REPORT: TIME SERIES ANALYSIS

Statistical Analysis HW 3

Junhui Liang, Miguel Pérez, Ángel Igareta

January 10, 2020

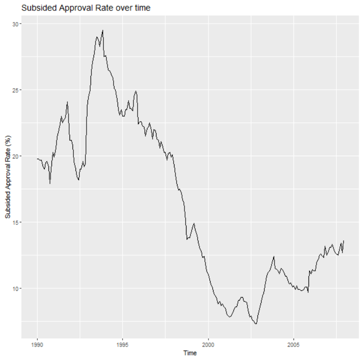
1 Dataset

1.1 Introduction

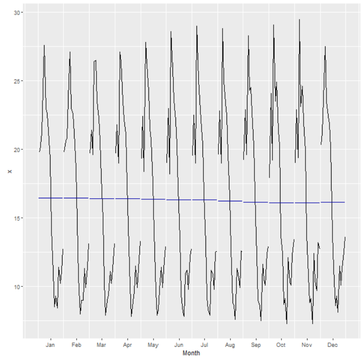
The aim of this report is to analyze a specific time series which has been assigned to our group 15, as well as to fit an ARIMA model following the Box-Jenkins Methodology. The time series assigned presents subsidised housing approvals as a percentage of the total approvals. The dates in the dataset range from January 1990 to December 2007 and the source is **Banco de España**.

1.2 Time Series Characteristics

In this first section, we will plot the time series (figure 1a) and analyze the components that can be observed, which are the trend, seasonal, random and stationary ones.



(a) Time Series



(b) Time Series monthplot

First of all, we can observe three main trends over the years. From 1990 to 1994 the trend in housing approval rate was to increase, from 1994 to 2002, the trend decreased rapidly and finally, from 2002 to 2007 the trend has been increasing. If observed from a global scope, the general trend has been to decrease, from 20% to 13.8%.

In order to analyze the seasonality, as it can not be appreciated in the general plot, we have used ggmonthplot, after loading the time series with the frequency value of 12. As observed in the figure 1b, the output of this function show a high seasonality in the data, where the mean has almost the same value in all the seasons.

With respect of the stationary property of the time series, we can see that the general time series is not stationary, because of the high variation of the housing approval rate mean over the years. However, to statistically prove it, we performed the Dickey-Fuller Test (*adf.test*) and the resulting p-value was of 0.41. Hence, we fail to reject the null hypothesis of the time series being non-stationary.

1.3 Decomposition

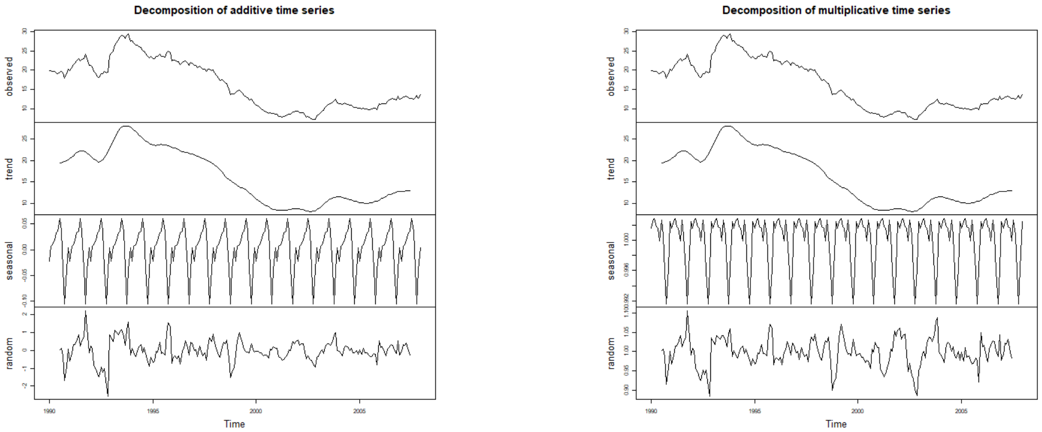
In this section, we will decompose the time series to divide it in *trend*, *seasonal* and *remainder*. For that, we will use the R function *stl()*. However, we need to choose between performing an additive decomposition or a multiplicative one.

The additive decomposition is recommended when the variation around the trend-cycle does not vary with the level of the time series. However, if that variation appears to be proportional to the level of the time series, a multiplicative decomposition would adjust better. As in our time series is difficult to analyze the variation around the trend-cycle, we will perform both decompositions and analyze the results [1].

In order to plot the additive decomposition ($y_t = S_t + T_t + R_t$), as it is the default from stl, we would only have to plot the result. However, the multiplicative decomposition follows a different formula: $y_t = S_t \times T_t \times R_t$. In order to use it with stl, we would need to first transform the data with log and then use an additive decomposition, because:

$$y_t = S_t \times T_t \times R_t \quad \text{is equivalent to} \quad \log y_t = \log S_t + \log T_t + \log R_t.$$

The components of both decompositions can be observed in the figure 2. It is important to note that the multiplicative decomposition components are transformed with log, in order to back-transform the data we could use the exp method.



(a) Additive Decomposition

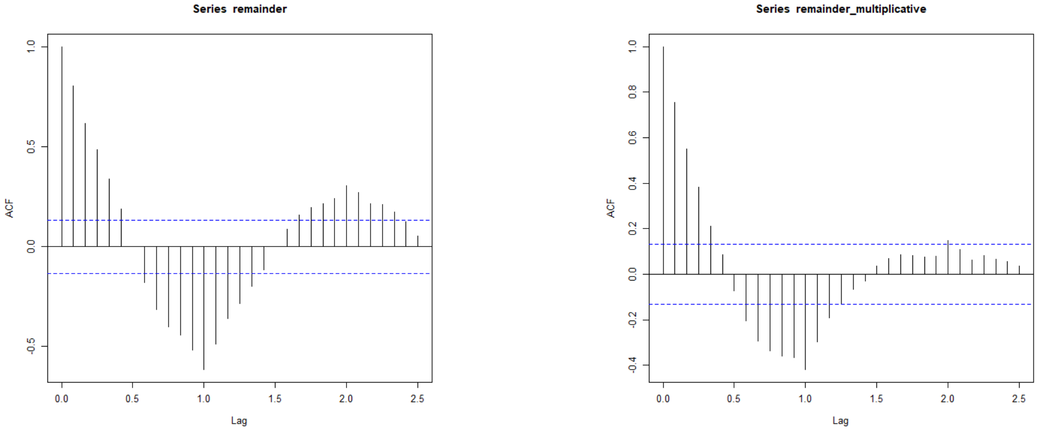
(b) Multiplicative Decomposition

Figure 2: Components of the Time Series

At first glance, we can appreciate a high difference in the remainder residuals for both plots. In the additive decomposition plot, the remainder does not appear to be *white noise* (constant variance and zero mean), as it should. Instead, it seems to follow a decreasing trend over time. On the other hand, the remainder for the multiplicative decomposition does seem to satisfy the white noise definition.

Now, we will statistically check the previous presented assumptions. The mean for both remainders (from additive and multiplicative decomposition) is of -0.043 and -0.009 respectively, so both of them satisfy the zero mean assumption. According to the constant variance assumption, we performed the Dickey-Fuller Test in both residuals and we got the same p-value with value 0.01, confirming that both of them satisfy the assumption.

On top of that, if we plot the Autocorrelation And Cross-Correlation Function Estimation (ACF) for both remainders, as observed in the figure 3, we can see that both plots are very similar, but the one for the multiplicative component has lower correlation with the past values.



(a) Additive Decomposition Remainder ACF

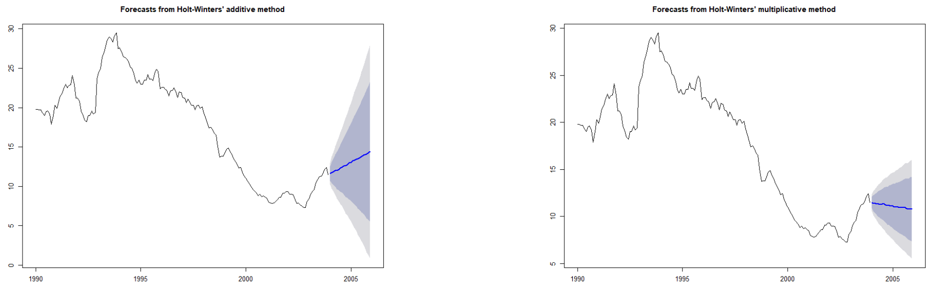
(b) Multiplicative Decomposition Remainder ACF

Figure 3: Remainders of the Time Series Decomposition

In conclusion, as presented in the above results, both decomposition present white noise in the residuals so both are valid.

1.4 Decomposition Models Forecasting

Here is a representation of a forecasting using multiplicative and additive decomposition from Holt-Winters HW approach. We can see the results are not successful, so we will try to use ARIMA to try to improve them.



(a) Additive Decomposition Remainder ACF (b) Multiplicative Decomposition Remainder ACF

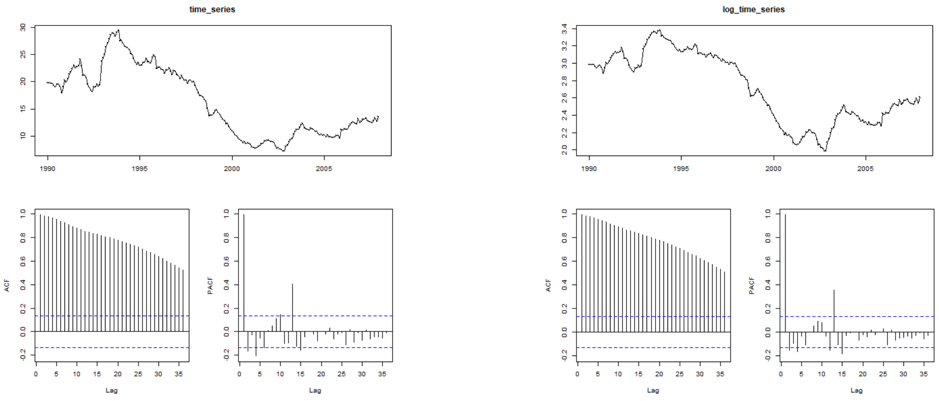
Figure 4: Remainders of the Time Series Decomposition

2 Model Selection

In the second section, we would fit an ARIMA model to our time series, presenting and justifying which components and values we are going to use for the model.

2.1 Transformation Selection

First, in order to decide which variable we are going to use between the original time series and the log transformed one, we decided to plot each one and see if the transformation contributes to a better variance. As observed in the figure 5, both time series look very similar, and neither there is a difference in ACF nor PACF plots.



(a) Original Time Series (b) Log Transformed Time Series

Figure 5: Transformation in Time Series

Due to this, in the result section, we will try the best configuration for both time series and analyze which one has better results.

2.2 Seasonal Component Selection

Based on the previous seasonality analysis, it is clear there is high seasonality in the time series. However, monthly or quarterly data does not mean it is necessary to include a seasonal component in the model.

In order to make this decision, we chose the empirical method, comparing models with and without the seasonal component, in other words, setting the (P, Q) parameters as 0 in ARIMA as the way of representing a model without seasonal component. We implement this process in the permutation function explained with details in the below. Experimentally, we conclude to introduce seasonal component is significant to enhance model performance.

2.3 Differences Selection

The next task is to decide which differences level to use to make the time series we are analyzing stationary. Before doing any transformation, we performed the Dickey-Fuller Test (*adf.test*) and the p-value was higher than 0.05, meaning that the time series is not stationary. On top of that, we performed the KPSS test and we obtained a p-value of 0.01 where the null hypothesis is that the series is stationary. So in both test we obtained the same preliminary conclusion. Besides, the standard deviation of the time series was of 6.24.

After differentiating with a value d of 1, we reduced the standard deviation to 0.62 and obtained a p-value of 0.01 in the Dick-Fuller Test and 0.1 in the KPSS test, statistically proving that our differentiated time series is now stationary.

Regarding the seasonal differences, we added a $D = 1$ to the previous differentiated result and we obtained a higher standard deviation of 1.1. Besides, the acf were very similar, so we decided not to use a seasonal difference, only $d = 1$.

2.4 Order Components Selection

In order to select which combination of order components obtained the better ARIMA model, we created a function that, given a list of possible parameters per order component p, q, P and Q, per each permutation of those values, it would train an ARIMA model and obtain the RMSE, getting the top 3 models with better results. After that, we would take into account the models' residuals and the parameters selected to choose the best model input.

The selected order components would depend on the time series that was used to train the model, the raw or the transformed one, as it would be presented in the results.

3 Results

3.1 Tests

As presented in the section 2.4, we use a method to automatically choose the parameters that better work with our time series and get a top x of the resulting models. We tried the permutations from 0 to 3 for each order component, being a total of $4^4 = 320$ permutations. This was repeated for the log transformed time series as input, transforming in this case the test dataset too.

On top of that, we tested training some models without the seasonal component and we obtained a better result when including it.

3.2 Final Combinations

After performing all the previous parameter permutations, the final selected models were:

- For the raw time series the best parameters found were p = 1, d = 1, q = 0, include seasonal component and P = 3, D = 0 and Q = 1. This combination obtained an RMSE of 0.68 in the time series forecasting of one year.
- For the log transformed time series the best parameters found were p = 1, d = 1, q = 1, include seasonal component and P = 1, D = 0 and Q = 0. In this case, the model obtained an RMSE of 0.08 in the same forecasting time window. However, this result is in log scale.

In the figure 6 we can observe the forecast for both models from 2004 to 2005. Note that the red forecast is the model trained with the log transformed time series but it has been back-transformed to plot it next to the other model.

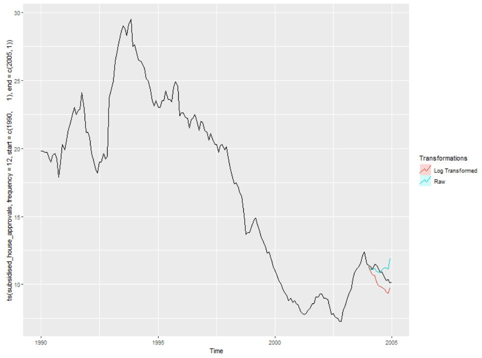


Figure 6: Time Series Forecast

In order to further compare the models with and without log transformer, it is imperative to check the assumption of models.

- For the raw time series, via the t-test, the p-value of the null hypothesis that true mean is equal to 0 is 0.02886, indicating it is high possible to reject the null hypothesis. On the other hand, the p-value of Box-test is 0.001325, a very small value as the evidence that there is dependence.
- For the lag transformed time series, the p-value of t-test is 0.4181, implying true mean is equal to 0 with high confidence. Furthermore, with respect to Box-test, the p-value is 0.7123, manifesting there is no dependence.

Above all, based on the consideration of the RSME and assumption verification, it is highly recommendable to apply log transformer to the raw time series. By means of "sarima" function, we can check more characteristics of the model in the figure 7, which reached the balanced performance .

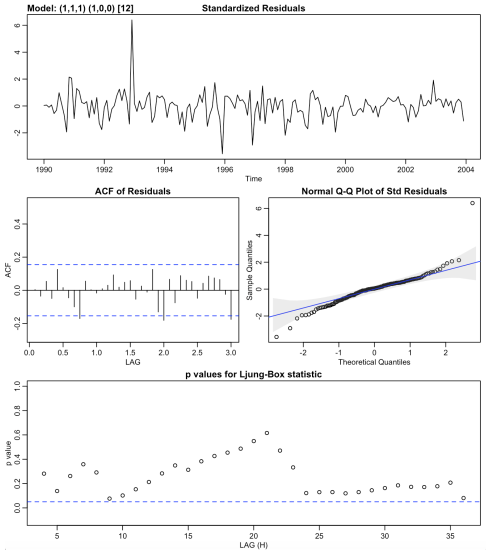


Figure 7: Residual analysis of log_time_series

3.2.1 Autoplot Approach

Finally, we performed the automatic ARIMA model training in order to compare the best resulting model with our final model. We obtained a model worse than our previous versions, with a RMSE of 1.07, so in this case, with Box-Jenkins Methodology we obtained more successful results.

Meanwhile, the confirmation of assumption is listed as the following. With regard to t-test, the p-value is 0.8792, proving it failed to reject the null hypothesis that true mean is equal to 0. What's more, the p-value of Box-test is 0.4757, which is high enough to be the evidence of independence. As the same time, p-value is very small in Jarque Bera test when removing some outliers, showing it is normality. According to residual plot, it also has constant variance.

References

[1] Hyndman, R. J., and Athanasopoulos, G. Forecasting: Principles and Practice. *https://otexts.com/fpp2/components.html*.