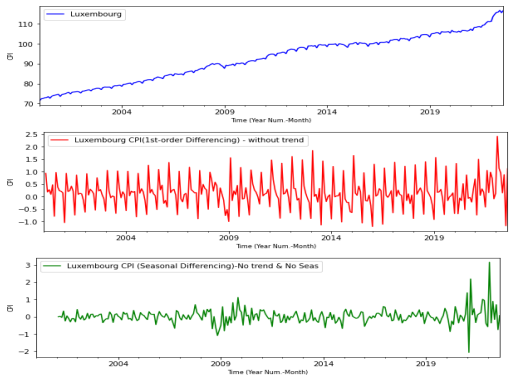

Fig. 22: Luxembourg with differences

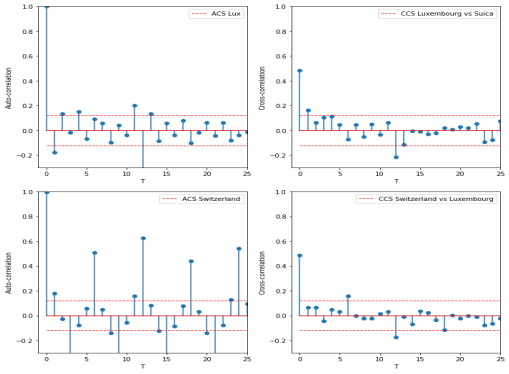

Fig. 23: Autocorrelation and Cross Correlation

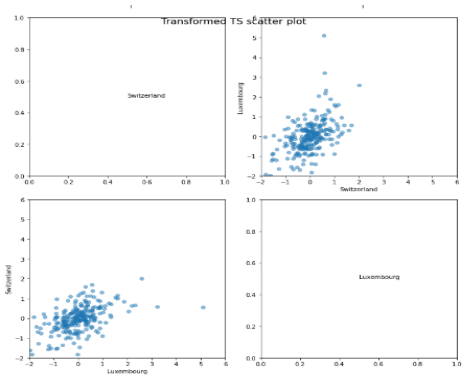

Fig. 24: Scatter Plot

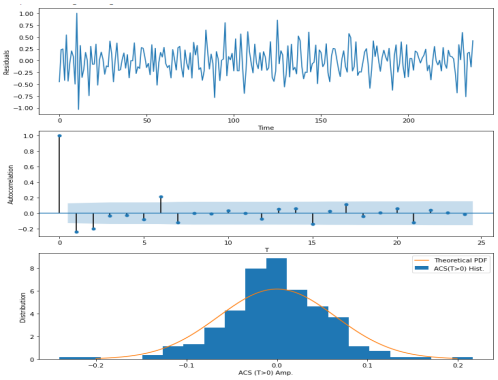

Fig. 25: SARIMAX model

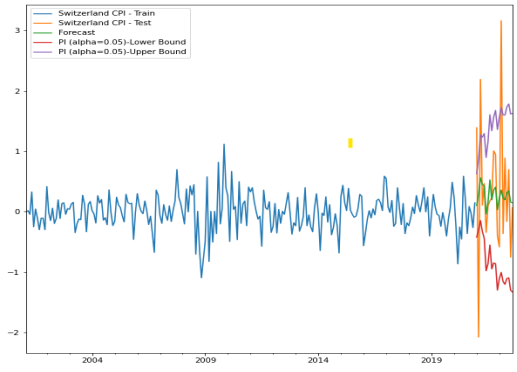

Fig. 26: Forecasting with SARIMAX model

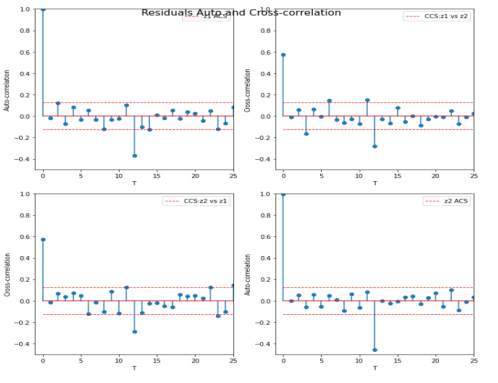

Fig. 27: Residuals Auto and Crosscorrelation with VARMAX

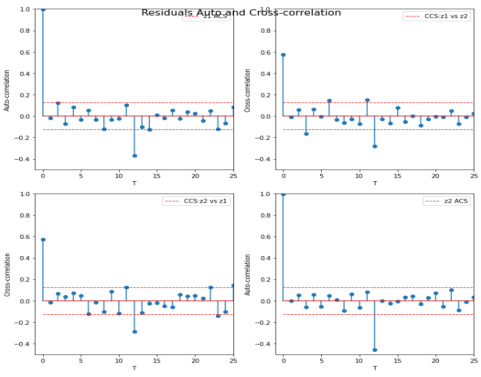

Fig. 28: VARMAX Forecast

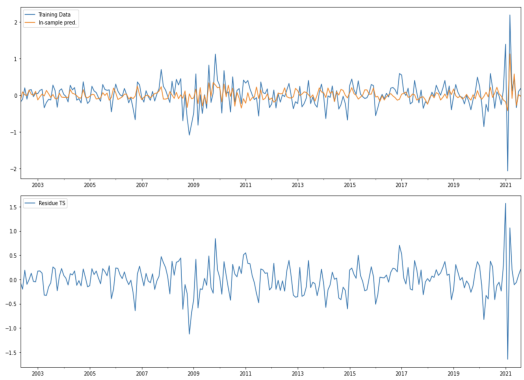

Fig. 29: Residual MLP

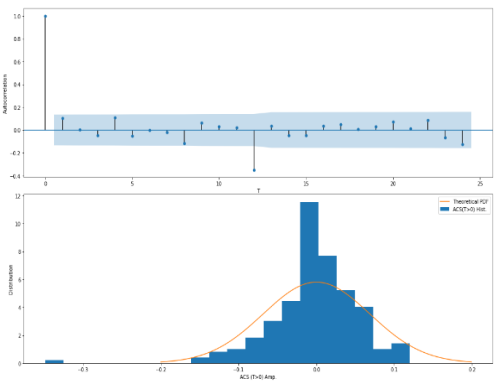

Fig. 30: Residual$_2$ MLP

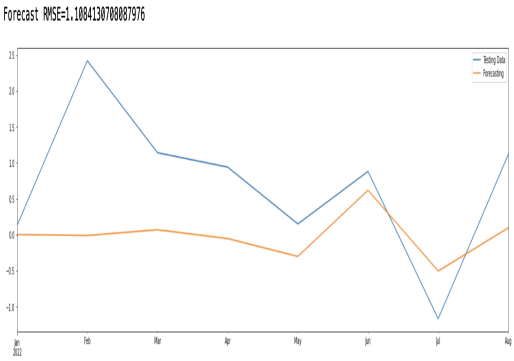

Fig. 31: One Step Test Small

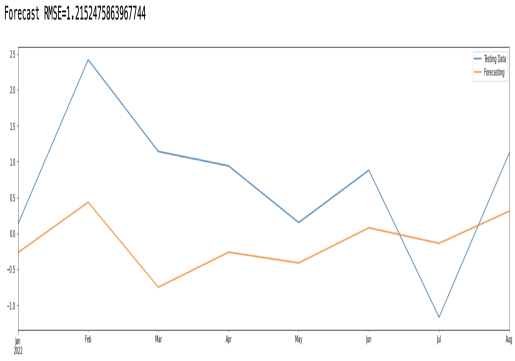

Fig. 32: Several Step Test Small

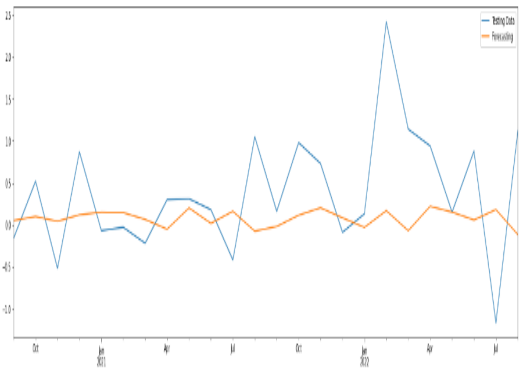

Fig. 33: One Step Test increased

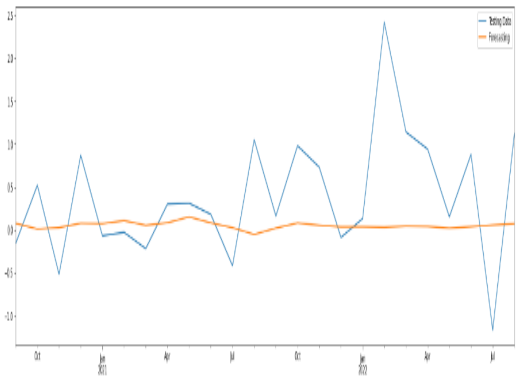

Fig. 34: Several Step Test increased

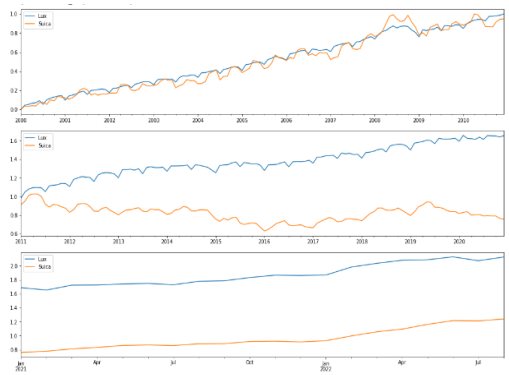

Fig. 35: TVT - Lux e Suiça

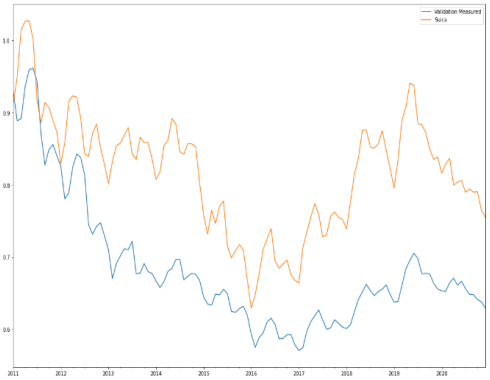

Fig. 36: Validation Forecast

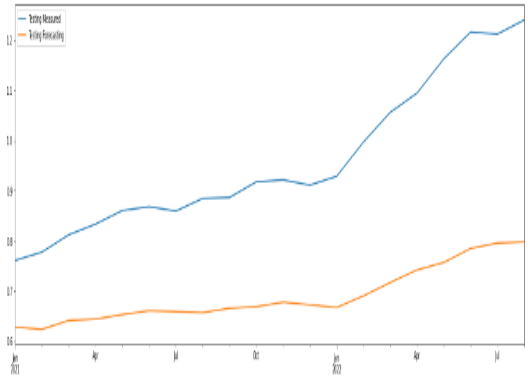

Fig. 37: Test Forecast