Overview

According to a study [1] conducted to analyze the economic costs of "parking pain":

- German drivers spend an average of 41 hours a year searching for suitable parking spots
- In monetary terms amounts to €896 per driver in wasted time, fuel and emissions
- and the country as a whole €40.4 billion

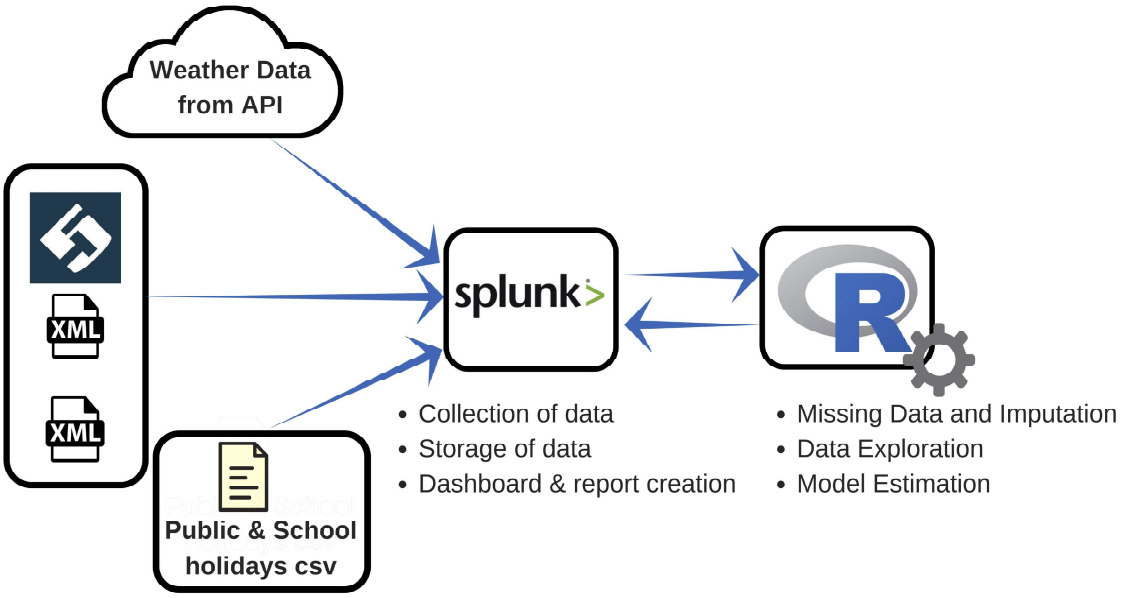
Accurate and timely prediction will help drivers and parking managers make better decisions

- "Open data" regarding the occupancy of the (non curbside) parking lots and the real time traffic situation in Torino is provided by 5T S.R.L.
- Consoft is exploring the opportunity to leverage this data.

- The goal of my work was to develop an actionable prediction model for predicting the occupancy of the parking spaces
- 4 models were experimented with:
  - Multiple Linear Regression
  - Seasonal ARIMA
  - Artificial Feed Forward Neural Networks
  - Support Vector Regression

The Data

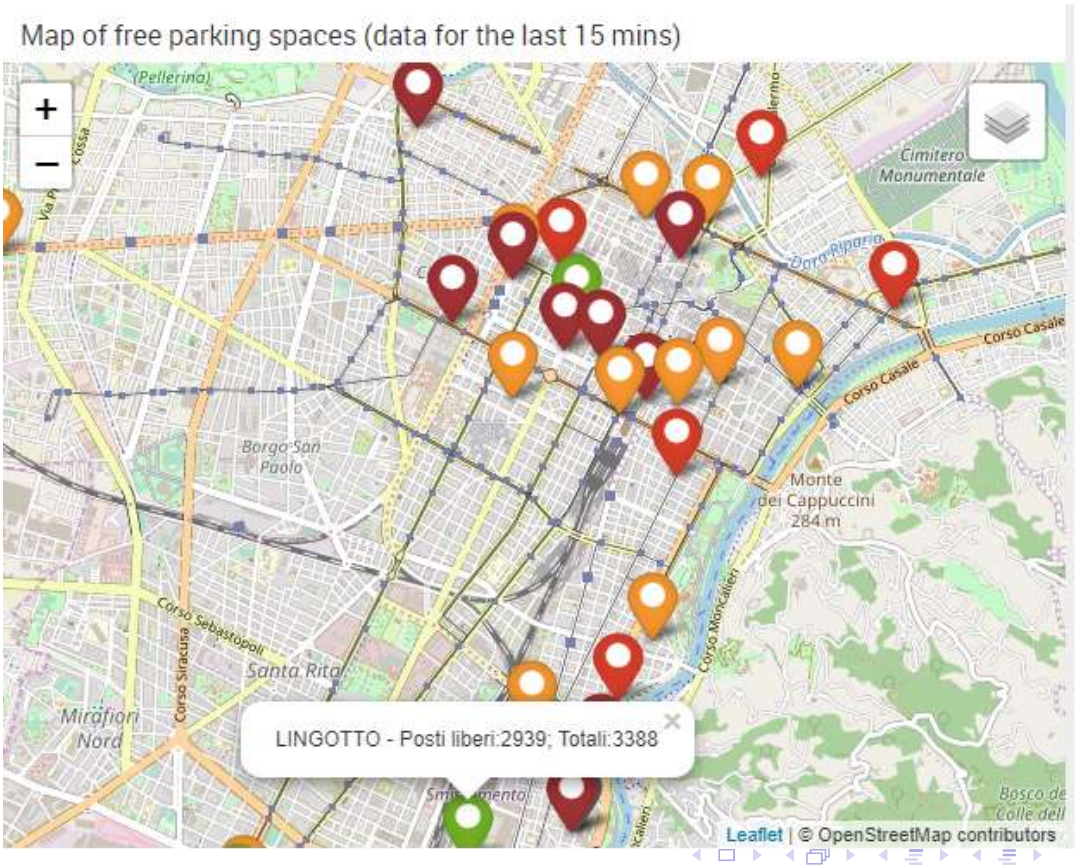
Data sources and flow



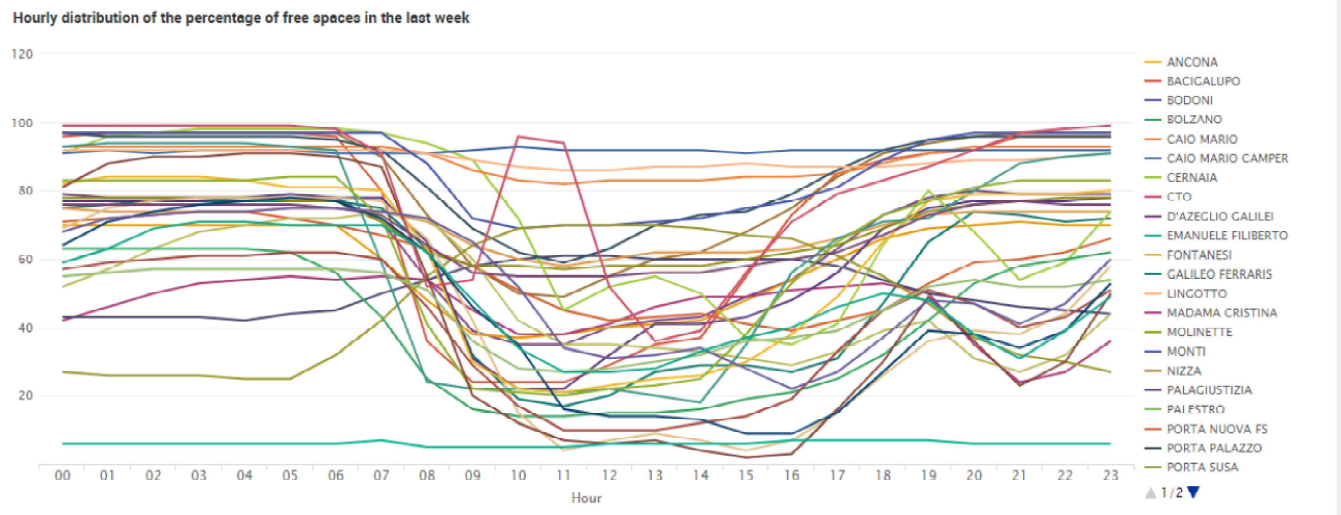
- Length of data : September 2017 - mid November 2018.
- Frequency of data : Every ten minutes aggregated to hourly.

Data Visualization before modelling

Availability of parking in various parking lots across Turin



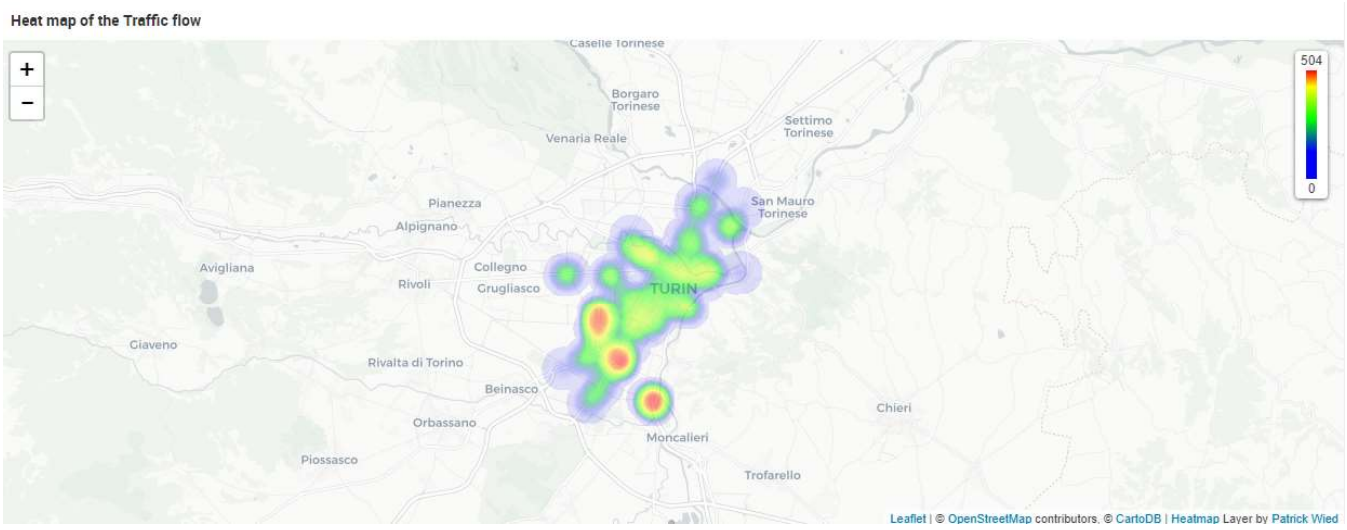
Hourly distribution of the availability of parking for different parking lots



The Data

Data Visualization before modelling

Heat map of the traffic situation on a particular day



The Data

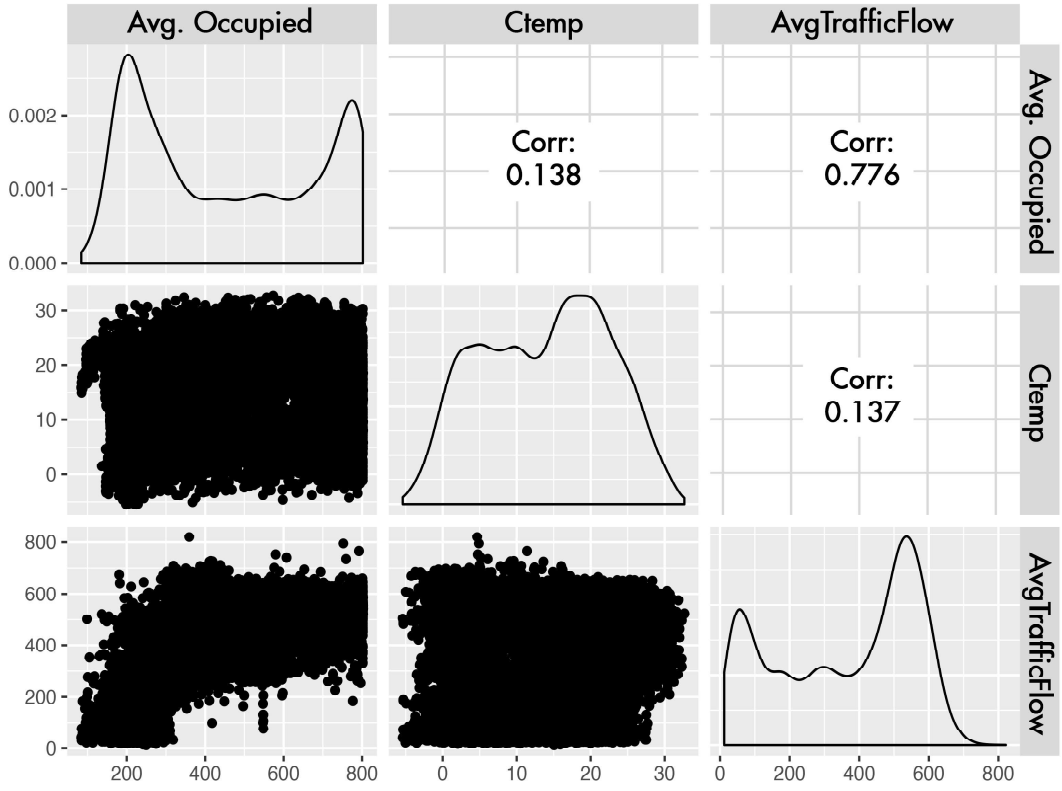
Data Visualization before modelling

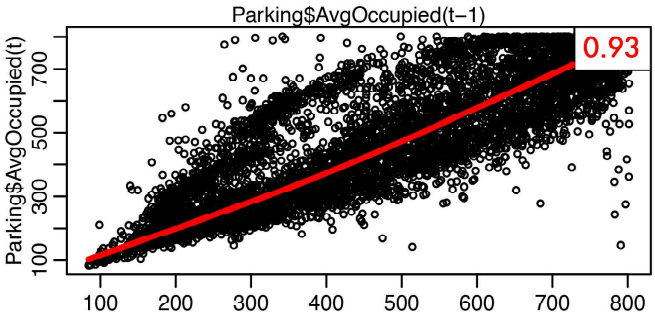
Traffic flow vs Speed on a particular day

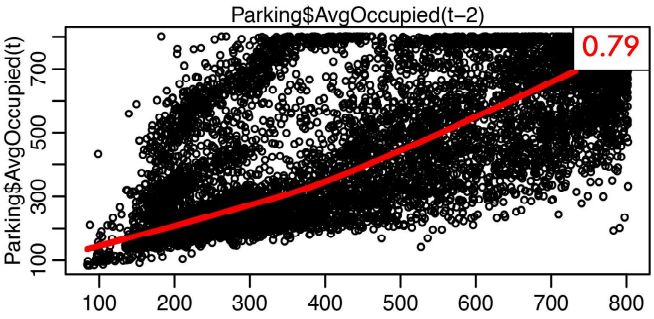
Significant regressors:

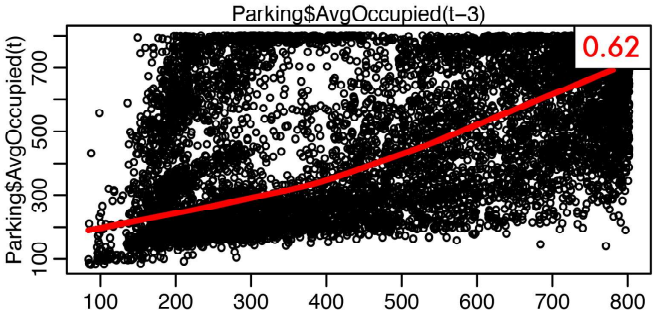
- Average traffic flow
- Hour of the day (0 through 23)*,
- Week day number (1 through 6 for Monday to Saturday and 0 for Sunday)*,
- Five lags of the number of occupied parking spaces,
- Public and school holiday binary indicators*,

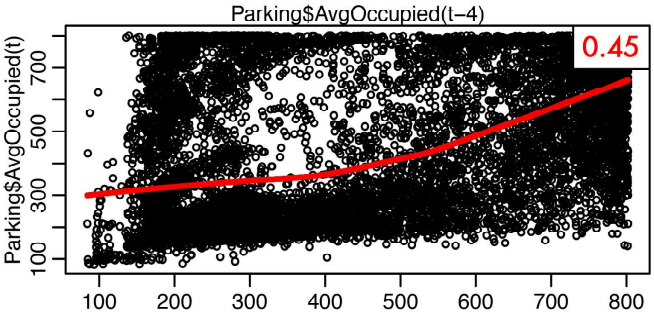
For continuous candidate regressors:

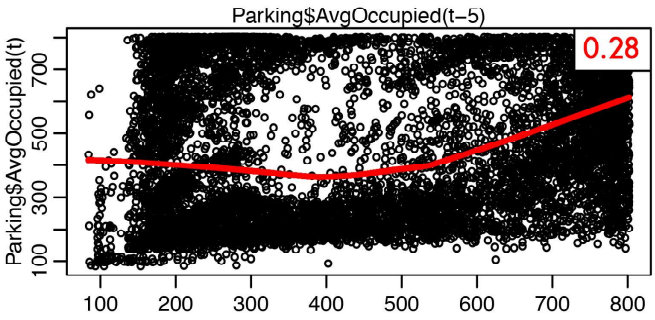



Parking$AvgOccupied(t−1)
0.93
Parking$AvgOccupied(t)
100
300
500
700
100
200
300
400
500
600
700
800


Parking$AvgOccupied(t−2)
0.79
Parking$AvgOccupied(t)
100
300
500
700
100
200
300
400
500
600
700
800


Parking$AvgOccupied(t−3)
0.62
Parking$AvgOccupied(t)
100
300
500
700
100
200
300
400
500
600
700
800

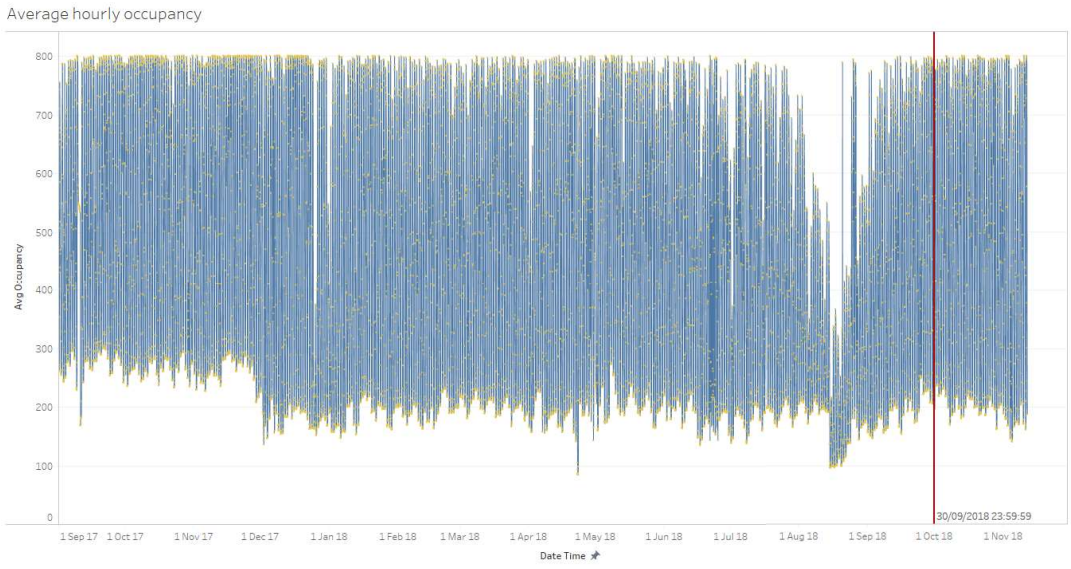

Parking$AvgOccupied(t−4)
0.45
Parking$AvgOccupied(t)
100
300
500
700
100
200
300
400
500
600
700
800


Parking$AvgOccupied(t−5)
0.28
Parking$AvgOccupied(t)
100
300
500
700
100
200
300
400
500
600
700
800

Model Estimation : Settings

The Dataset

- Training set : from 1^{st} September 2017 till 30^{th} September 2018 (9480 data points)
- Test set : from 1^{st} October 2018 to 11^{th} November 2018 (1008 data points)



Model Estimation : Settings

Performance Criteria

- Mean Absolute Percentage Error (MAPE)

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{yi} \right|$$

- Root Mean Square Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Model Estimation

Model I : Multiple Linear regression

Before model estimation, need for transformation of data is examined. For this purpose the Box-Cox transformation was employed with $\lambda = 0.3$

Model Estimation

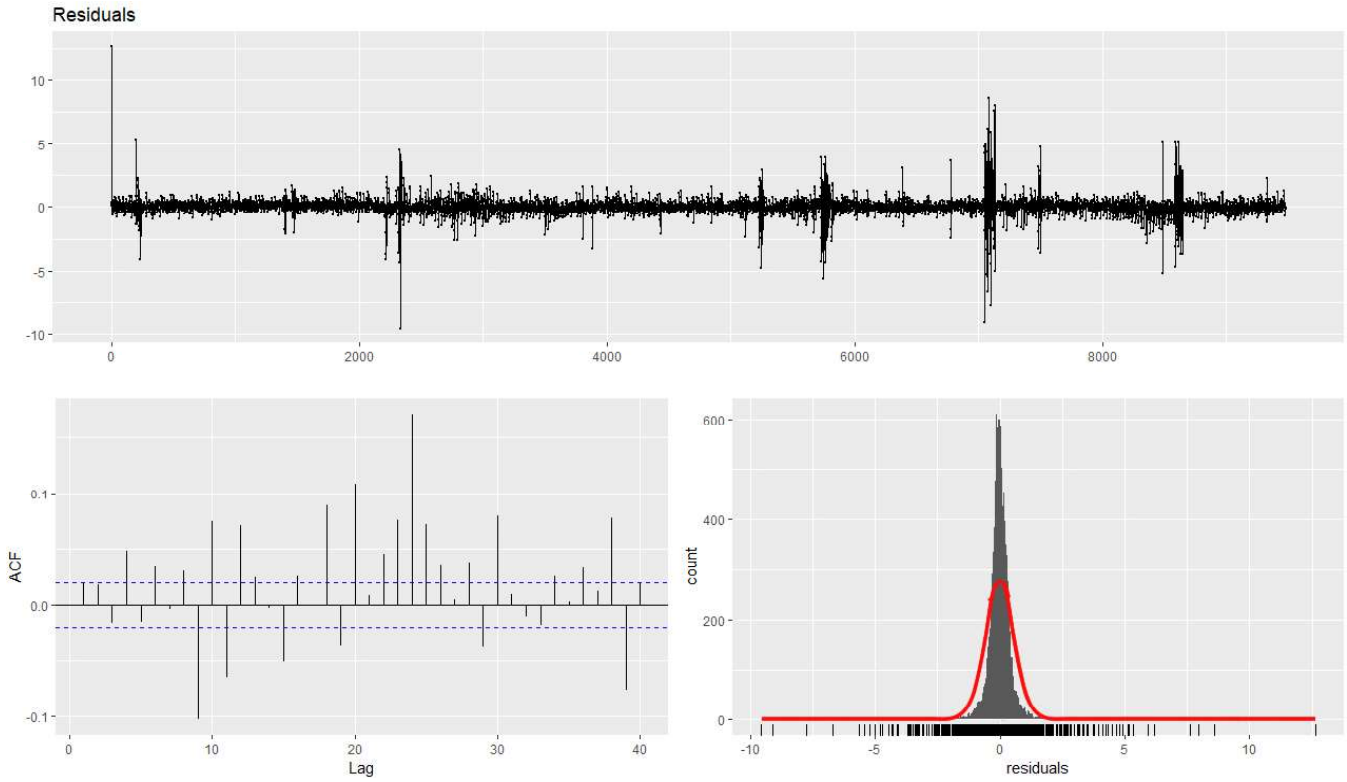
Model I : Multiple Linear regression

The model that yielded the highest $R^2 = 0.967$ along with the highest prediction accuracy, RMSE=0.4101 and MAPE=1.74 is:

$$Y_t = I_{(\text{Hour Indicator * Week Day Indicator})} + Y_{t-1} + Y_{t-2} + \\ + I_{(\text{School Holidays Indicator})} + \text{Traffic Flow}$$

Where Y_t denotes the transformed parking occupancy at time t

Residual Diagnostics:



H_0 : There is no serial correlation of any order p (where p is the order of the lags of the dependant variable used as independent variables)

```
Breusch-Godfrey test for serial correlation
of order up to 174
data:  Residuals
LM test= 1344.6, df= 174, p-value< 2.2e-16
```

At $\alpha = 0.05$, we reject the H_0, concluding that the residuals violate the assumptions of the linear regression model, thus rendering the model untrustworthy for accurate prediction.

Model Estimation

Model II : Seasonal ARIMA

- Classical regression is more often than not inefficient in explaining all of the interesting dynamics of time series.
- While SARIMA$(p, d, q)(P, D, Q)_m$ where:
  - p : Non seasonal AR order
  - d : Non seasonal difference
  - q : Non seasonal MA order
  - P : Seasonal AR order
  - D : Seasonal difference
  - Q : Seasonal MA order
  - m : frequency

  captures the correlation due to lagged linear relationships.

Model Estimation

Model II : Seasonal ARIMA

Two methods were used in order to check which one yielded a better SARIMA$(p, d, q)(P, D, Q)_m$:

- Box and Jenkins method
- Hyndman-Khandakar algorithm [2]

Both of the above are systematic methods of identifying, fitting, checking, and using ARIMA time series models, however they differ in the implementation. The *auto.arima()* function in *R* software uses the second method.

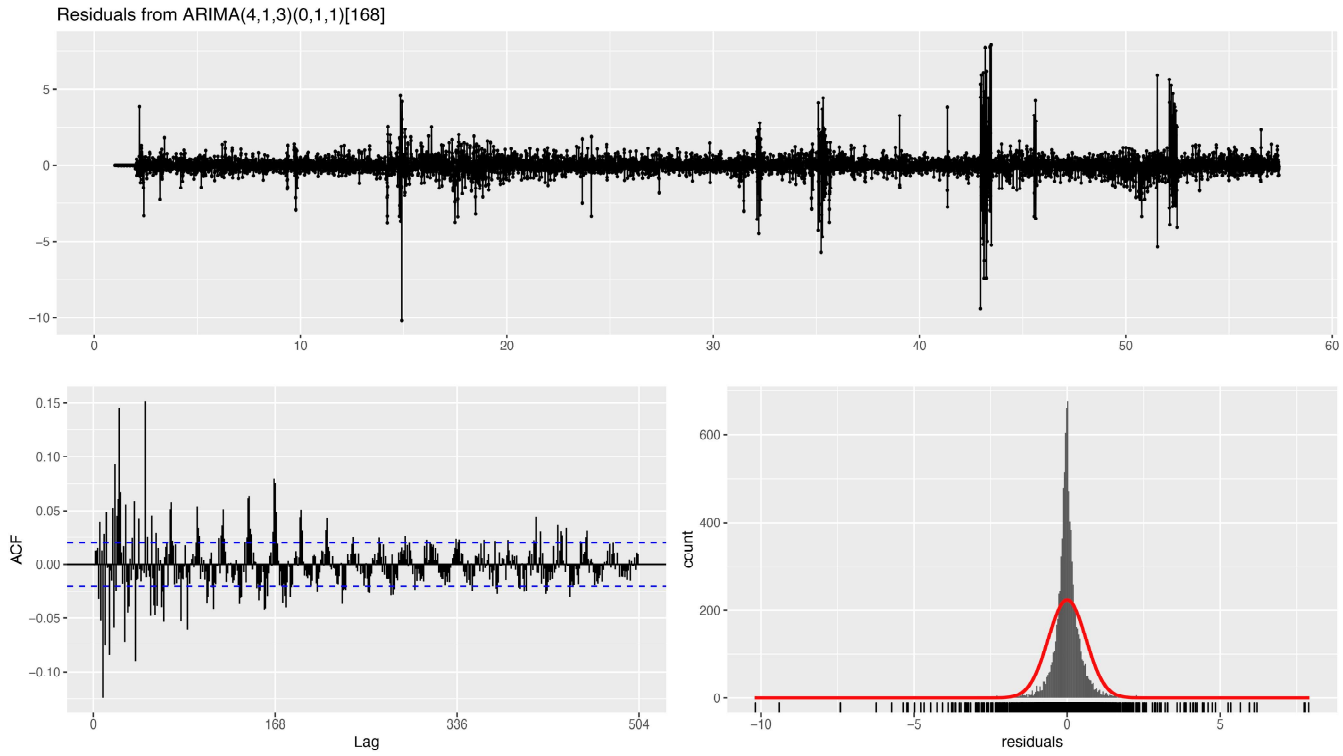
Model by Box and Jenkins approach $\mathrm{ARIMA}(4,1,3)(0,1,1)_{168}$:

$$(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3-\phi_4 B^4)(1-B)(1-B^{168})y_t \\ =(1+\theta_1 B+\theta_2 B^2+\theta_3 B^3)(1-\Theta_1 B)\epsilon_t$$

By Hyndman-Khandakar approach $\mathrm{ARIMA}(4,0,3)(0,1,0)_{168}$:

$$(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3-\phi_4 B^4)(1-B^{168})y_t = \\ (1+\theta_1 B+\theta_2 B^2+\theta_3 B^3)\epsilon_t$$

Residual Diagnostics:



Model II : Seasonal ARIMA

H_0 : Data are independently distributed (i.e., absence of serial auto-correlation)

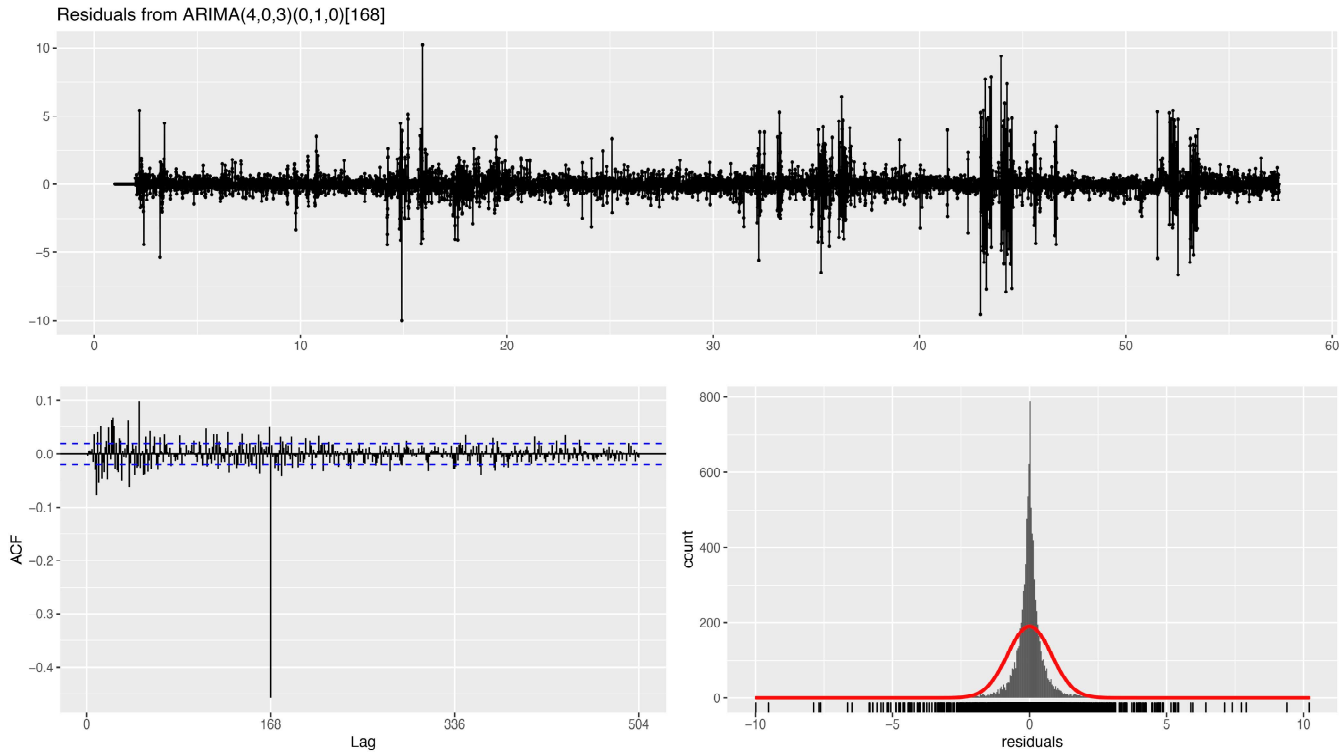
```
Ljung-Box test

data:  Residuals from ARIMA(4,1,3)(0,1,1)[168]
Q* = 2566.2, df = 328, p-value < 2.2e-16

Model df: 8.   Total lags used: 336
```

At $\alpha = 0.05$, we reject the H_0, thus indicating that the model does not conform to the assumption of the white noise residuals. However, if we proceed to forecast with this model, the performance metrics are MAPE=11.55 and RMSE=72.35, but, we cannot trust this model to give accurate predictions.

Model Estimation

Model II : Seasonal ARIMA

Residual Diagnostics:



Model Estimation

Model II : Seasonal ARIMA

H_0 : Data are independently distributed (i.e., absence of serial auto-correlation)

```
Ljung-Box test

data:  Residuals from ARIMA(4,0,3)(0,1,0)[168]
Q* = 3460.3, df = 329, p-value < 2.2e-16

Model df: 7.   Total lags used: 336
```

At $\alpha = 0.05$, we reject the H_0, implying that the model does not conform to the assumption of the *WN* residuals. However, if we proceed to forecast with this model, the performance metrics are MAPE=14.03 and RMSE=75.58, but, we cannot trust this model to give accurate predictions.

Model III : Artificial Feed Forward Neural Networks

- Real life time series are noisy, non-stationary, uncertain and irregular
- Fitting the ARIMA class of models becomes an uphill task:
  - A priori knowledge of form of relationship
  - Problem of multiple seasonalities
  - Assumptions on residuals

Model III : Artificial Feed Forward Neural Networks

ANNs are universal approximators that can approximate any computable function without a-priori assumptions about the data. ANNs are:

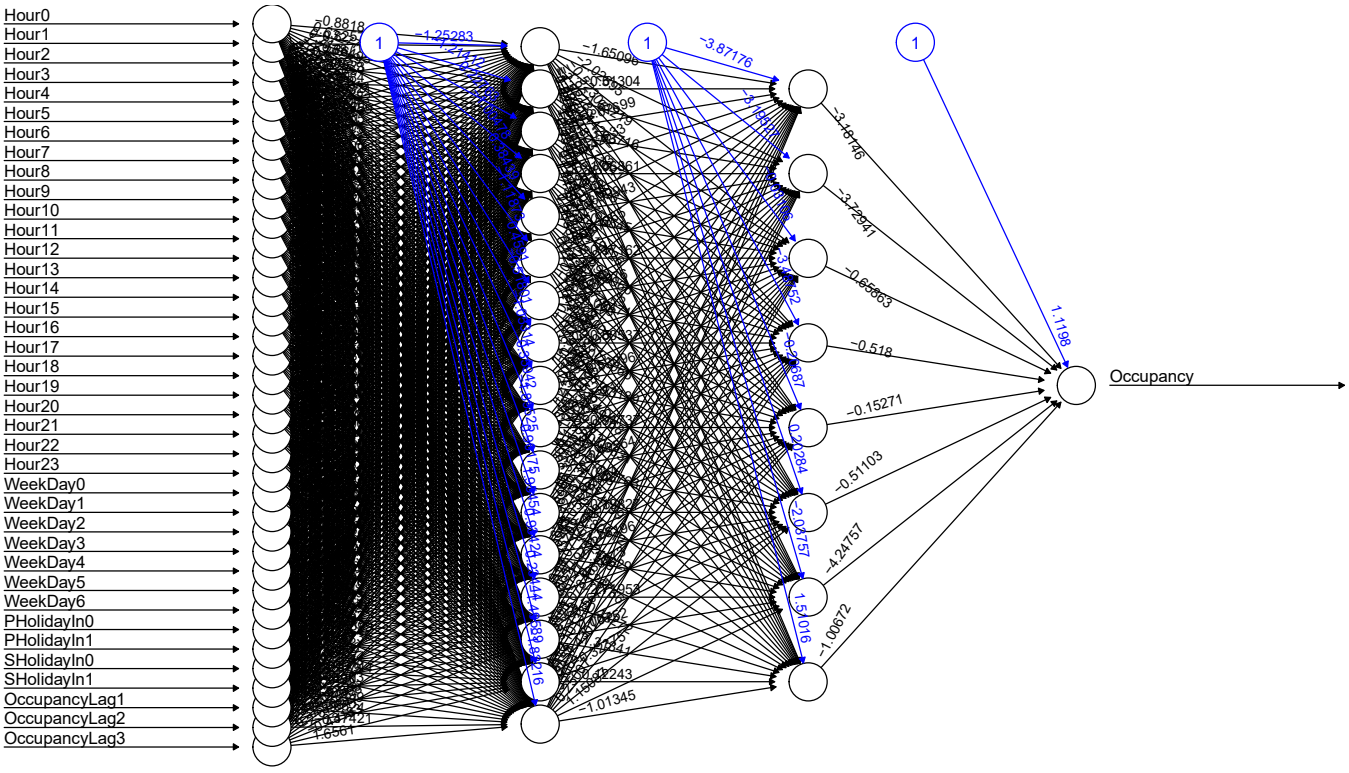
- Non parametric
- Assumption free
- Noise tolerant
- Adaptive

Model III : Artificial Feed Forward Neural Networks

Setting of ANN yielding MAPE=5.08 and RMSE=29.97

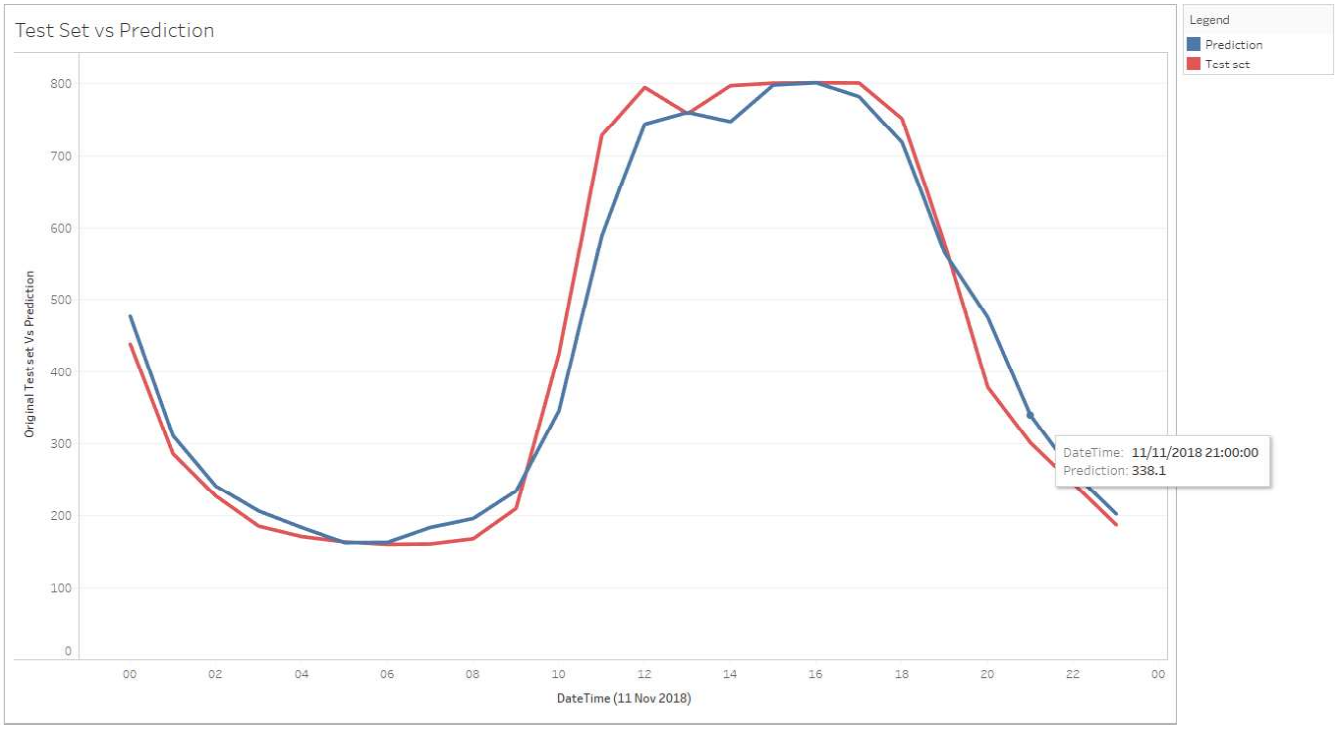
- Activation function : sigmoid
- Number of Hidden layers : 2 (with 17 & 8 neurons respectively)
- Cost function : Sum of squares of errors (SSE)
- Number of repetitions (or epochs) for the NN's training : 5
- Learning rate : 0.0001
- Threshold for partial derivatives of the error function as stopping criteria : 0.05

Model III : Artificial Feed Forward Neural Networks



Model III : Artificial Feed Forward Neural Networks

Prediction vs Test set for the last 24 hours



Model III : Artificial Feed Forward Neural Networks

Jon von Neumann, a Hungarian-American mathematician, physicist, computer scientist and inventor said that:

"*With four parameters I can fit an elephant, and with five I can make him wiggle his trunk*"

Simply put, with enough parameters one can fit any data set. In ANNs, we have to determine the hyper-parameters and the parameters which can practically be hundreds of millions (the weights and biases of the artificial synapses, depending on the size of the ANN).

Model Estimation

Model IV : Support Vector Regression

- Supervised machine learning algorithm, developed by Vladimir Vapnik and Corinna Coates in 1995
- Can be used for both classification and regression problems
- Is an approach to improve the generalization properties of neural networks
- Originally SVMs were developed for pattern recognition problems and recently have been extended to solve non-linear regression problems
- It relies on the kernel trick (avoids the explicit mapping that is needed to get linear learning algorithms to learn a nonlinear function or decision boundary)

Model IV : Support Vector Regression

ANN	SVM
Extensive experimentation preceding theory	Sound theory first followed by implementation and experimentation
Can suffer from multiple local minima	Solution is global and unique
Computational complexity depends on dimensionality of input space	Not the case
Prone to overfitting	Uses regularization to combat overfitting

Model Estimation

Model IV : Support Vector Regression

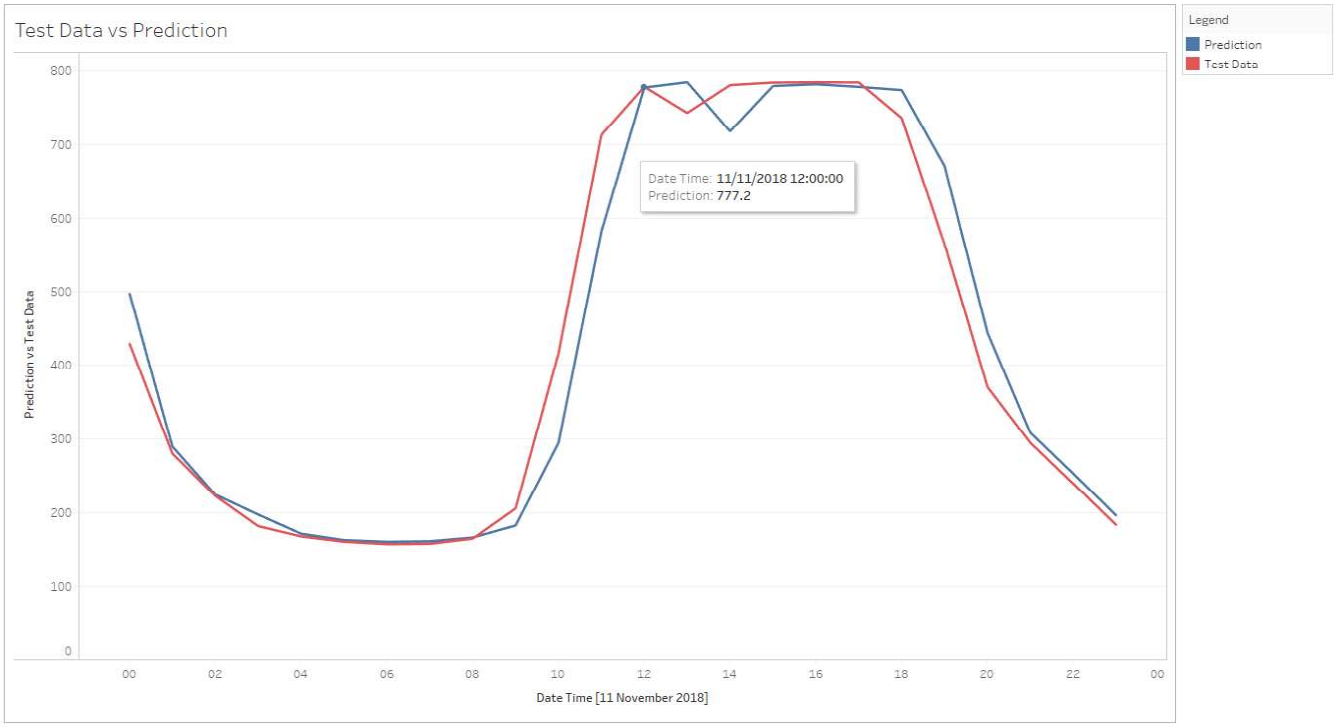
After performing grid search, the below setting was used, yielding MAPE=6.05 and RMSE=41.06

- Non linear kernel used : Radial Basis Function (RBF)
- γ for RBF kernel : 0.00006
- Regularization parameter or Cost (C) : 100
- Tolerance parameter (ϵ): 0.001

Model Estimation

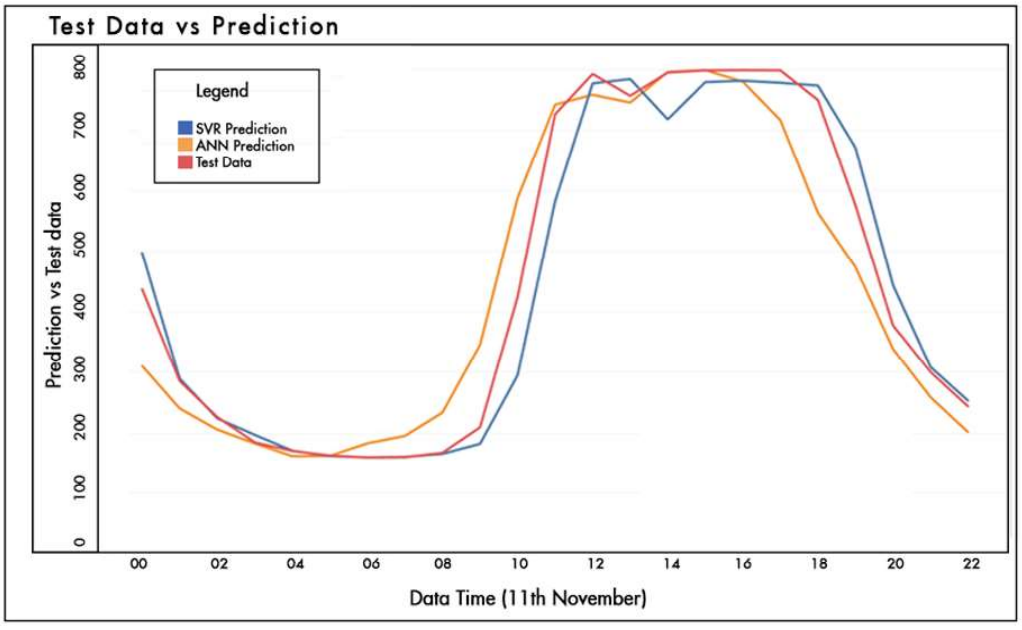
Model IV : Support Vector Regression

Prediction vs Test set for the last 24 hours



Conclusion

MODEL	RMSE	MAPE
Multiple linear regression	0.41	1.74
SARIMA	Model 1: 72.35 Model 2: 75.58	Model 1: 11.55 Model 2: 14.03
Artifical Feed Forward Neural Networks	29.97	5.08
Support Vector Regression with RBF Kernel	41.06	6.05



Conclusion and further work

Further Work

- Explore TBATS and Dynamic harmonic regression models to deal with multiple seasonalities
- Experiment further with ANNs by adjusting the hyper parameters
- Explore LSTMs (Recurrent NNs)
- Experiment with SVR by using different kernels

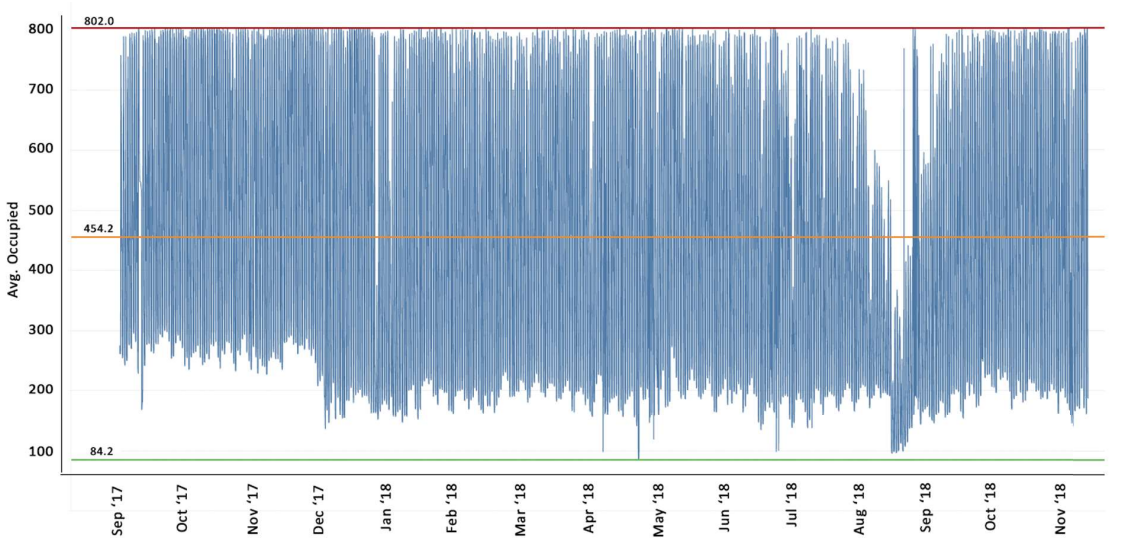
Thank you for your time and attention!

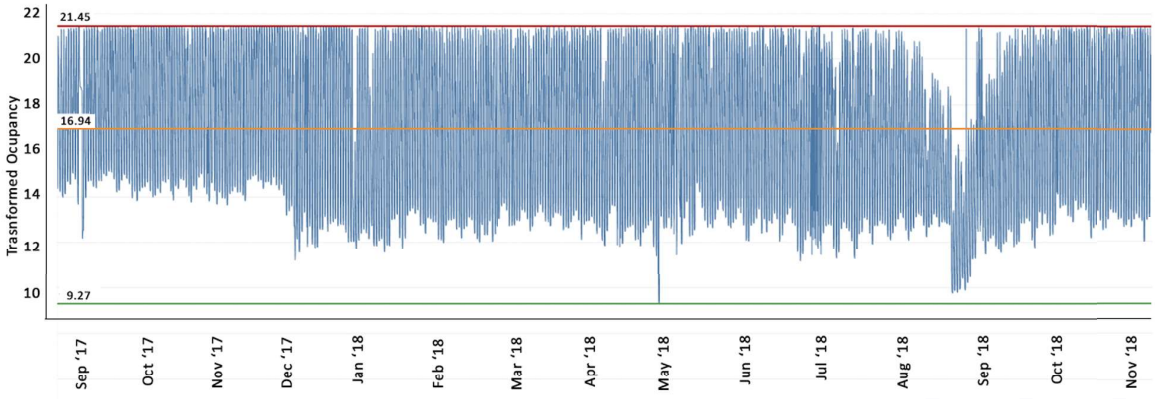
References

Graham Cookson and Bob Pishue, *The impact of parking pain in the US, UK and Germany*, https://sevic-emobility.com/images/news/INRIX_2017_Parking_Pain_Research_EN-web.pdf

Hyndman, R. J., and Khandakar, Y. *(2008). Automatic time series forecasting: The forecast package for R. Journal of Statistical Software, 27(1), 1–22.* https://doi.org/10.18637/jss.v027.i03

Box Cox Transformation





ACF for frequency