Data challenge Citadel

Khaoula Belahsen, Louis Lapassat, Louis Serrano, Jean-Noël Tuccella

March 7, 2019

1 Introduction

The Citibike bike share system was introduced in New York City throughout summer 2013, starting from the borough of Manhattan and then expanding towards the areas of Brooklyn and Queens. The traffic from one station to another depends on various parameters, such as their locations and their neighborhood demographics, and thus makes the network's structure highly challenging to analyze. One key idea of such a sharing network is to offer a system that matches the geographical demand with an appropriate offer in bikes at any time. Building on that, identifying saturated and deserted areas inside the network would be of particular interest in order to optimize the distribution of bikes across the city.

This issue is taken very seriously by large cities with bike systems. For instance, Paris through its 'Velib' bike offer, has developed an app that suggests to users in real time to end their trip at particular stations to avoid costly overnight bike transfers. In our study today, we wish to answer the following question :

Topic question: What suggestions can we provide in order to optimize the management of bike allocation based on an in-and-out flow analysis across the bike share network ?

2 Executive Summary

First, we focused on identifying clusters of stations based on their incoming and outgoing bike flows. We therefore calculated the absolute difference between the cumulated in and out flows for a fixed period of time, and then grouped the stations into three clusters using quantiles over this absolute difference: going from the most balanced one in green to the most unbalanced one in red. Comparing station to neighborhood demographics through colourful maps, comforted our clustering method, showing some overlaps between low population and low average income areas with balanced bike zones.

Subsequently, starting a trip from each cluster we computed the frequency of staying within the same cluster, and leaving for the two others. Supposing these clusters were stationary, we used them to forecast for each cluster the difference between the in and out flows for the next day. We used an ARIMA model with parameters inferred from auto-correlation and stationarity plots.

This could be a major step forward to adapt the distribution of bikes between the clusters overnight. The bike share company would use this information in order to reallocate bikes, for instance from low-stress clusters to high-stress ones. Furthermore, the company could collaborate with users to efficiently target an equilibrium between the different stations in high-stress areas. One way to do it would be to ask a customer arriving at a high-stress station to go to a close-by available station. To do that, we could look at stations within the most unbalanced (red) cluster, and identify the nearest neighbours with the lowest difference between check-ins and check-outs. The later described neighbour stations would constitute the top suggestions of alternative parking stations.

In our forecasting analysis, we assumed that the clusters were invariant in time. What remains to verify would be to keep these clusters as they are and train our forecast model on new unseen temporal data, evaluate the prediction and compare the results to validate the hypothesis.

3 Technical Exposition

Our analysis started by restraining the data to the last 6 months. This enabled us to achieve the following goals, while still providing a sufficient temporal historic for our future time series prediction:

- Make the study computationally feasile on our computers due to the large dimension of rows
- Get ride of some Montréal outliers that were present in the dataset

We also dropped missing data in the stations dataset, leaving us with a cleaned nyc_bikeshare dataset. We created several metrics to provide a quantitative analysis such as :

- difference of in and out flows for each station for the whole period of time selected:

$$diff\_in\_out = N_{check\_ins} - N_{check\_outs}$$

- for each cluster of stations : the frequency of trips staying in the cluster and going outside of it

3.1 Investigation of the Checks-Ins and Checks-out flow

To obtain first insights of the situation regarding the problematic of in-and-out flow of shared bikes, we first looked at the most popular stations of NYC quantified by the total number of trips (in + out). We analysed the global flow of bike for the latest 6 months of the dataset.

As we can see in Figure 1, most of this station are not balanced when we look at in and out flow during the period.

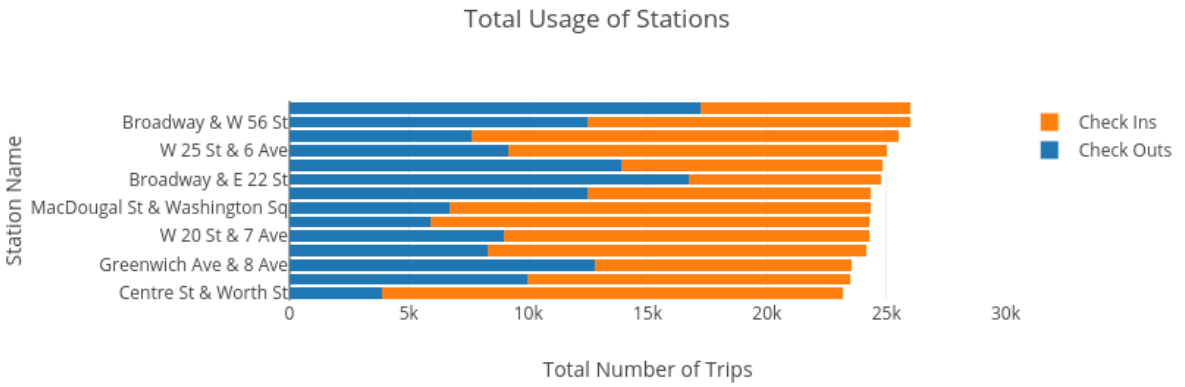

Figure 1: Check Ins and Check Outs proportion on the 15 most popular stations

Figure 1 demonstrates that the network is generally unbalanced and that some re-balancing must be done on this kind of station in order for the network to keep on working efficiently.

This stress on station can also be seen on Figure 2 when looking at the relative flow of bike on a station, i.e the difference between check-in and check-out for the last 6 months.

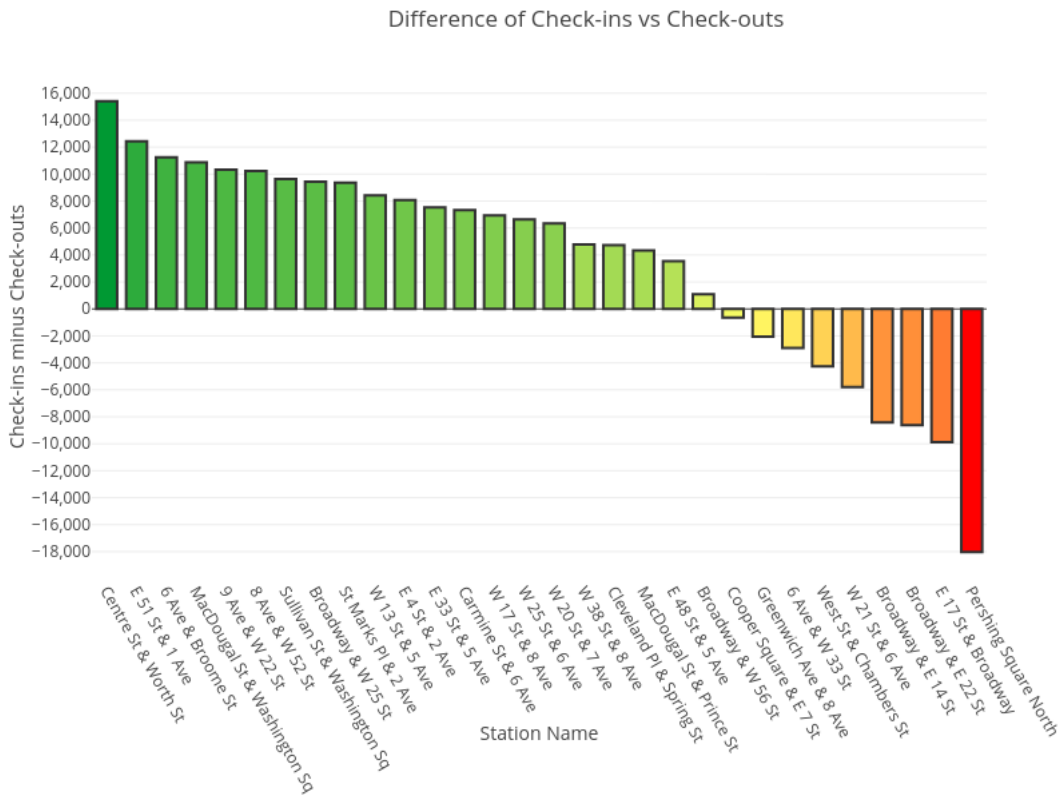

Figure 2: Difference between in and out flows for the busiest stations

This initial analysis clearly shows that this the in-and-out flow of a bike share network is unbalanced by nature. The network has to deal with this problem by physically move bikes from a station to an other. This implies an inventory management strategy. Based on this observations, we decided to focus on finding tool to optimise this strategy.

Finally, in order to have a clear overview of the NYC situation we decided to group station by clusters. We defined three clusters by looking at the absolute value of the in and out flow difference (using the terciles):

- red : highly unbalanced (1st tercile)
- orange : moderately unbalanced (2nd tercile)
- green : balanced (3rd tercile)

The results can be seen in Figure 3:

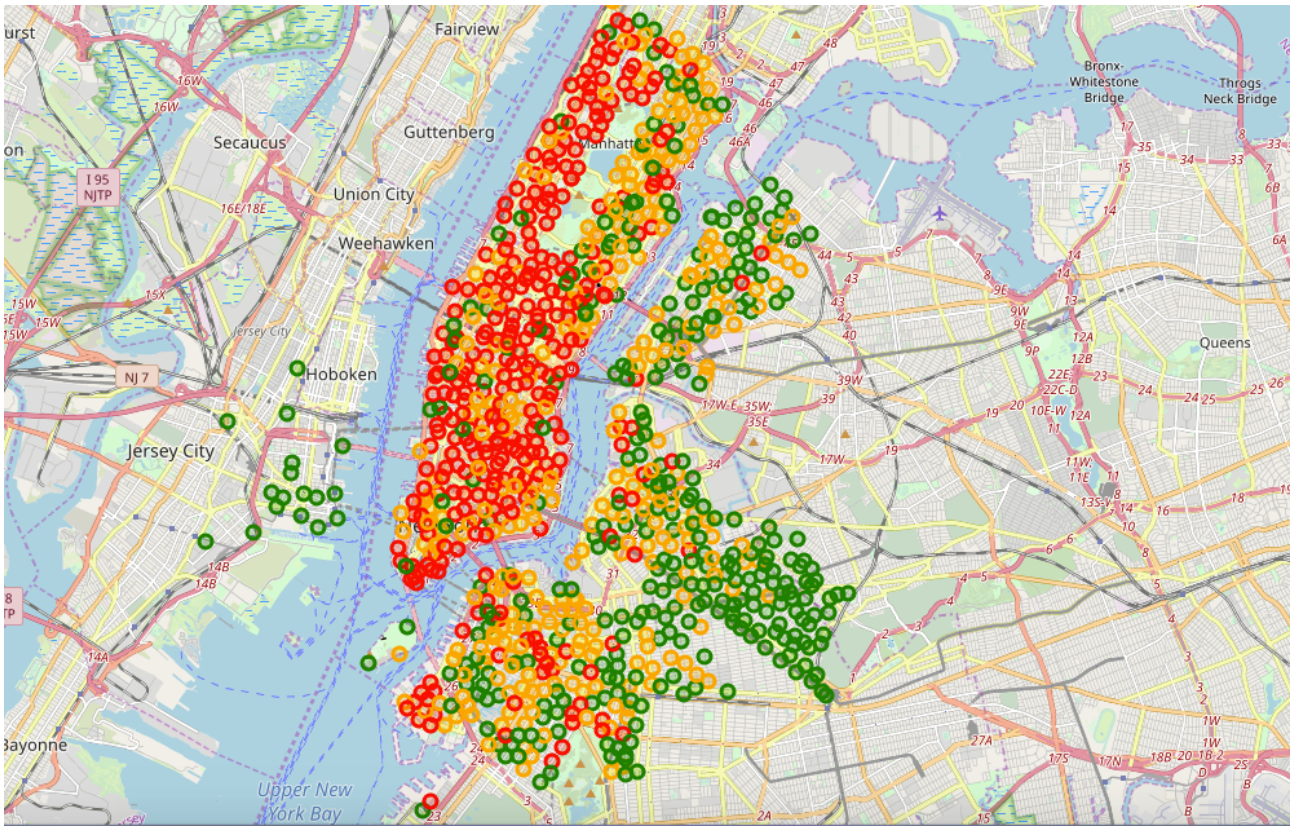

Figure 3: Station clusters in term of balance between inflows and outflows

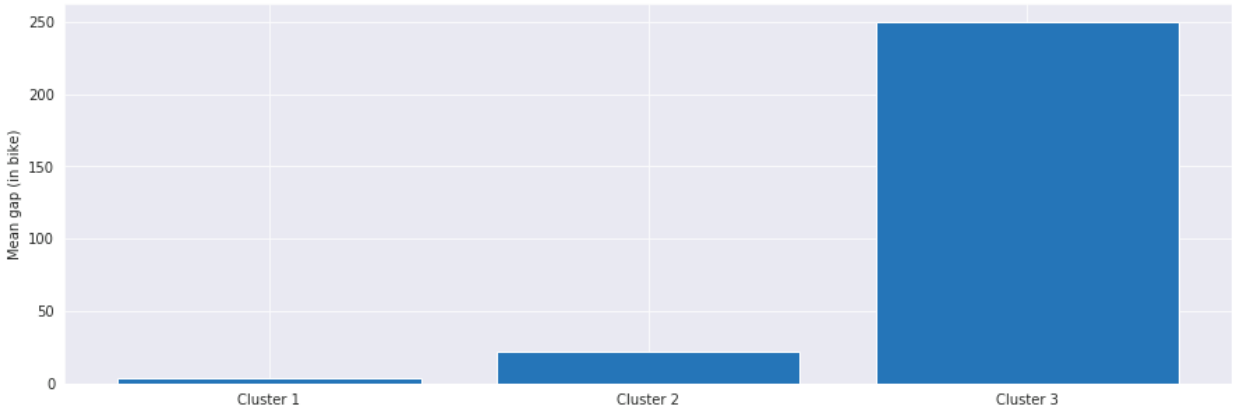

Figure 4: Mean gap in bikes across the clusters for the whole period of time

This tells us that there is a high spatial correlation on the level of stress applied to a station. We can particularly see that the Manhattan is a high-stress area. It is followed by moderate-stress area such as Brooklyn and then by balanced area in more suburbans area.

This dynamic is confirmed by the mean gap in bikes across the clusters: cluster 1 which represents the green cluster has almost a mean of 0, translating the equilibrium in that area, as opposed to cluster 3 (red one).

We then analyse the dynamic between clusters by looking at the bike flow between them:

CLUSTERS	from green	from orange	from red
to green	26%	17%	8%
to orange	33%	30%	22%
to red	41%	53%	70%

Table 1: Flow frequencies from clusters to others

We can clearly see that some the green and the orange clusters are putting pressure on the red one (almost 50% of all bikes coming from the green and the orange area are going to the red one).

This dynamic between clusters gives us important information useful for network regulation. This can be used by the network operator to re-balance the network by "artificially" moving bikes from an area to another.

3.2 Demographic analysis

In order to gain more insight about what drives bike use across different areas of NYC, we analysed the demographics of each neighborhood. Especially, we looked at segmentation of population density and mean income groups across the different neighborhoods. The results are shown in the figures below:

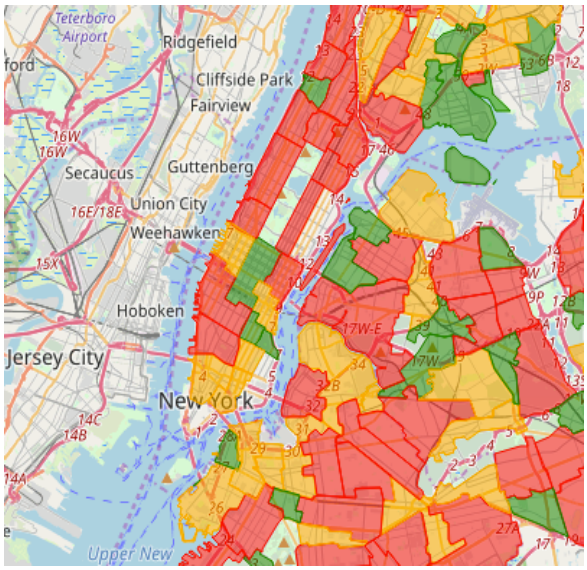

Figure 5: Population density

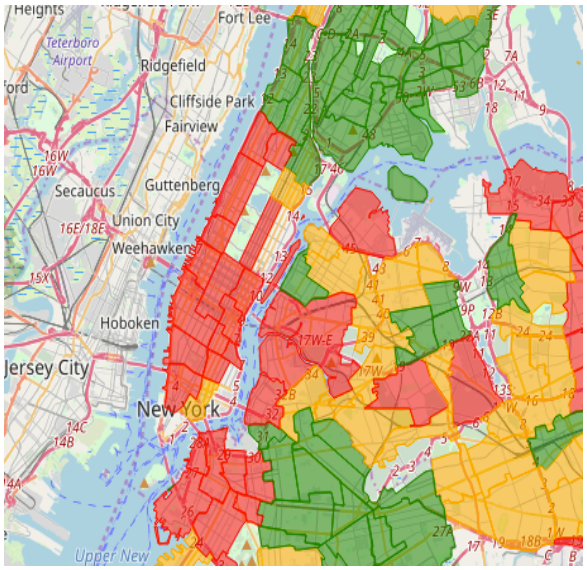

Figure 6: Income

We can notice some correlation between our demographic correlation and flow cluster : it seems that high-stress area overlap with high income and high-density zone and, on the other hand, low-stress area corresponds to low-income high-density zone. We however decided not to push further on this direction for the rest of our analysis.

Instead, we wanted to provide a prediction at time $day + 1$ for the check in and check out flows for each clusters, as it would give the bike share company a metric for their bike network optimization.

4 Forecasting

Now that we know which area are at "high stress" we want to build a model in order to deal with this (lot of in or out flow). The main idea is the following: suppose the clusters do not change over time, can we predict for each clujjeee"ster (a further idea would be to predict for each stations)

the difference of flows ($in - out$) in order to redistribute the bikes over the network (mainly to "low stressed" stations). To do so, we just split our data in $80\% - 20\%$ train/test sets and we choose an ARIMA model (mainly because we expect seasonality in our data, coming from season, holidays, etc.). But first let's take a look at our data:

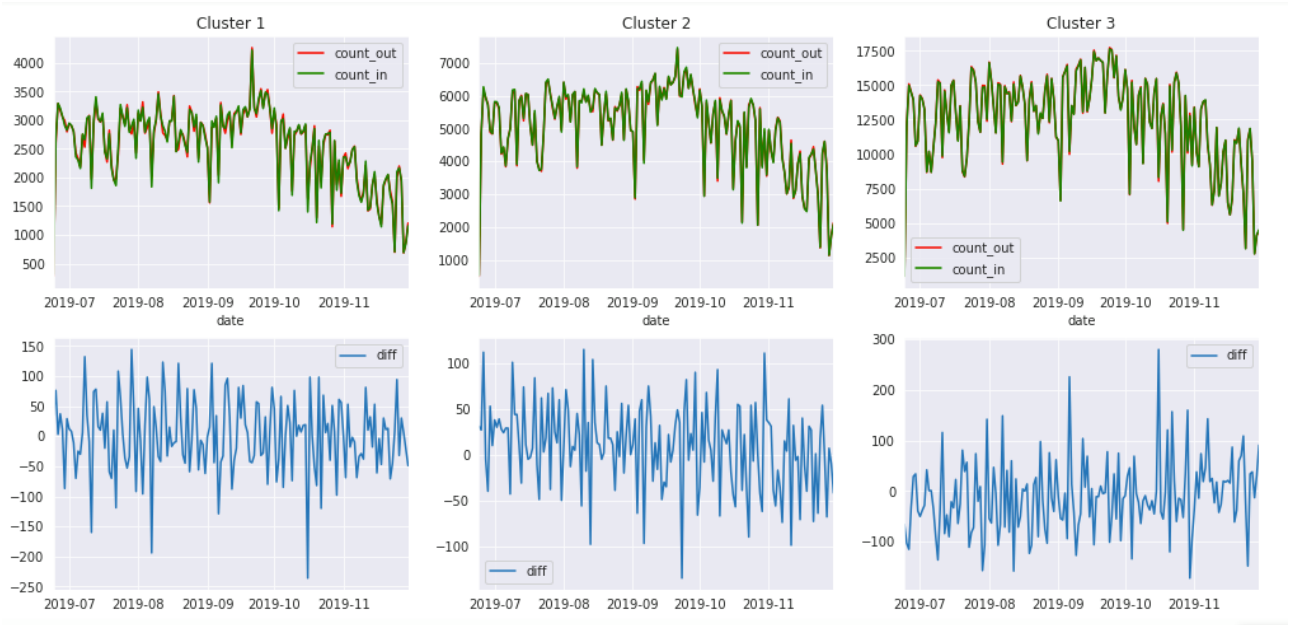

Figure 7: Up: In and out number of trips per cluster / Bottom : Difference between in and out

The above figures present (in the first line) the in/out count of bikes for each day per cluster and finally (last line) what we want to predict, the difference $in - out$. This enables us to check if the time series is sufficiently well-behaved to apply some time series models. The bottom plot does not show seasonal patterns nor upwards or downwards trends and suggest that our time series is stationary (and maybe our first idea of using an ARIMA model what a bit too much, in fact we only have here 160 days, therefore not that much to see some seasonality from season, holidays, etc.). However, the shape of the curve might reflect auto-correlation within the time series. To further investigate this and choose the ARIMA parameters, we plot the auto-correlation:

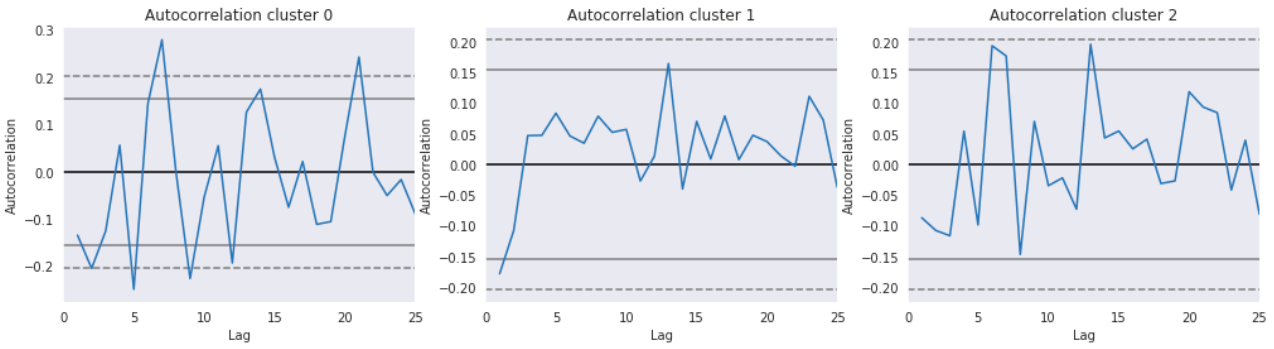

Figure 8: Autocorrelation per cluster

The figures demonstrate that there is not much auto-correlation and the first "break point" could be around $3 - 5$ (this is typically our AR parameter for ARIMA). Finally and as the above plots shown, we can assume that the series are stationary and thus the second parameter of our ARIMA

model could be $0-1$. Finally we run and fit the same model on our data (we train and test according to each cluster, so we have 3 different models sharing the same hyper-parameters):

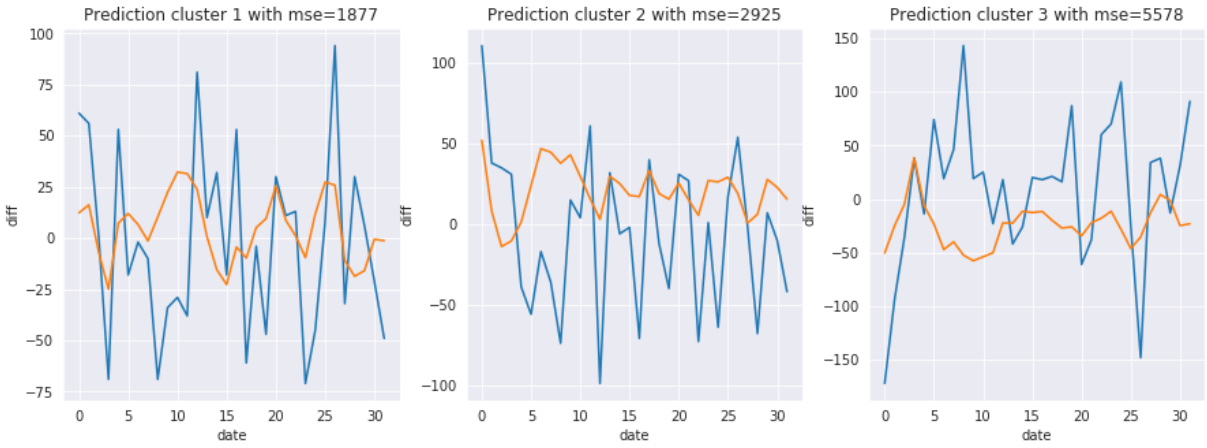

Figure 9: Up: In and out number of trips per cluster / Bottom : Difference between in and out number of

The figures illustrate that we might face some difficulties to predict the difference, $in-out$, for all clusters. It seems like the overall trend is captured (if we have a surplus of bikes or not, with the sign) but the models are not really accurate (mean squared error is quite big). Finally we think that using a grid search and more data would help enhancing the models.

5 Conclusion

In conclusion, our analysis was successful in finding clusters of stations based on a 'stress' factor, which was calculated by taking the overall absolute difference between in and out trips for each station. This is insightful as it provides understanding to the distribution of zones in equilibrium and those which highly need intervention to maintain it.

In order to provide a solution to this problem, we developed an ARIMA model to forecast for each cluster the next day difference between in and out trips. This would help citibike re-allocate bikes overnight to provide the best possible coverage of the city. Though our model is rather simple and achieves moderate performance, it provides a baseline and seems to capture overall moves of the target time series.