Central African Republic Exports Time Series Analysis

Authors: Siddharth Das, Christine (Yinyin Li), and Martin Topacio

Introduction

Central Africa's economy is heavily reliant on agriculture, which comprises over fifty percent of its GDP due to the fertile land and abundance in natural resources. These natural resources include water and minerals (diamonds, gold, copper, iron) that make up roughly half of Central Africa's export money. These exports, alongside other goods such as timber, coffee, and cotton, are the foundation that Central Africa relies on for its economy. In this report, we will incorporate various statistical methods to analyze the Central African Republic Exports time series data from 1960-2017. It is important to remember that time series analysis is dynamic, moving forward and backward between steps due to experimentation, rather than solely moving forward. We intend to build an ARIMA model to analyze the data by following a general procedure that includes plotting the data in order to identify unusual observations, transforming the data if necessary, taking differences of the data if they are not stationary, examining the ACF and PACF, testing our chosen models based off of the ACF and PACF, checking the residuals to see if they resemble white noise, then finally finding the forecasts of future outcomes.

Step 1: Plot the data and identify any unusual observations

The general process of data analysis initiates by visualizing the data to detect any unusual observations or distinct patterns. This visualization is titled "Central African Republic Exports Over Time". Observing the time series plot, we consider seasonality, and notice a lack of stationarity as the exports exhibit a declining trend over time. Seasonality can be indicated by systematic variations or patterns in the data that repeat at fixed intervals within each cycle, such as daily, weekly, monthly, or quarterly patterns. Since the export observations do not show repeating fluctuations or patterns within the year, we determined that seasonality is not present in this time series. On the other hand, stationarity in a graph is indicated by the absence of trend and seasonality. A time series achieves stationarity when its mean remains constant over time, and its autocovariance function shows no time dependence. According to the original time series plot, it is evident that exports are consistently decreasing over time, indicating this time series is non-stationary. We implemented other methods for identifying stationarity, such as ACF examination, ADF, and KPSS tests. The ACF provides valuable insight into the behavior of time series data, specifically shining a light on candidate models and the correlation between the observations and their past values. The ACF plot of the original time series illustrates a decreasing linear trend, which further supports our hypothesis that the exports are non-stationary. Ideally, we hope to see an exponential decay in ACF plots to represent stationarity.

ADF and KPSS tests were also utilized in identifying stationarity. For the ADF test, the null hypothesis is Ho: The data has unit root and is non-stationary vs the alternative hypothesis Ha: The data does not have unit root and is stationary. The conclusion for this test is based on the comparison of the p-value and the assumed value of alpha = 0.05. If the p-value is greater than .05, we fail to reject the null hypothesis. Alternatively if the p-value is less than or equal to .05, we reject the null hypothesis. The Augmented Dickey-Fuller Test resulted in p-value = .1006, which is bigger than the assumed value of alpha = .05. Thus, Augmented Dickey-Fuller Test supports the conclusion that exports is a non-stationary time series.

Regarding the KPSS test, the null hypothesis is Ho: the data has unit root and is non-stationary, versus the alternative hypothesis Ha: The data does not have unit root and is stationary. The conclusion for this test is based on the comparison of the p-value and the assumed value of alpha = 0.05. If the p-value is greater than or equal to .05, we reject the null hypothesis. Alternatively, if the p-value is less than .05, we fail to reject the null hypothesis. The KPSS Test for Level Stationarity resulted in a p-value of .01, which is less than the assumed value of alpha = .05. Hence, the KPSS Test for Level Stationarity also supports the conclusion that exports is a non-stationary time series.

After multiple plots and testing to identify stationarity, we can claim that the export data is nonstationary. We now want to transform the time series to ensure stationarity. Models for estimation and forecasting tend to assume consistent statistical properties such as mean and variance over time. If these assumptions are not met, inferences will be less interpretable, and conclusions about the behavior of the time series will be less accurate.

Step 2: Data transformation

Before proceeding with our time series analysis, we must confirm the stabilized variance property of stationary time series. Heteroskedasticity, or non-constant variance, is often present for time series data from the real world. The opposite of heteroskedasticity is homoscedasticity, referring to constant variance. Since perfectly constant variance is relatively rare in time series, we want to proceed with the form of data where the variance is as stable as possible. The most commonly utilized methods for stabilizing the variance are data transformations and differencing. We will explore both methods in this analysis. Data transformations are often utilized in time series modeling to improve stationarity, homoscedasticity, normality, forecasting (prediction) accuracy, etc. Since our goal is to ensure variance stability, we will consider Box-Cox, Logarithmic, and square-root transformations.

In time series, white noise is independent and identically distributed (i.i.d.) with a constant mean and variance. In simple terms, white noise should appear random without displaying any distinct patterns. Residuals represent the errors made by the model predictions, calculated as the difference between the predicted values of the model and the true, observed values. Similarly for a linear regression model, the residuals are assumed to be independent, following a normal distribution, with a constant variance. Since both white noise and residuals have very similar properties, we can check for heteroskedasticity by utilizing the residuals vs. fitted and normal Q-Q linear regression plots and identifying normality. In addition, we will examine the ACF and PACF plots to further support our understanding of whether transformations improve stationarity (variance stability). Through the ACF and PACF model we could identify any trends or patterns that yield different results from the transformations.

After performing Box-Cox, log, and square-root transformations on the data, we sought to evaluate each transformed data for its stationarity. The residuals vs. fitted and normal Q-Q plots both revealed that the variance had not changed in terms of being unconstant for all transformations, meaning they all resembled the original data. The ACF and PACF also did not show any significant improvement in terms of stationarity upon the original data. The ACF and PACF models for the transformation show a similar linear trend that is shown on the original export data. The one transformation that potentially rose above the others was the log transformation, as it had a lower AIC and BIC value than the other transformations. These results lead us to declare that performing transformations upon the data is not necessary, and instead, we will now seek to take differences of the data to make it stationary.

Step 3: If the data are non-stationary, take first differences of the data until the data are stationary

After concluding that the transformation is not necessary, the next step is to ensure the time series becomes stationary before proceeding with the modeling process. To achieve stationarity, we will apply differencing to the time series to stabilize the data by adjusting its mean, thereby eliminating fluctuations and reducing or eliminating trends and seasonality within the dataset. In other words, we will compute the first difference by subtracting the exports of each year by the exports of the previous year. After taking the difference, we created a time plot, ACF and PACF to again check if data is stationary. A stationary time plot will show the series to be roughly horizontal with constant variance. The first difference time plot showed constant variance throughout time compared to the original data, which shows a downward trend through change in variance. The ACF and PACF also show an exponential decay pattern indicating stationarity of data. We also applied the KPSS test to prove stationarity. Our KPSS output p-value is 0.1, which is greater than .05, so we reject the null hypothesis and accept the alternative that the data does not have unit root and is stationary. Now that the data is stabilized through the first difference, we will go ahead and move onto step 4.

First differences formula: (X_t) - (X_(t-1))

Step 4: Examine the ACF/PACF: MA or AR for the difference?

When examining the ACF/ PACF models, we want to identify which potential models we would utilize to help forecast future data. From the ACF plot, we identified that MA(3) will be a chosen model since the ACF cuts off (passes the blue dotted line) at lag three. MA (moving average) models are effective in capturing short-term fluctuations or irregularities in the data by modeling the relationship between the current observation and past white noise error terms. ARIMA(0,1,3) will also be another potential candidate for forecasting. ARIMA models will be taken into consideration, as they are flexible and handle a wider range of time series data. AR(1) and AR(2) will be chosen as a potential model from the PACF. We picked AR(1) and AR(2) from lag 1 and 2 showing a coefficient drop to statistically insignificant levels. The alternative candidates for ARIMA models would be ARIMA(2,1,0), ARIMA(1,1,0), ARIMA(2,1,3). Now that we have all the potential candidate models, we will proceed to step 5.

AR(1) model: X_t = aX_(t-1) + W_t, |a| < 1
MA(3) model: X_t = W_t + Θ_1W_(t-1) + Θ_2W_(t-2) + Θ_3W_(t-3)

Step 5: Try your chosen model(s) and use the AIC/BIC to search for a better model.

In step 5, our objective is to select the most suitable model among those identified as potential candidates in step 4. To determine the best fit model, we analyze the AIC and BIC values of each candidate. Lower AIC and BIC values indicate a better fit. In addition, the criteria for selecting the better model is determined by the model with the smallest difference between its AIC and BIC values. Amongst the MA(3), AR(1), and AR(2) models, AR(2) exhibits the smallest AIC (275.2535) and BIC (283.4257), suggesting superior fit. Among the ARIMA models considered - ARIMA(2,1,0), ARIMA(1,1,0), ARIMA(0,1,3), ARIMA(2,1,3), ARIMA(2,1,0) with drift, ARIMA(1,1,0) with drift, and ARIMA(0,1,3) with drift, ARIMA(2,1,3) with drift - ARIMA(2,1,0) yields the lowest AIC (274.5368), BIC (280.666), and difference, indicating the best fit within this subset. With the top two potential candidate models identified, we proceed to compare their AIC and

BIC values to determine the optimal model. ARIMA(2,1,0) displays lower AIC and BIC values, making it the preferred model for forecasting future predictions.

Step 6: Check the residuals from your chosen model by plotting the ACF of the residuals and doing a portmanteau test of the residuals. If they do not look like white noise, try a modified model.

Now that we have classified our ideal model in ARIMA(2,1,0), we move on to checking the residuals of the model by plotting the ACF of residuals and performing the Ljung Box test to identify if the model shows prominent white noise. For forecasting time series, we ideally want the residuals of the chosen model to represent white noise, as it signifies independence, constant mean and variance. An ACF plot of the residual should not show any significant spikes beyond the confidence intervals (the blue dotted line). If the lags are within the intervals, it indicates no remaining autocorrelation, hence a white noise model. In our ACF of residuals, all of our lags are within the confidence interval. The normal Q-Q residual plot also looks approximately normal. In our performance of a portmanteau test of residuals using the Ljung Box test, we claim that: H0: The residuals are independently distributed,HA: The residuals are not independently distributed; they exhibit serial correlation. Our p-value result is 0.2194, which is greater than the significance level of .05. As a result, we fail to reject the null hypothesis and claim that the residuals are independently distributed, indicating that there is no significant autocorrelation in the residuals. With these conditions being satisfied, we may conclude our model to be a white noise model.

Step 7: Once the residuals look like white noise, calculate the forecasts.

After confirming that our selected ARIMA(2,1,0) model resembles white noise in the form of constant mean and variance, we now move on to calculating forecasts for the exports data with our model. We achieve these forecasts using the forecast package in R, and we furthermore visualize the forecasts in a time series format with an autoplot. Our ARIMA(2,1,0) model, which we have determined to be the optimal model, has the ability to forecast future export data for the Central African Republic. Our forecast graph predicts that the exports will be relatively stable over the next ten years, without any drastic increases or decreases. This stability over the next ten years falls in line with the level left at 2017, the most recent year in the dataset, giving the Central African economy a beneficial estimate on what level of success to expect from its exports. The country should neither expect any dramatic downturns nor major spikes from their exports within the next ten years. Our autoplot also features a prediction interval for possible fluctuations in the data over the predicted years. This prediction interval accounts for the potential in fluctuations from the predicted line in the graph, with the interval at the beginning of the ten years being roughly a little over 10 exports wide, then growing to roughly over 20 exports wide by the end of the ten year prediction. Furthermore, these results should serve as a boon to the Central African economy, as their leaders are able to predict future results of their business, giving them an idea on what to expect in terms of potential failure or success.

Conclusion

In this project, we aimed to forecast future time series data for the Central African Republic exports industry by building an ARIMA model. We first determined the data to be non-stationary, which led us to attempt transformations on the data to make it stationary. After taking various forms of transformations, including Box-Cox and log, into consideration, we came to the conclusion that transformation was not necessary because any attempt failed to stabilize the variance. To achieve stationarity, we took differences of the data. We verified the stationarity of the data post-differencing using ACF and PACF plots and the KPSS test. The next step focused on determining candidate models for the forecast, where we looked at different AR and MA models, as well as ARIMA models. After testing out each model and evaluating them by the AIC and BIC best fit criteria, the ARMA(2,1,0) model ultimately emerged as our best model due to its lower AIC and BIC values. After looking at ACF and residual plots and conducting the Ljung Box test, we confirmed that our selected model is, in fact, a white noise model. Once this was verified, we calculated the forecasts using the model, finding that ten years after the data, the exports will be relatively stable with room for possible fluctuations. The main predicted line in the plot indicates that Central Africa need not worry about any downturns in their economy, but also not to expect any major spikes over the next ten years as well. In determining these predictions, we have completed our main objective for this project in forecasting future data for the Central African Republic exports industry. We hope that this project allows the Central African Republic to gain a greater understanding of the future potential of their economy and how being able to analyze time series data can prove to be beneficial to their country.

Sources

1. Final project.pdf - Instructions
2. W7-3_annotated.pdf - 7 Steps
3. W7-2_annotated.pdf - ARIMA Differencing
4. W7-1_annotated.pdf - Forecasting for ARMA
5. Discussion 9 - AIC/BIC
6. Discussion 8 -
7. https://otexts.com/fpp2/arima-r.html - auto.arima()
8. https://datascienceplus.com/time-series-analysis-in-r-part-2-time-series-transformations/ - Data Transformations
9. https://search.r-project.org/CRAN/refmans/feasts/html/portmanteau_tests.html - portmanteau test
10. Statistical Tests to Check Stationarity in Time Series (analyticsvidhya.com) KPSS and ADF test

Code Appendix