Public Bike Usage in Jalisco, Mexico: A Box-Jenkins Forecasting Study

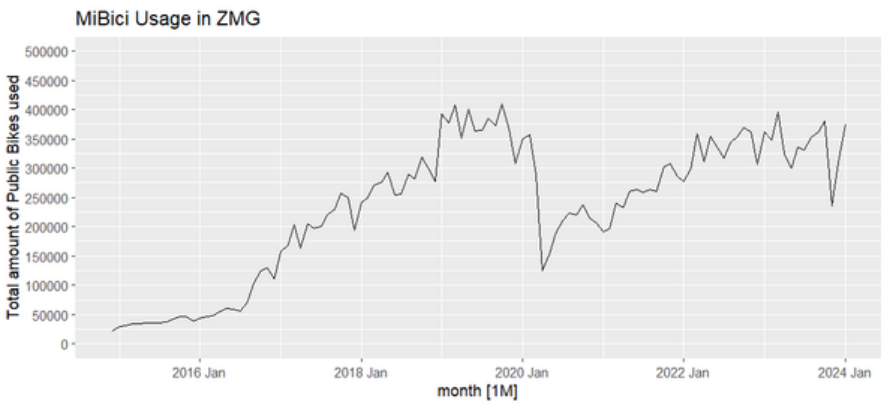
Elena Sutkuté (20232018) | Vidhi Rajanikante (20221982)

Introduction

Available to us is the captured data from 25,863,690 public bike trips and 372 public bike stations of *MiBici* in Guadalajara's metropolitan area (ZMG) which is located in Jalisco, Mexico, over a time span of nearly a decade, from December 2014 to January 2024.

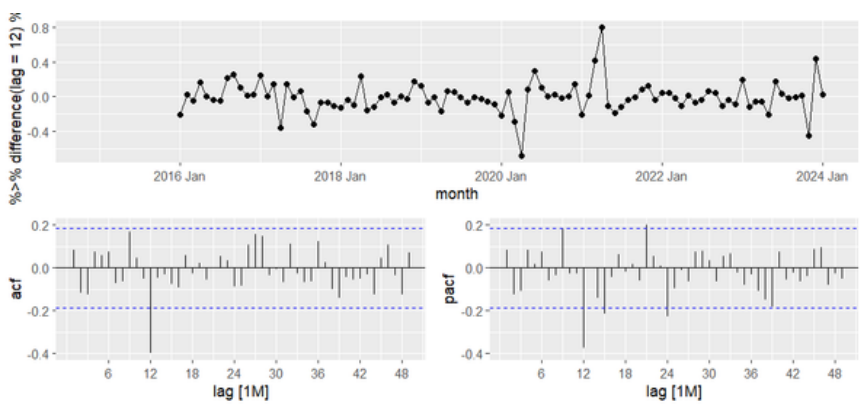
With the goal of understanding whether this trend, of using public bikes through *MiBici,* we decided to use this data as a means to apply the ***Box-Jenkins Methodology*** to forecast the past year as well as the following year to check the continual behaviour or trend in ZMG. The data, ***Public bike use data 2014-2024 (MiBici)***, was acquired from kaggle. [1] [2]

As the acquired dataset was captured daily, we summarised it into monthly data, thus having 110 observations - each observation representing 1 month, with the total amount of public bikes used. Clearly the inevitable *COVID-19*'s influence is present and shown as quite the shock in our initial plot of the dataset, after being converted to a tsibble.



ARIMA Model Identification

Dividing the dataset into a training set and a test set was important for a good forecast, as we need to train our models to show the best fit, where our training set was defined to be from December 2014 to January 2023 and our test set to be the last year of the dataset [4]. ***Log transformation*** was applied in order to deal with the variance within the dataset for stability. Applying a seasonal difference and then a normal difference to the dataset confirmed its stationarity, backed up by the unit root (***ADF***) test with type='none' and lags=0.



As both the autocorrelation and partial autocorrelation funcitons suggest the order of the models cut off after lag 12, we decided to use an AR model of either order 1 or 2 (as the peak at lag 24 is slightly above the mean) and a MA model of order 1 as the other lags portray white noise, not forgetting the fact that we have initially applied, both, a seasonal and normal differences. Apparently another portrayal, where we only applied a normal difference was shown to be stationary and suggested an AR model of order 2 and a MA model of order 1, similar to the one we mentioned above, which we also decided to use. After the selection of our candidate models, the ***Ljung-Box*** test was performed which resulted to no autocorrelation in the residuals and being normally distributed, thus failing to reject the null hypothesis, concluding that the model is a good fit for the data. [3]

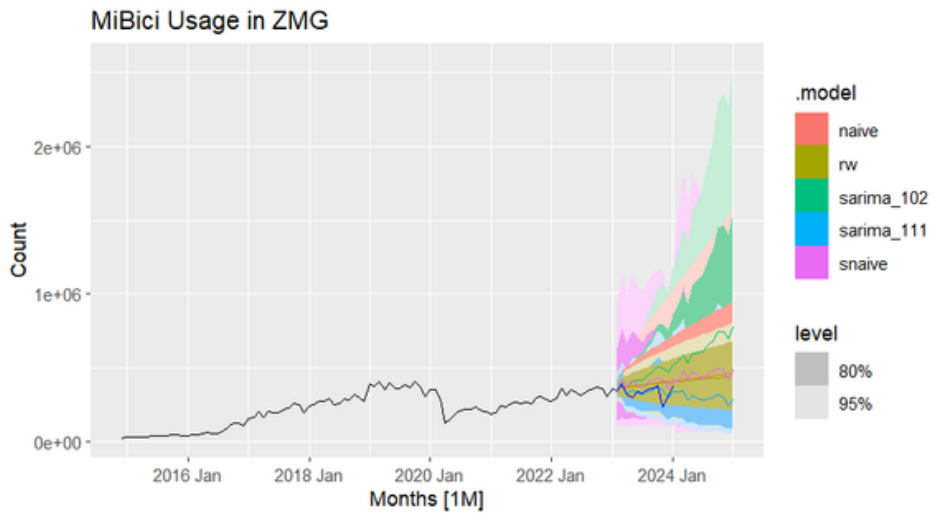
Results

For testing we chose to include not only our chosen models, but also the 'base' ones like naïve, seasonal naïve and random walk to see which one(s) showed a plausible fit to the data. Accuracy scores [4], as shown below, were calculated on the models applied for the forecasting of our test set.

	.model	.type	RMSE	MAE	MAPE
1	naive	Test	69787.74	56576.37	18.751618
2	rw	Test	65200.82	51764.27	17.262247
3	sarima_102	Test	123850.18	105749.82	33.890771
4	sarima_111	Test	38901.08	24519.23	8.478595
5	sarima_211	Test	38389.18	29453.41	9.726866
6	snaive	Test	76196.07	60974.53	20.089975

According to the resulted *accuracy scores*, 'sarima_111' and 'sarima_211' have similar values we considered to ignore 'sarima_211' due to its' worse forecast it on the test data. Additionally, we plotted forecasts for 'base' models to see if any provide better forecasts than our suggested models.

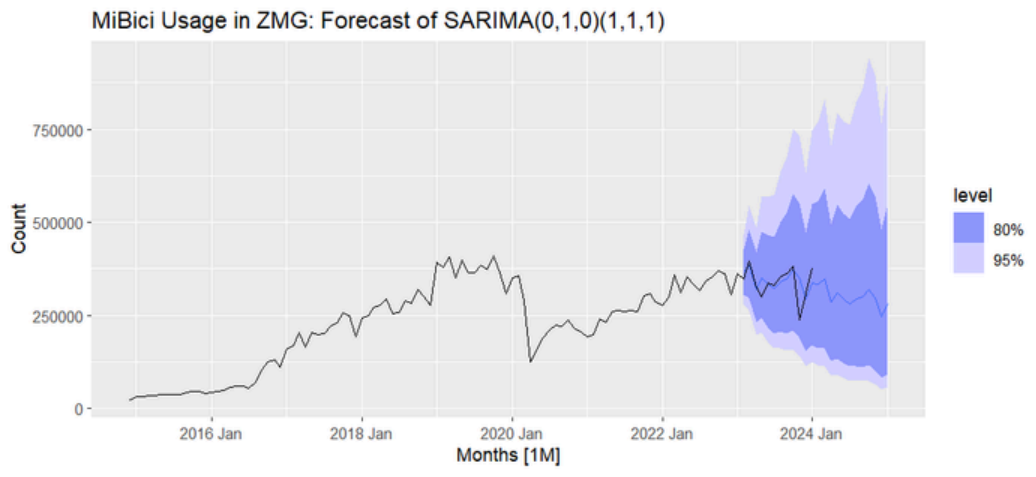
Considering the remaining models at hand, the forecasts illustrated were plausible as they portrayed the usage of public bikes in year of 2023 better when compared to other suggested models but still not as good as the dataset, which was also the reason to increase out forecasting period from the test period (12 months) to the test period plus another 12 months.



All the models except '*sarima_102*' show valid estimates for the data. Yet, the ranges of *Count* in the '*sarima_102*', '*snaive'* and '*naive'* seems to be very large, in comparison to the other models, whereas '*sarima_111*' seems to be more stable with the data. We can clearly note that all the models except '*sarima_111*' seems to overestimate the test period plus 1 year, '*sarima_102*' with the largest overestimation.

Conclusions

In regards to the graph above, we came to a conclusion that considering the range of the model, at 80% or 95% levels, not being as large in comparison to the other model and the fact that it predicts similarly to the actual data, the '*sarima_111*' (our **SARIMA(0,1,0)(1,1,1)**) is a good enough model to predict the public bike usage in Guadalajara's metropolitan area as more people tend to commute to places further away from the metro for efficiency and time saving.



Though '*sarima_111*' is not really reliable, as shown above, this may be due to the shock which is the *COVID-19* influence in the dataset - clearly showing a huge drop due to less usage of public amenities, in general. As the shock has not been treated, we can assume that this have biased our estimate towards the forecast of the data. The shock may have decreased the public bike usage but it is gradually picking up in the coming months, by aproximately 33%.

References

[1] Kaggle. (2024). *Public bike use data 2014-2024 (MiBici)*. https://www.kaggle.com/datasets/sebastianquirarte/over-9-years-of-real-public-bike-use-data-mibici/code

[2] MiBici. (2024). *Datos ABiertos*. https://www.mibici.net/es/datos-abiertos/

[3] Minitab. (Unknown). *Interpret the key results for forecast with Best ARIMA Model.* https://support.minitab.com/en-us/minitab/help-and-how-to/statistical-modeling/time-series/how-to/forecast-with-best-arima-model/interpret-the-results/key-results/

[4] Hyndman, R. J., & Athanasopoulos, G. (2023, June 8). *Evaluating forecast accuracy*. https://otexts.com/fpp2/accuracy.html