Forecasting mobile network traffic

Report

By Sudharshan PJ

1 Hardware and Software Setup

1.1 Hardware:

The exploratory data analysis and predictions were carried out on: Intel Core i7-8750H with Memory: 16GB, Samsung 850 EVO SSD for loading the data.

The MapReduce [1] jobs were run on Dell PowerEdge R430 with 96GB of memory, 6TB of SATA disk and two 10-core Xeon E5-2650 v3 2.3GHz CPUs. More information about this can be found in this link.

Note: No dependencies on GPUs

1.2 Software:

The codes are written in Python 3. The other library requirements are mentioned in *requirements.txt* file. Install the requirements by running the following:

```
pip install -r requirements.txt
```

For the task I, an alternative solution has been proposed which has a dependency on Apache Pig [2][1], which in turn requires Hadoop Map-Reduce to be installed in fully-distributed mode.

1.3 Steps to run the code:

The file *eda.py* contains the code for the Exploratory Data Analysis and this can also be visualized in Jupyter Notebook file *eda.ipynb*. Similarly, the file *prediction.py* contains the code for forecasting and this can also be visualized in Jupyter Notebook file *prediction.ipynb*. These can be executed as below:

```
python eda.py
python prediction.py
```

The above codes does not require any Command Line Arguments.

The file *finalsum.pig* contains the Pig commands to process the Big Data and it can be executed as below:

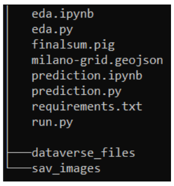
```
pig -x mapreduce finalsum.pig
```

2 Short Description of the Dataset

The Telecommunication Activity dataset [3] is a part of Telecom Italia Big Data Challenge which is an aggregation of telecommunications, weather, news, social networks and electricity data from the city of Milan and the Province of Trentino. This dataset has been released to the research teams under the Open Database License (ODbL) and is maintained by Harvard Dataverse.
In this exploration, we mainly focus on the telecommunication data of Milan city. The telecommunication data for the city is available as .txt files with tab-delimited values (TSV). There are totally 62 files consisting of Call Detail Record (CDR) collected from Nov 1, 2013 to Jan 1, 2014, one file for each day. More details on the dataset can be found on the official website.

The city of Milan's spatial distribution is aggregated in a grid with 10,000 square cells. These 10,000 squares has a size of about 235×235 meters. The Grid dataset [4] provides the geographical reference for each square. Hence, the location for the gridID provided in the Activity dataset can be deciphered from the grid dataset. Below figure shows a tree which provides a view of how the dataset is to be stored in the execution directory:

[1] Apache Pig is a high-level primitive which translates Pig-Latin code to optimized sequence of Map-Reduce jobs



```
│       eda.ipynb
│       eda.py
│       finalsum.pig
│       milano-grid.geojson
│       prediction.ipynb
│       prediction.py
│       requirements.txt
│       run.py
│
├───dataverse_files
└───sav_images
```

Figure 1: Directory Tree

3 Data Wrangling

3.1 Method I (Python)

- The file is read into a dataframe and then daily, hourly data were extracted by grouping.
- The *timeInterval* column was converted to Milan's local timezone for easy processing of data and stored as a new column.
- *countryCode* column and *redundant* columns timeInterval were then dropped as they are not required for the given tasks.

The above operations were carried-out on each file and then combined, which results in a faster and efficient data loading as compared to combining the files and then processing them. The Two-month traffic data in each geological location was obtained through the above process.
Time taken to obtain the sum of the traffic data over the period of two-months using this method: ∼15 minutes

3.2 Method II (Apache Pig)

For the task of combining and summing the traffic data for each geological location over the period of two-months, Pig Latin (SQL-like) queries were written to obtain the data. This method is much faster than the previous one, as it parallelize the tasks through several Map, Combine and Reduce tasks at the same time.
Time taken to obtain the sum of the traffic data over the period of two-months using this method: ∼1.2 minutes.
But, it requires the presence of dedicated multiple nodes for MapReduce jobs which may not be always possible as it requires decent amount of computation power along with a cluster setting.
So, for the following tasks, we will stick to the fully Python-based implementation.

4 Task I

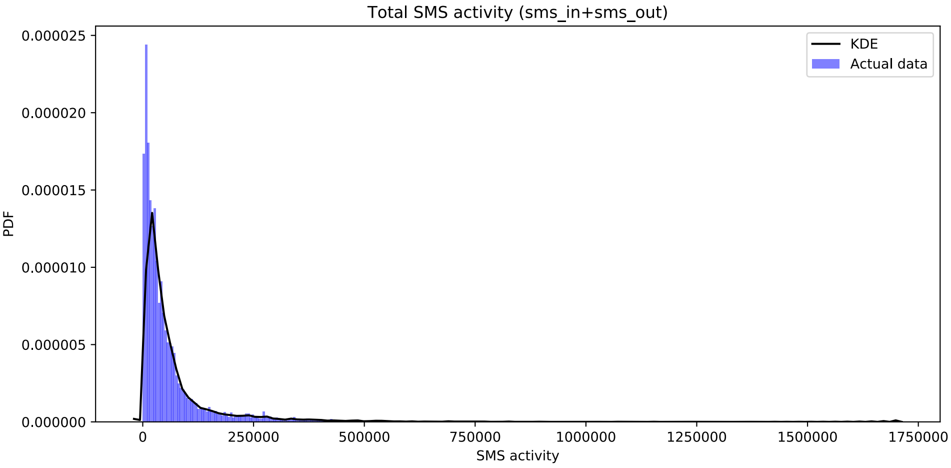


Figure 2: Heatmap for Total SMS activity (includes incoming and outgoing SMS)

Figure 2 depicts the Probability density function (PDF) for Total SMS activity which includes incoming and outgoing SMS for 10,000 geographical locations over the period of two months. We can see that most of the values lie between 0 and 120,000 but the maximum SMS activity extended up to 1,750,000. Only a few regions of Milan have very high SMS activity. It is interesting to note that the SMS activity is mildly diverse and is densely populated in the lower range.

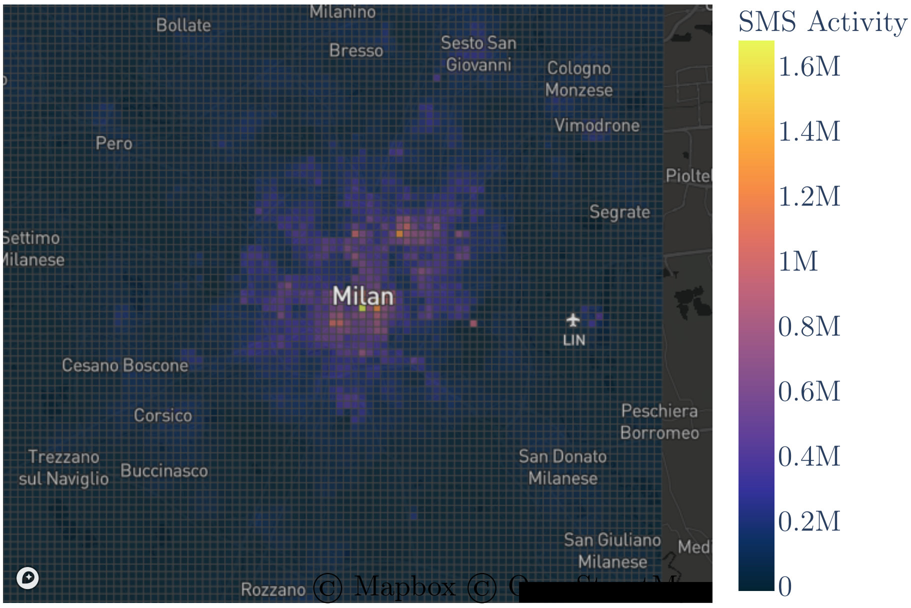


Figure 3: Probability density function for Total SMS activity (includes incoming and outgoing SMS)

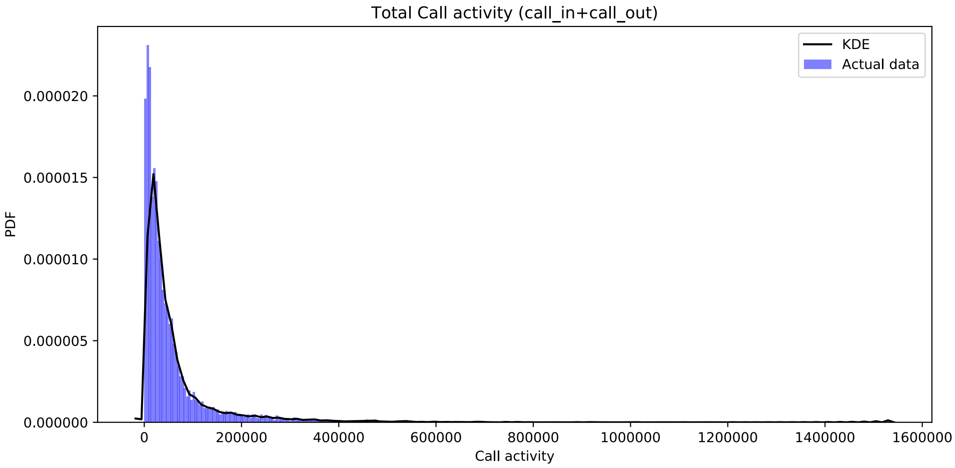


Figure 4: Probability density function for Total Call activity (includes incoming and outgoing Call)

Figure 4 shows the PDF for Total call activity in 10,000 regions. We can see that the plot is almost similar to Figure 2 and most of the traffic data lie between 0 and 100,000. Similar to SMS activity, only a few regions of Milan have very high Call activity and the Call activity is mildly diverse and is densely populated in the lower range.

Figure 6 depicts PDF for the internet activity and it can be seen that it is quite similar to that of the Call and SMS activity. This might be due to the fact that the grids with the higher probability density for SMS, Call and Internet activity are either densely populated or likely a larger gathering spot viz., touristic places, work places, recreational centres, etc. From the figure 3, 5 and 7, it is comprehensible that the most of the dense grids are situated in the metropolitan area of Milan.

Figure 8 shows top 10 gridIDs for the total SMS activity over the two-month period. Figure 9 shows their location on the map. We can see that most of the grids are situated in the metropolitan area of Milan as expected.

Figure 10 shows top 10 gridIDs for the total Call activity over the two-month period. Figure 11 shows their location on the map. We can see that most of the grids are situated in the metropolitan area of Milan as expected. Most of the regions are overlapping the top 10 gridIDs for SMS activity.

Figure 12 shows top 10 gridIDs for the total Internet activity over the two-month period. Figure 13 shows their location on the map. We can see that most of the grids are situated in the metropolitan area of Milan as expected. Most of the regions are overlapping the top 10 gridIDs for SMS activity as well as the top 10 gridIDs for Call activity.

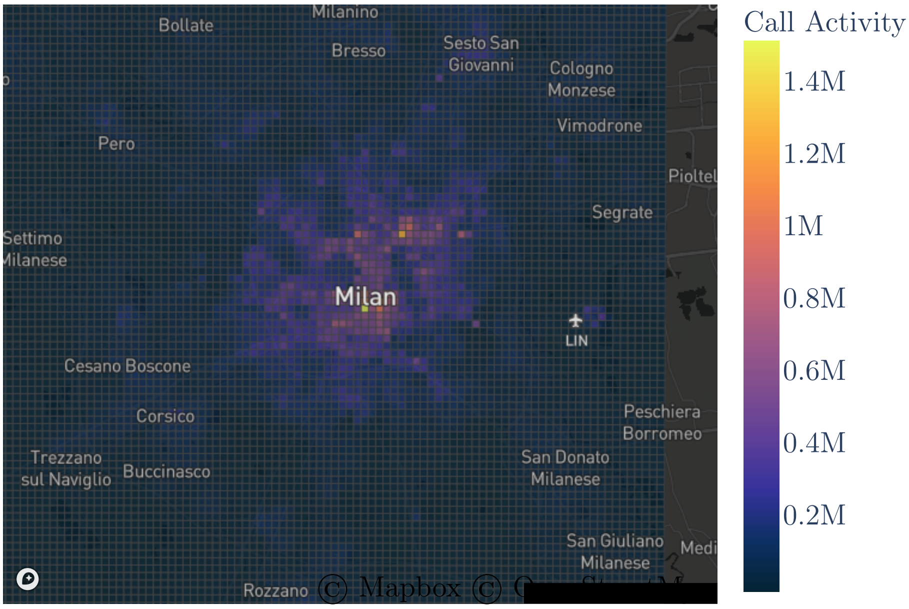


Figure 5: Heatmap for Total Call activity (includes incoming and outgoing Call)

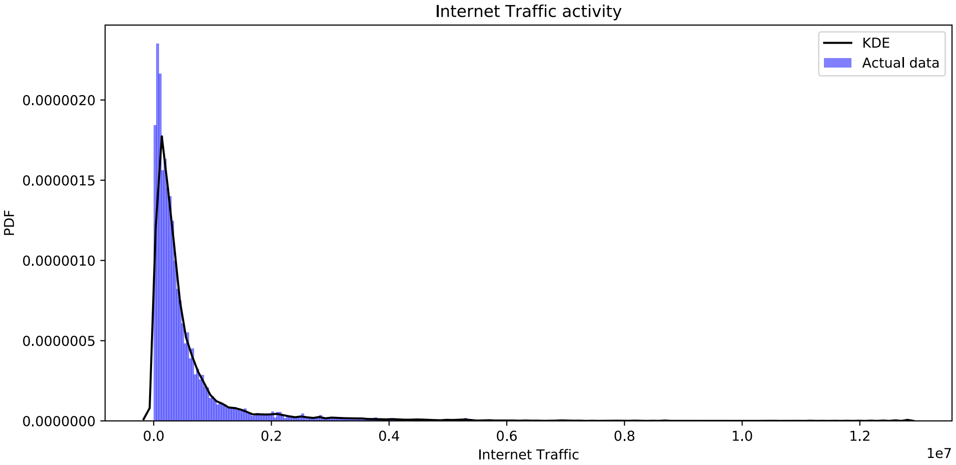


Figure 6: Probability density function for Internet activity

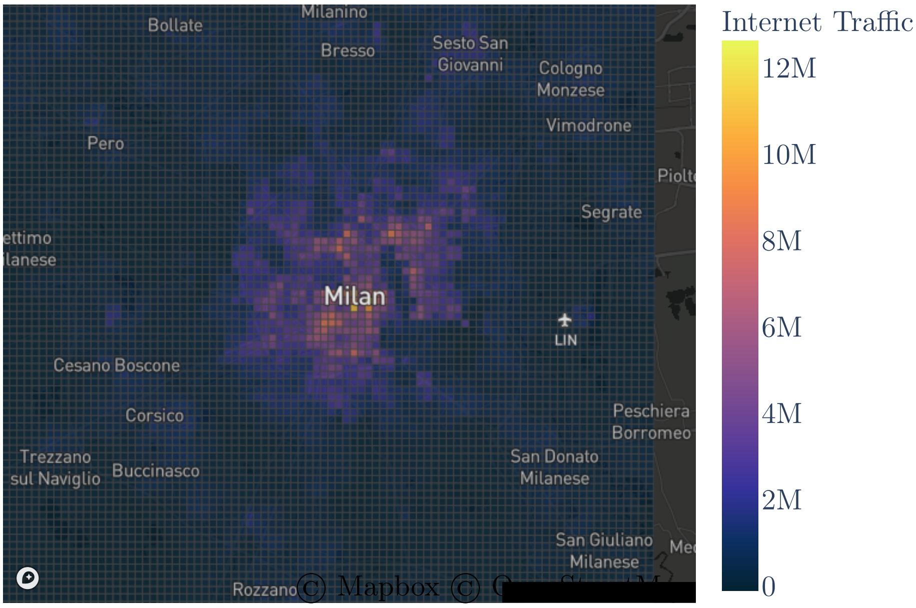


Figure 7: Heatmap for Internet activity

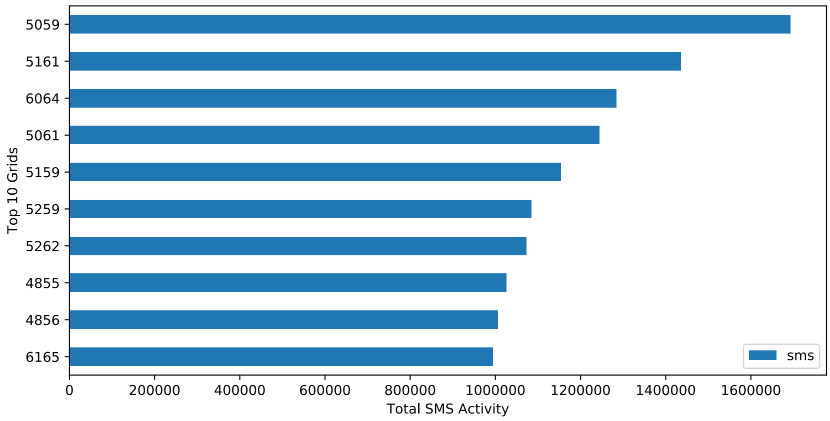


Figure 8: Top 10 grids for SMS activity (includes incoming and outgoing SMS)

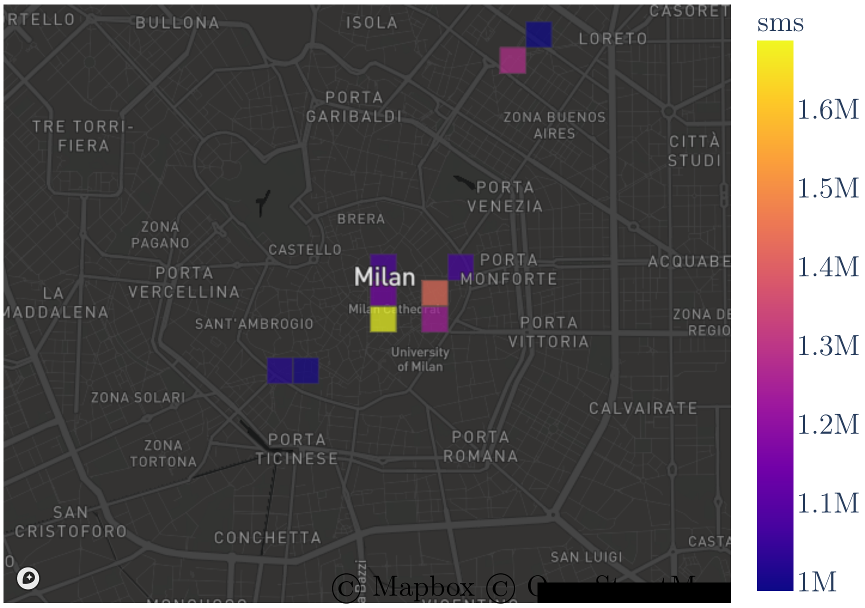


Figure 9: Location of Top 10 grids for SMS activity (includes incoming and outgoing SMS)

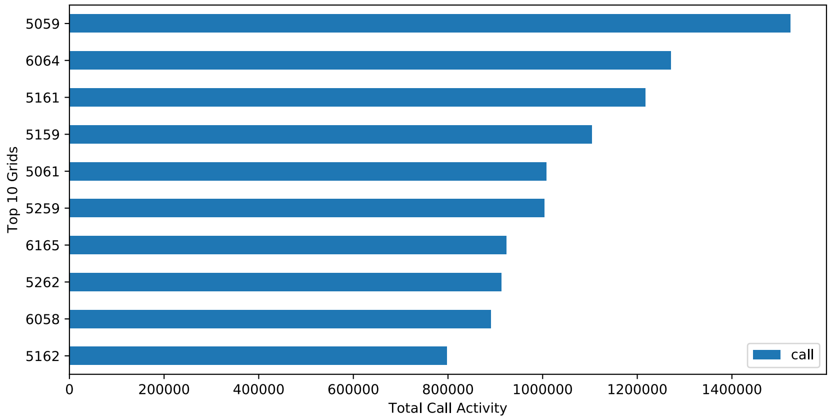


Figure 10: Top 10 grids for Total Call activity (includes incoming and outgoing Call)

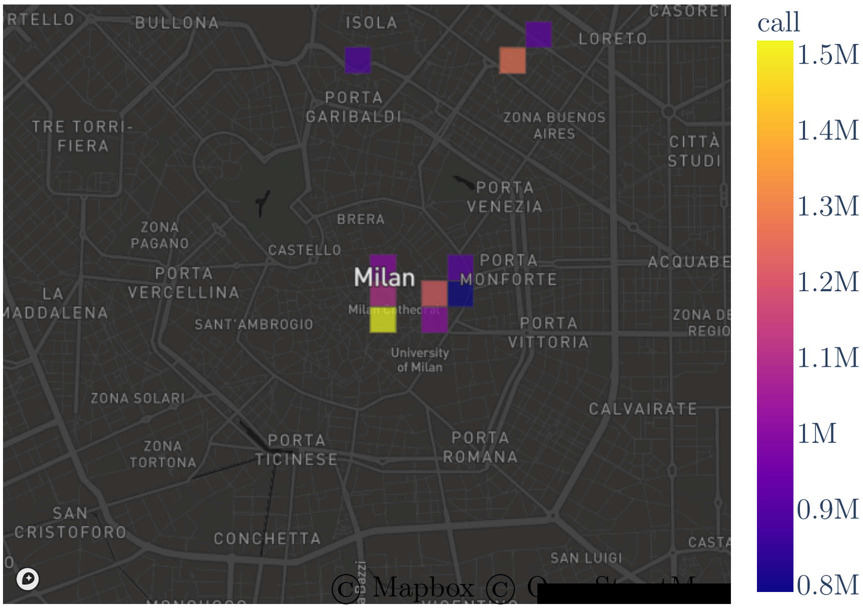


Figure 11: Location of Top 10 grids for Call activity (includes incoming and outgoing Call)

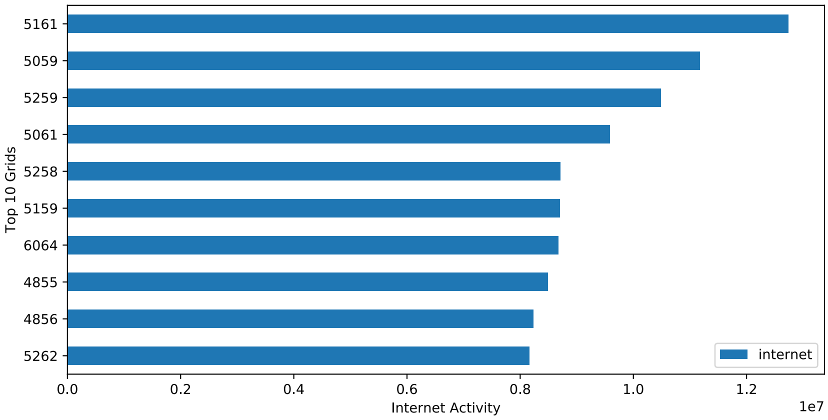


Figure 12: Top 10 grids for Total Internet activity

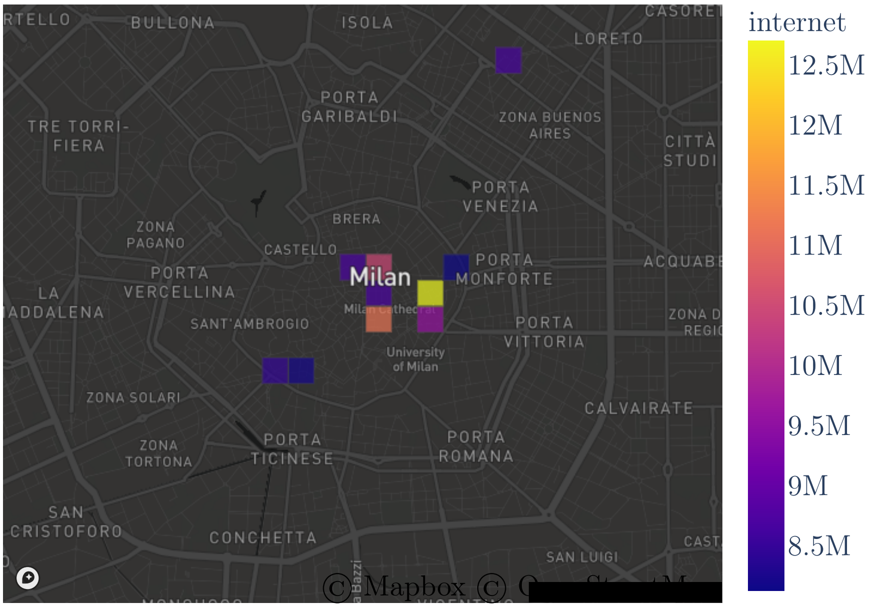


Figure 13: Location of Top 10 grids for Internet activity

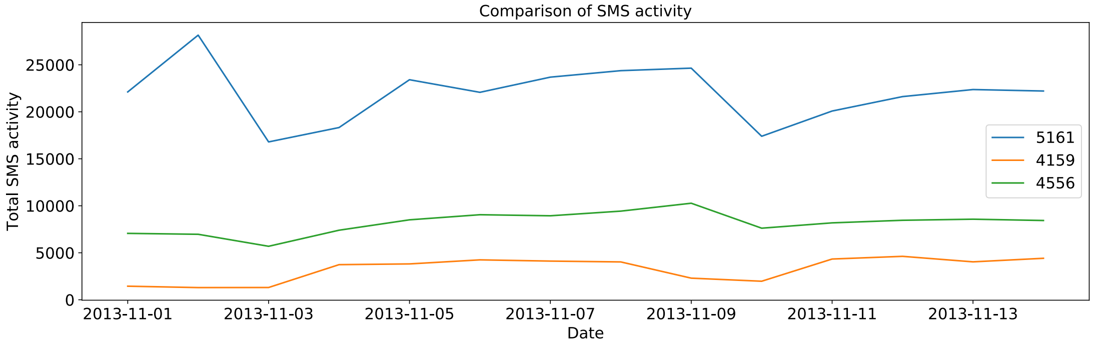


Figure 14: Time series of Total SMS activity

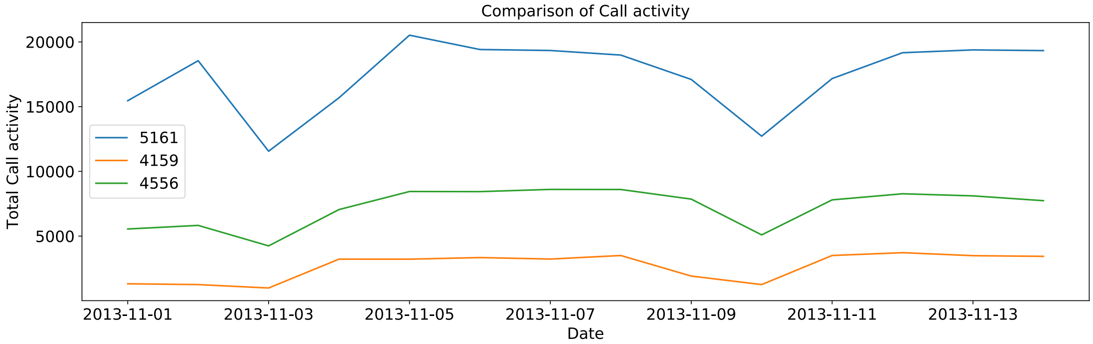


Figure 15: Time series of Total Call activity

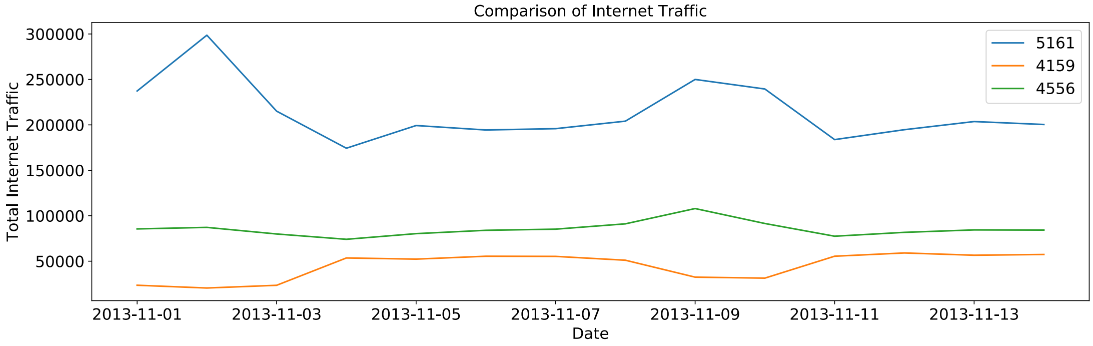


Figure 16: Time series of Internet activity

Figure 14 shows the time series of the total SMS activity over the first two weeks of November 2013 for the grids 5161 (which has the highest total Internet traffic as seen in figure 12), 4159 and 4556. Throughout this period, Grid 5161 has a higher amount of SMS activities followed by Grid 4556 and then Grid 4159.

We can see that for the grids 4159 and 4556, the total SMS activity is fluctuating less and stay quite low during this period compared to the grid 5161. For the grid 4159 that has the lowest total SMS activity for this period, the highest traffic has been reported during the period between 4th November - 8th November and the period starting from 11th November. For the grid 4556, the highest total SMS activity took place on 9th November. The grid 5161 displays another type of variation: top traffic is recorded on 2nd November followed by 9th November; lower traffics are recorded for 3rd November and 10th November. Overall, there is a seasonal behavior of ups (following the pit) and downs (following the peak) in the Grids 5161, 4159 and 4556 for the SMS activity.

Figure 15 shows the time series of the total Call activity over the first two weeks of November 2013 for the same grids as before namely the grids 5161, 4159 and 4556. Throughout this period, similarly to the time series for total SMS activity, Grid 5161 has a higher amount of Call activities followed by Grid 4556 and then Grid 4159.

We can see that for the grids 4159 and 4556, the total Call activity is fluctuating moderately during this period compared to the grid 5161. The time series for three grids follows the same pattern between two pits on 3rd November and 10th November. For the grid 4159 that has the lowest total Call activity for this period, the highest traffic has been reported during the period between 4th November - 8th November and the period starting from 11th November. For the grid 4556, there is a seasonal behavior of ups (following the pits on 3rd November and 10th November) and downs (following the peak period between 3rd November and 10th November). The grid 5161 fairly steeps up from 3rd November and peaks on 5th November

Figure 16 shows the time series of the total Internet activity over the first two weeks of November 2013 for the same grids as before namely the grids 5161, 4159 and 4556. Throughout this period, similarly to the time series for total SMS and Call activity, Grid 5161 has a higher amount of Internet traffic followed by Grid 4556 and then Grid 4159.

We can see that for the Grid 4556, the total Internet traffic is fluctuating very less during this period compared to Grid 4159 which in turn is fluctuating less than Grid 5161. For the grid 4159 that has the lowest total Internet traffic for this period, the highest traffic period has been reported between 4th November - 8th November and the period starting from 11th November. For Grid 4556, no seasonal behaviour can be noticed except for a local increase on 9th November. Grid 5161 has a peak value on 2nd November followed by 9th November and a lower value on 4th November followed by 11th November. For this grid, each peak value is followed by a decrease until it reaches the lower value which in turn is followed by an increase until it reaches the local peak value.

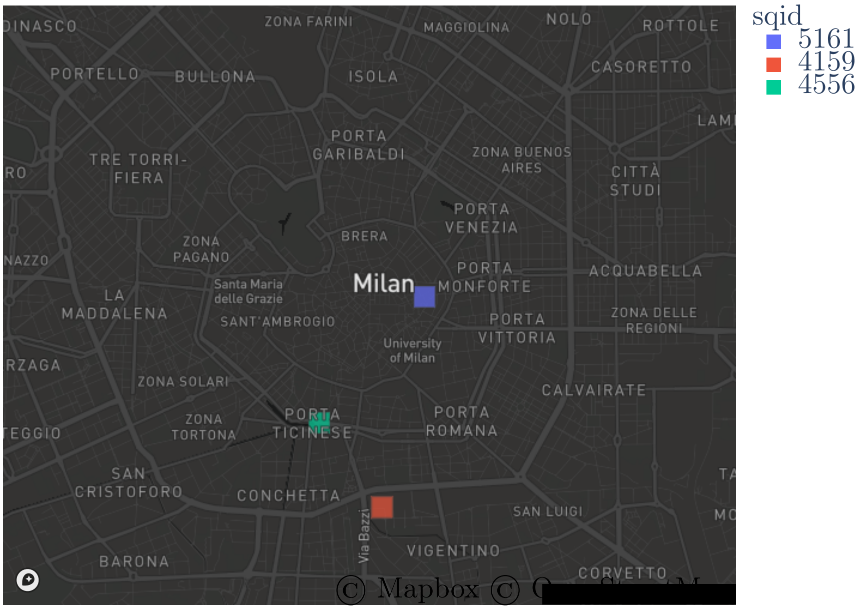


Figure 17: Location of Grid 5161, 4159, 4556

From the figure 17, we can see that the Grid 5161 is located in the centre of Milan city, this explains the reason for high level of SMS, Call and Internet traffic compared to the other two grids. Grid 4159 is further away from the city centre compared to the other two grids and we noticed earlier that it had the least volume of activity for SMS, Call and Internet. From this, we can observe that, as we move away from the centre of the city we tend to find lesser activity. It may not be true in general as there may be a gathering spot further away from the city centre which might result in high degree of activity as seen in figures 3, 5 and 7.

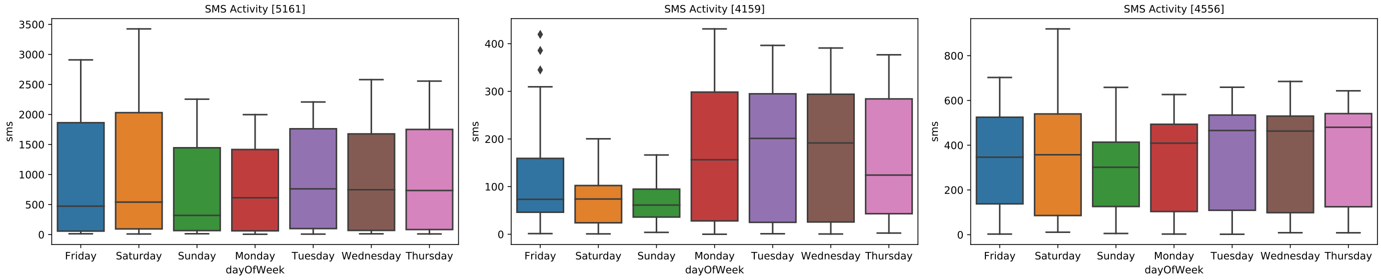


Figure 18: Box plot of SMS Activity for each day of the week

Figure 18 gives the average daily SMS activity while figure 19 gives the average hourly SMS activity for the same grids as before - grids 5161, 4159 and 4556. For Grid 5161, the average daily SMS activity is quite

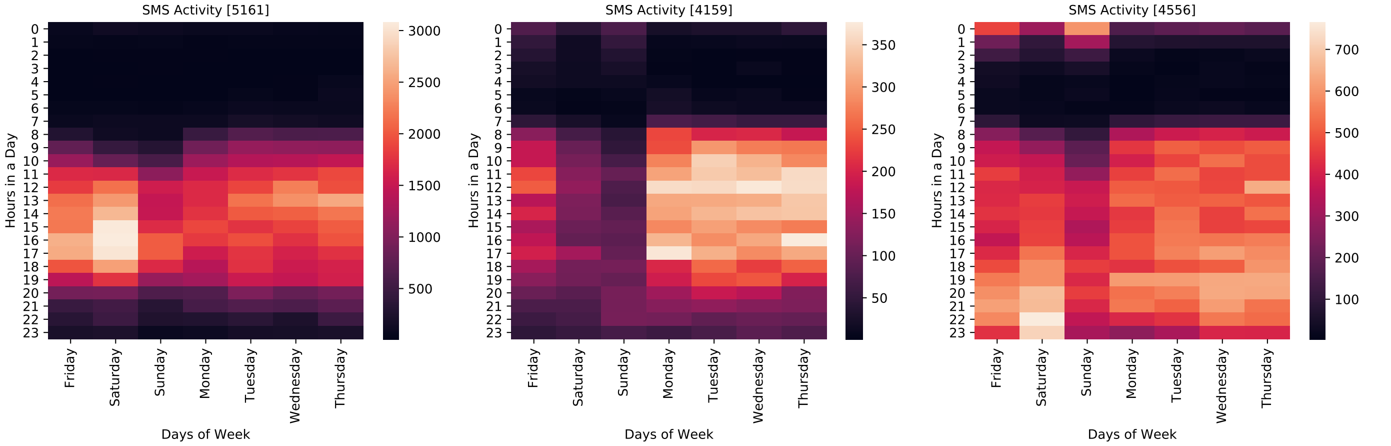


Figure 19: Average Hourly SMS Activity

constant over the whole week. However, there is a maximum of daily SMS activity on Saturdays. In addition, the highest hourly activity has been recorded on Saturdays in the afternoon between 2:30 pm and 5:30 pm. From 9:30 pm to 7:30 am, the SMS traffic is the lowest and this interval is slightly increased during the weekends compared to weekdays. The Grid 4159 shows most activity during the weekdays, especially from Monday to Thursday between 7:30 am to 7:30 pm. The activity in the Grid 4556 is more or less evenly distributed across the all the days. For the grids 4159 and 4556, it is interesting to note that the SMS traffic is extended until early morning on Fridays and Sundays but with a higher proportion for Grid 4556.

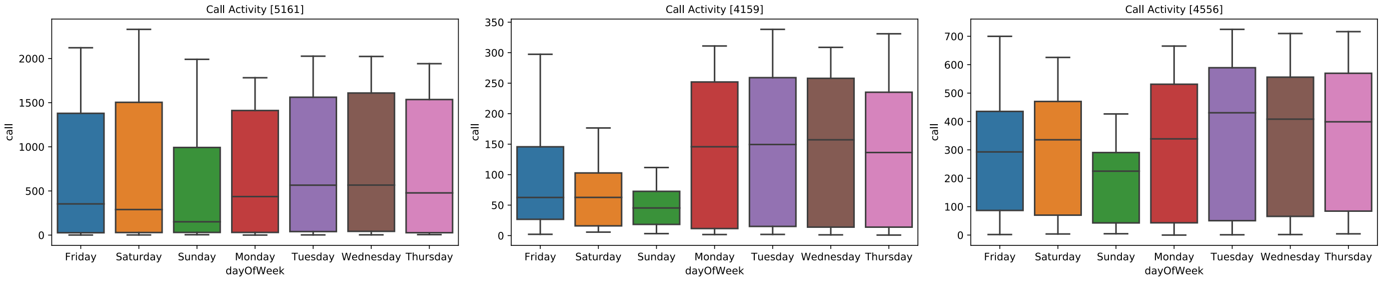


Figure 20: Box plot of Call Activity for each day of the week

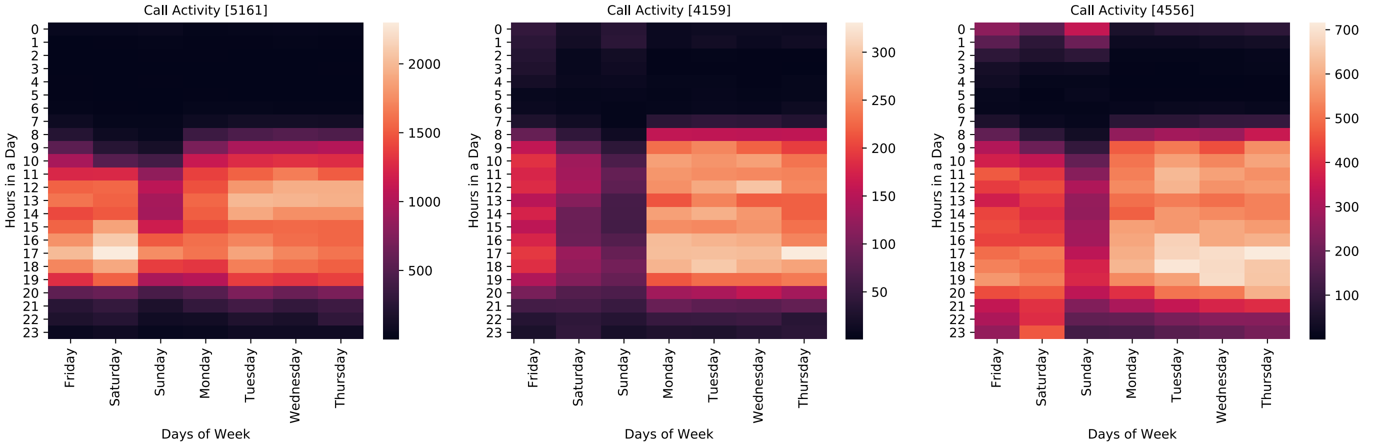


Figure 21: Average Hourly Call Activity

Figure 20 gives the average daily Call activity while figure 21 gives the average hourly Call activity for the grids 5161, 4159 and 4556. For Grid 5161, the average daily call activity is more or less evenly distributed across all the days with a noticeable decrease on Sundays. It showcases similar characteristics as seen for its SMS activity. For this grid, the highest hourly activity has been recorded on Saturdays in the afternoon between 4:30 pm and 5:30 pm. From 10:30 pm to 7:30 am, the hourly Call traffic is the lowest and this interval is slightly increased during the weekends. The Grid 4159 shows the same activity to that of its SMS activity: most of the call activity is concentrated during the weekdays, especially from Monday to Thursday between 8:30 am to 7:30 pm. In addition, there is a noticeable decrease of calls during the weekend especially on Sundays, with hourly call activity close to none. For this grid, from 9:30 pm to 7:30 am, the call traffic is the lowest. The activity in the Grid 4556 is mainly concentrated during the weekdays and Saturdays. For this grid, from 11:30 pm to 7:30 am, the call traffic is the lowest. Additionally, for the grids 4159 and 4556, we can notice that the call traffic is extended until early morning on Fridays and Sundays but with a higher proportion for Grid 4556.

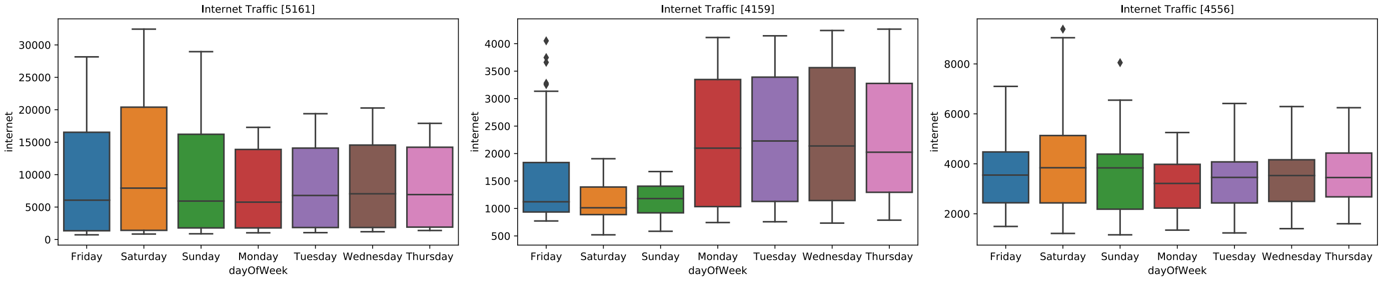


Figure 22: Box plot of Internet Activity for each day of the week

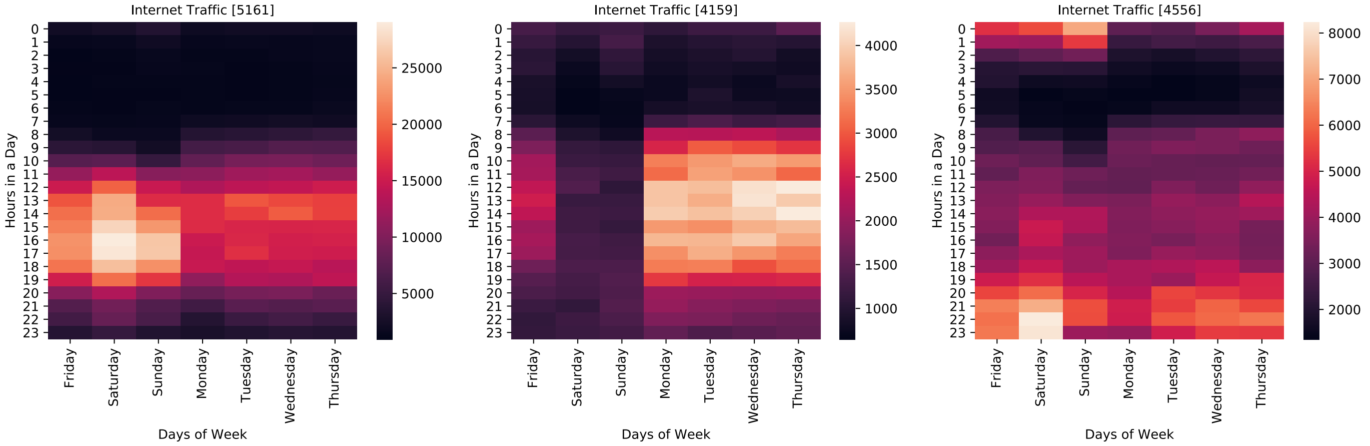


Figure 23: Average Hourly Internet Activity

Figure 22 gives the average daily internet activity while figure 23 gives the average hourly Internet activity for the grids 5161, 4159 and 4556. For Grid 5161, the average daily internet traffic is more or less evenly distributed across the all the weekdays with a noticeable increase from Fridays to Sundays. In addition, the highest hourly activity has been recorded on Saturdays in the afternoon between 3:30 pm and 5:30 pm. For this grid, from 11:30 pm to 7:30 am, the hourly Internet traffic is the lowest and this interval is slightly increased during Sundays. The Grid 4159 shows the same internet activity to that of its SMS and call activity: most of the internet activity is concentrated during the weekdays, especially from Monday to Thursday between 7:30 am to 7:30 pm. However, the internet traffic activity is more extended after 7:30 pm and stays quite high until 11:30 pm. In addition, there is a noticeable decrease of internet traffic during the weekend, with hourly internet activity close to none. For this grid, from 11:30 pm to 7:30 am, the internet traffic is the lowest. The internet activity in the Grid 4556 is more or less evenly distributed across the all the week with a steep increase on Saturdays. The highest hourly activity has been recorded on Saturdays at night between 9:30 pm and 11:30 pm. The highest internet traffic is mainly concentrated between 7:30 pm and 11:30 pm, with the exception of Mondays. For this grid, from 2:30 pm to 7:30 am, the internet traffic is the lowest. Additionally, for the grids 4159 and 4556, we can notice that the internet traffic is extended until early morning but with a higher proportion for Grid 4556 especially for Fridays, Saturdays and Sundays.

From the above analysis, Grid 5161 is more likely to be a touristic spot or shopping centre; Grid 4159 is more likely to be an University, educational centre or a workplace; Grid 4556 is more likely to include nightlife and entertainment as well as shopping centre.

5 Task II

5.1 For Grid 5161

For the following forecasting task, we deal only with the Internet activity data.

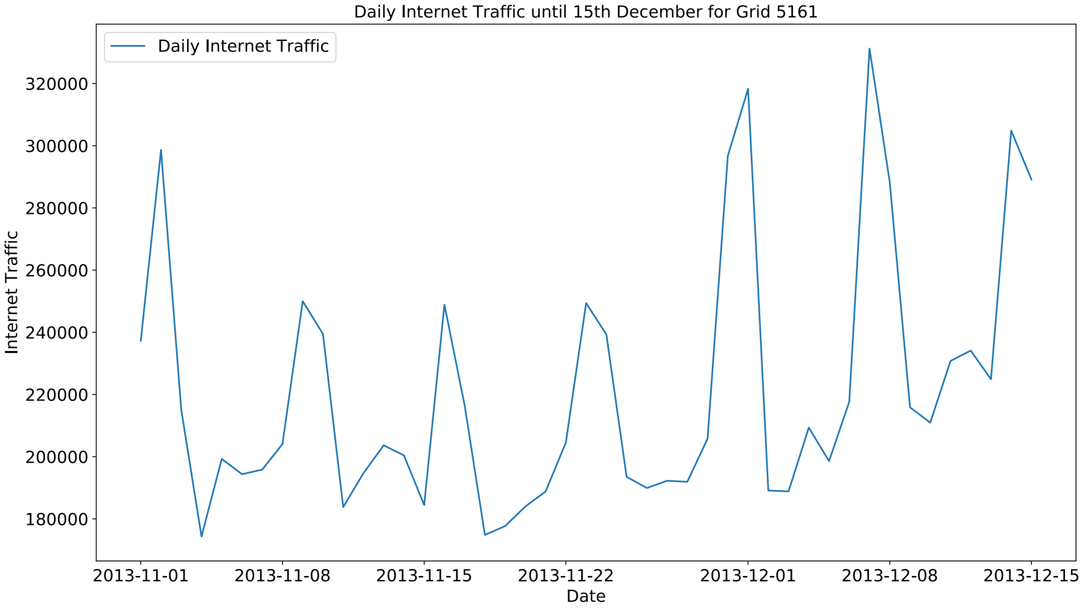


Figure 24: Time Series of Internet activity until December 15 (Training data) - Grid 5161

Figure 24 shows the time series of the internet activity which is used as the training data to forecast for the next 7 days - December 16 to 22. Figure 25 depicts the decomposition of the training data into Trend, Seasonality and Noise. It is clear from the figure that the time series has an increasing trend towards the end of the training data. In addition to the Trend, we can also observe that the seasonality has a frequency of 7 days which is to be expected as the daily activities are continual every week. Internet activity reaches new peaks during weekends and goes down during weekdays. In order to incorporate this seasonality, we will use ARIMA model with seasonal component, **SARIMA** [Seasonal Autoregressive Integrated Moving Average].

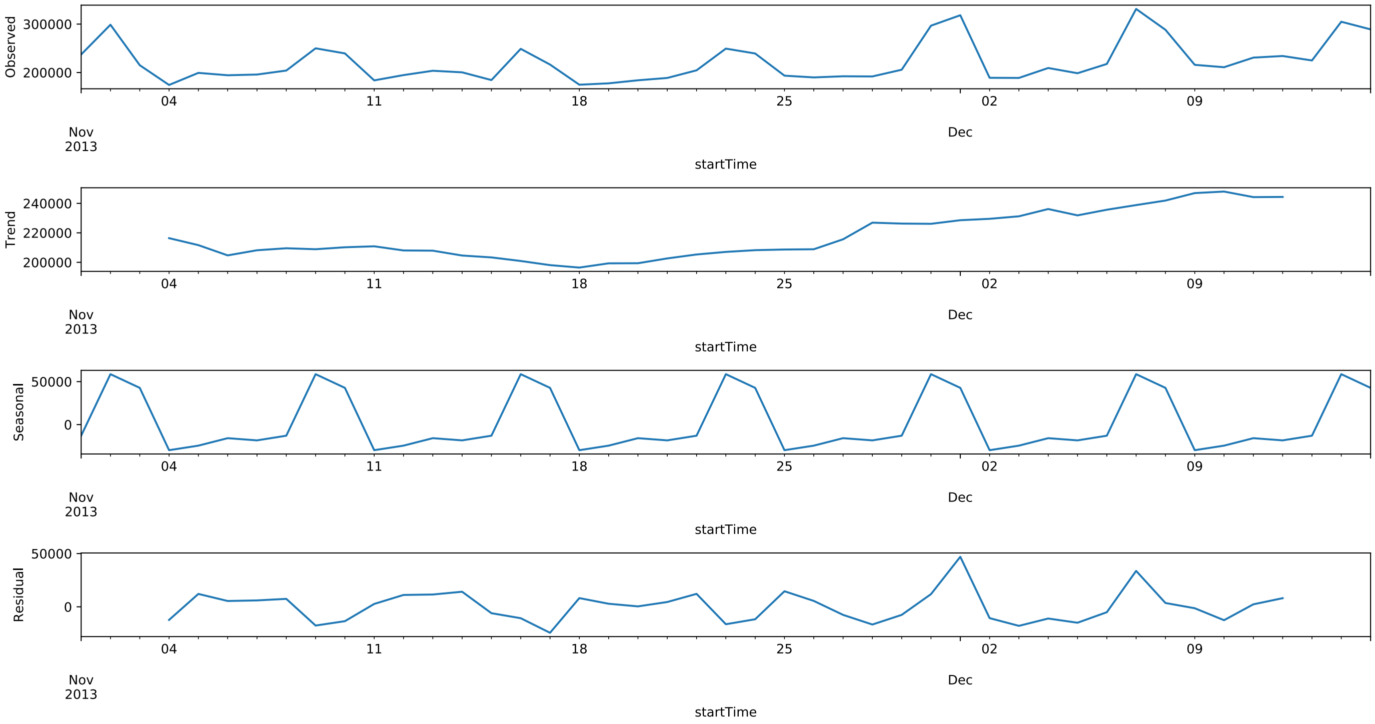


Figure 25: Time Series decomposition of Training Data for Grid 5161

5.1.1 Tuning

- In order to find the parameters for the SARIMA model, we need to find the parameters p,d,q of the Trend elements and P,D,Q,m of the Seasonal elements. We already know that m (number of time steps for a single seasonal period) = 7. Now, we need to obtain the other parameters for SARIMA.
- In order to check if the Trend is stationary or not, we perform Augmented Dickey Fuller Test (ADF). We obtain p-value = 0.9753 which is >0.05, so we cannot reject the null hypothesis and hence it is non-stationary. First Order differentiation can make it stationary. After performing the first differentiation, we obtain p-value = 8.4985 E-17 which is <0.05, so we can reject the null hypothesis of ADF. Thus, we now have a stationary timeseries.
- We now set d=1 and rest of the parameters of SARIMA model is obtained using *auto_ arima* function from the package *pmdarima*.
- The parameters of SARIMA obtained through the above process: SARMIA (p=0, d=1, q=0) (P=1, D=0, Q=0, m=7).

5.1.2 Prediciton

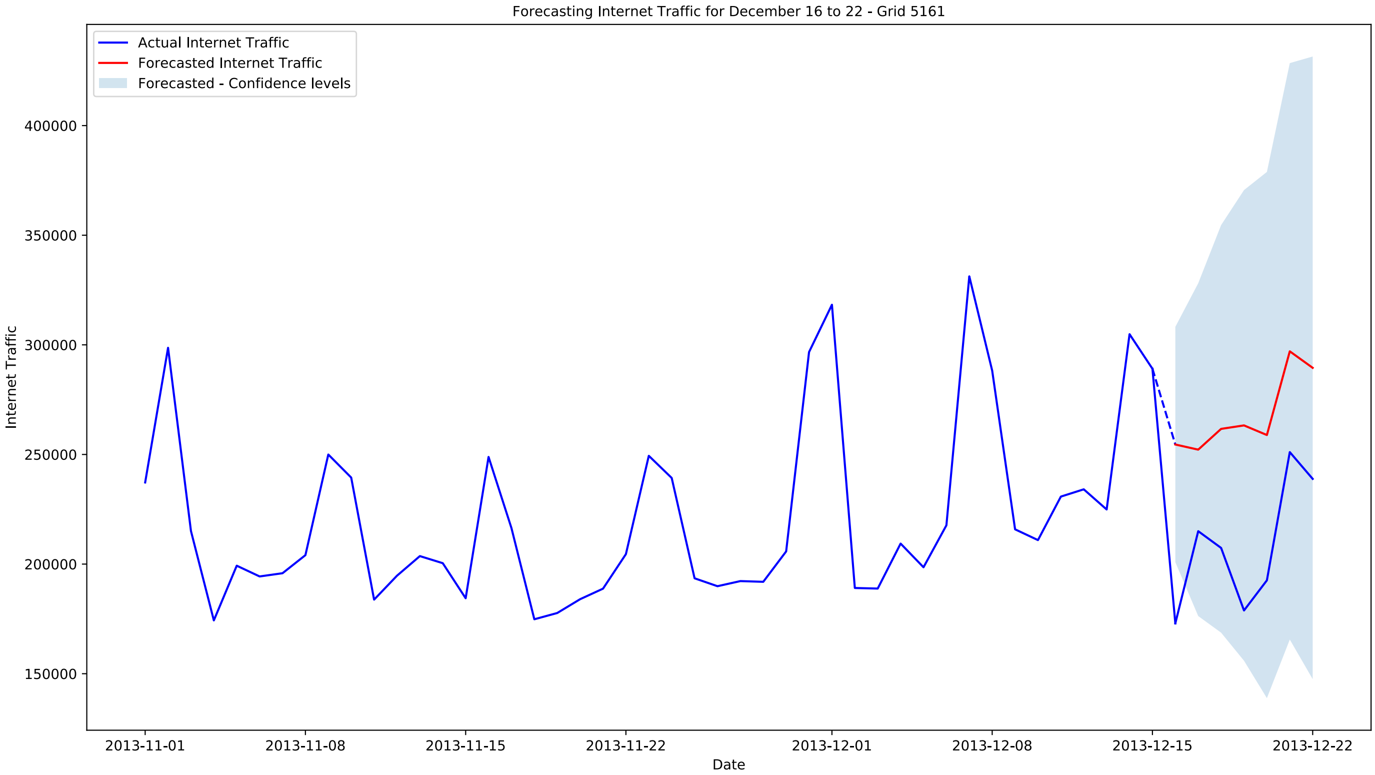


Figure 26: Forecasted data for Grid 5161

From figure 26, we can observe that the forecasted data is significantly deviated even though it has captured the seasonality element well. In order to find the exact reason, we can decompose the whole data and analyze.

It is very much clear from the from the figure 27 that the trend component went down all of a sudden during the forecasting period. This might be due to the holiday period during this time. In order for the algorithm to take this into account, we need to feed in the data for the past year so that the model can learn the generalized version of trend and seasonality for the whole year. Another way is to take into account whether we are predicting for the holiday period or non-holiday period and then train it on the similar scenario where we have a long holiday period.

5.2 For Grid 4159

Similarly, we follow the above steps and tune the SARIMA model for the Grid 4159. The forecasted data has been plotted in figure 28. This makes us curious on why the model provides a perfect prediction. Decomposition of the training data can reveal some details on why it is the case.

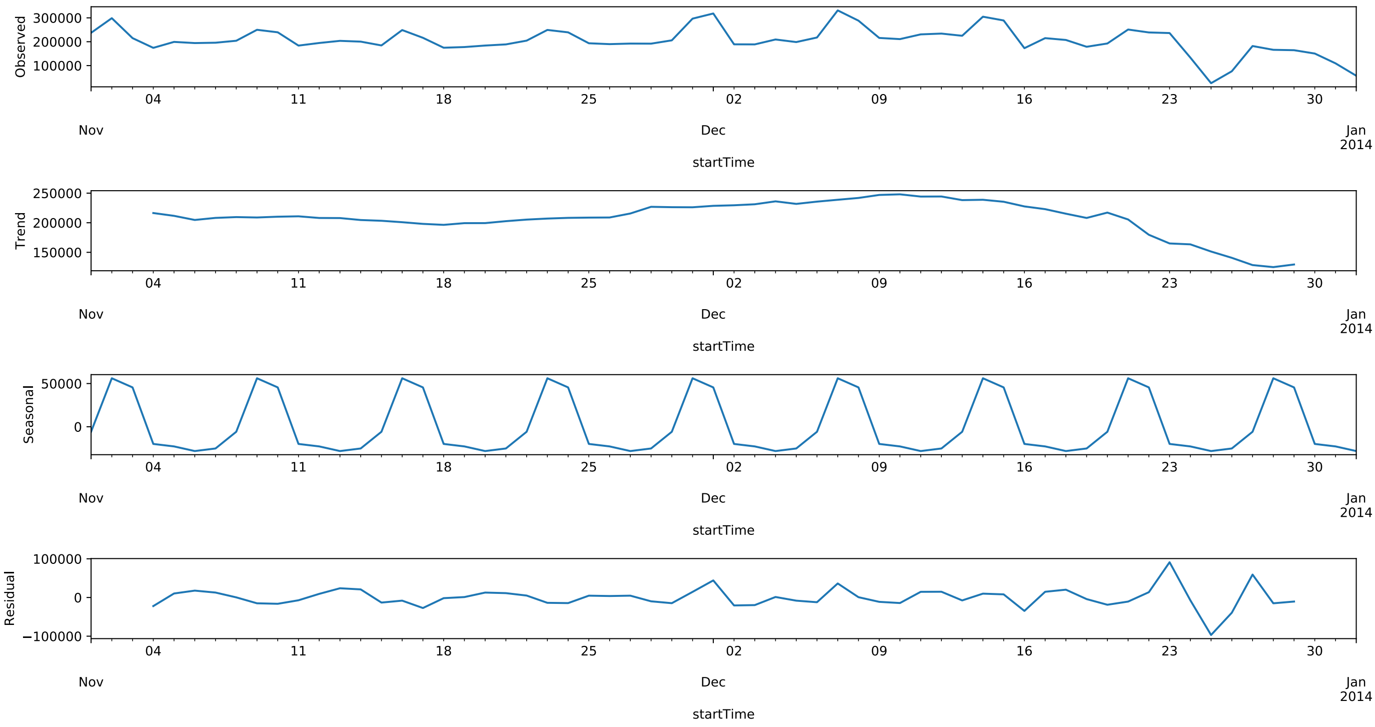


Figure 27: Time Series decomposition of Whole Data for Grid 5161

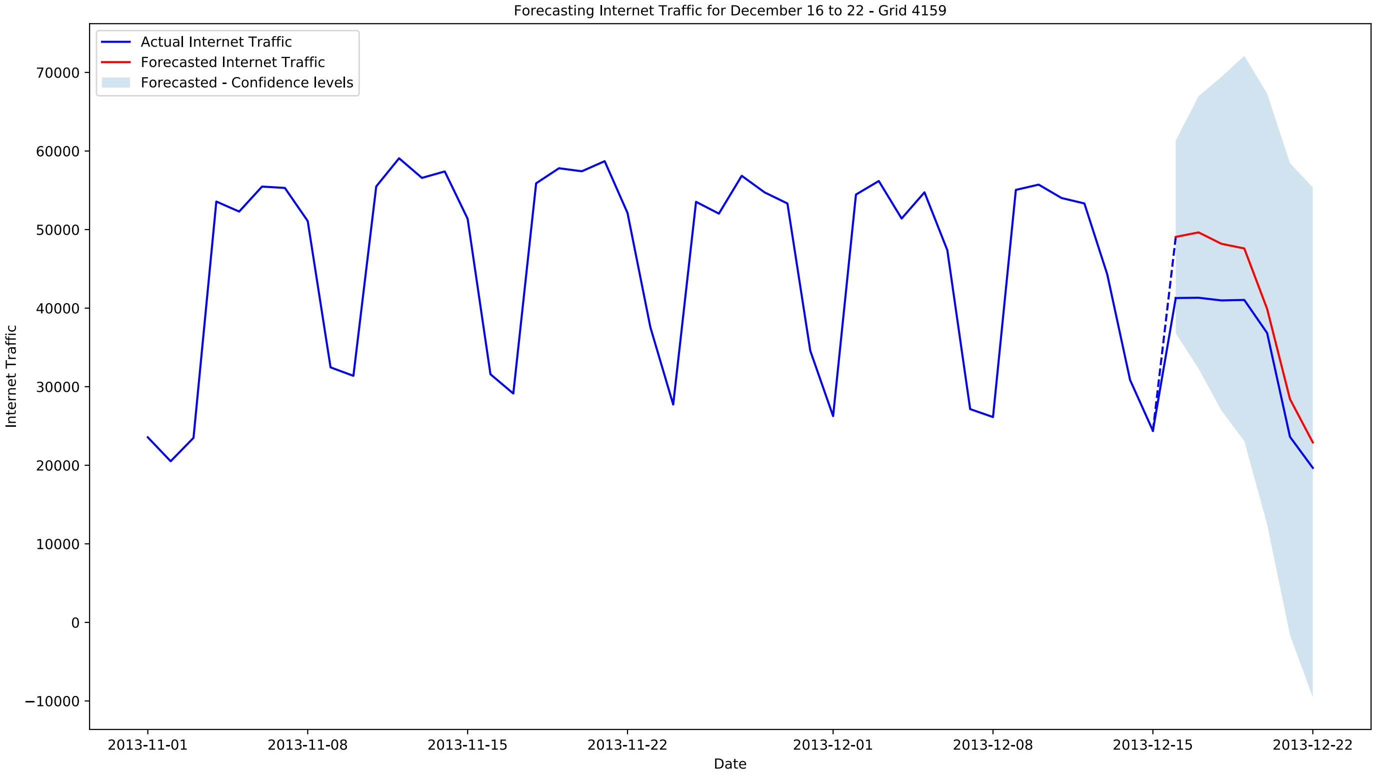


Figure 28: Forecasted data for Grid 4159

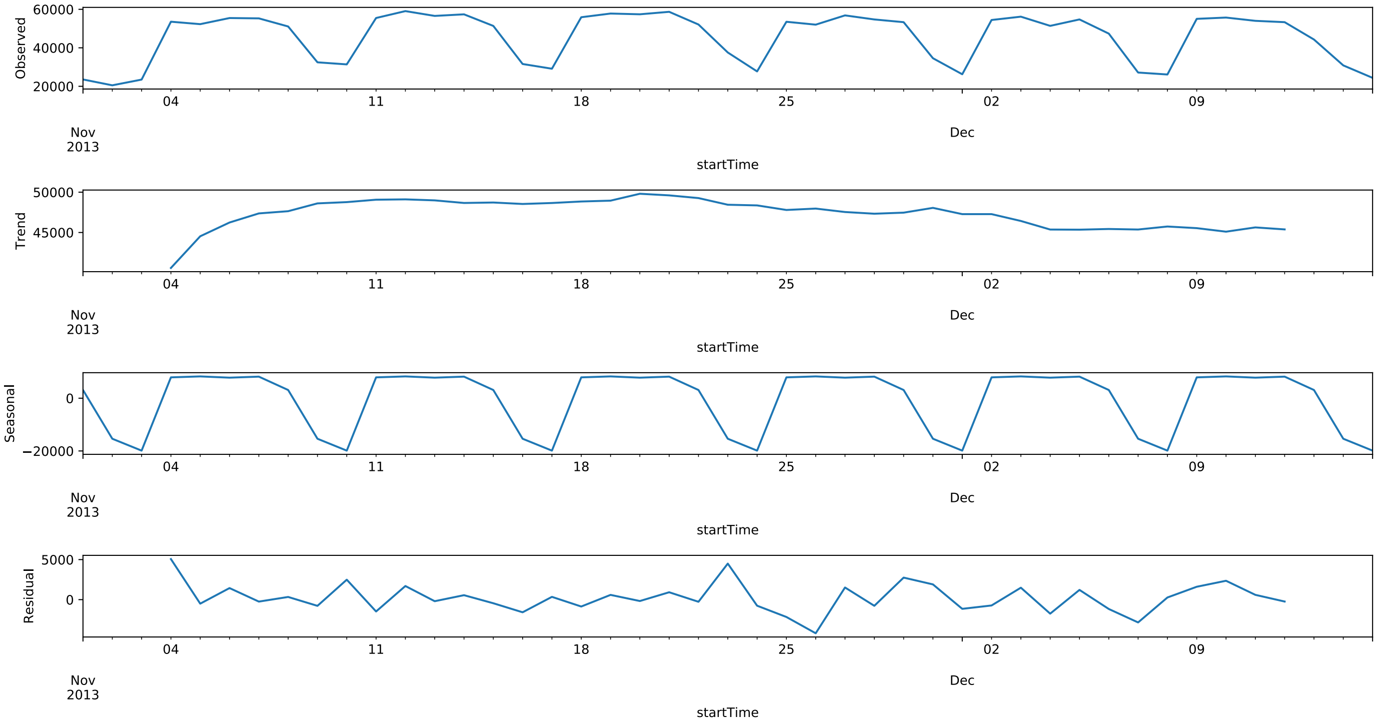


Figure 29: Time Series decomposition of Training Data for Grid 4159

Unlike the previous grid where the training data had an upward trend but the test data had a downward trend, here, from the figure 29, we can decipher that the training data already had a downward trend towards the end of the training data which has favored the testing data forecasting as they have a downward trend which can be seen from 28.

5.3 For Grid 4556

Similarly, we follow the above steps and tune the SARIMA model for the Grid 4556. The forecasted data has been plotted in figure 30.

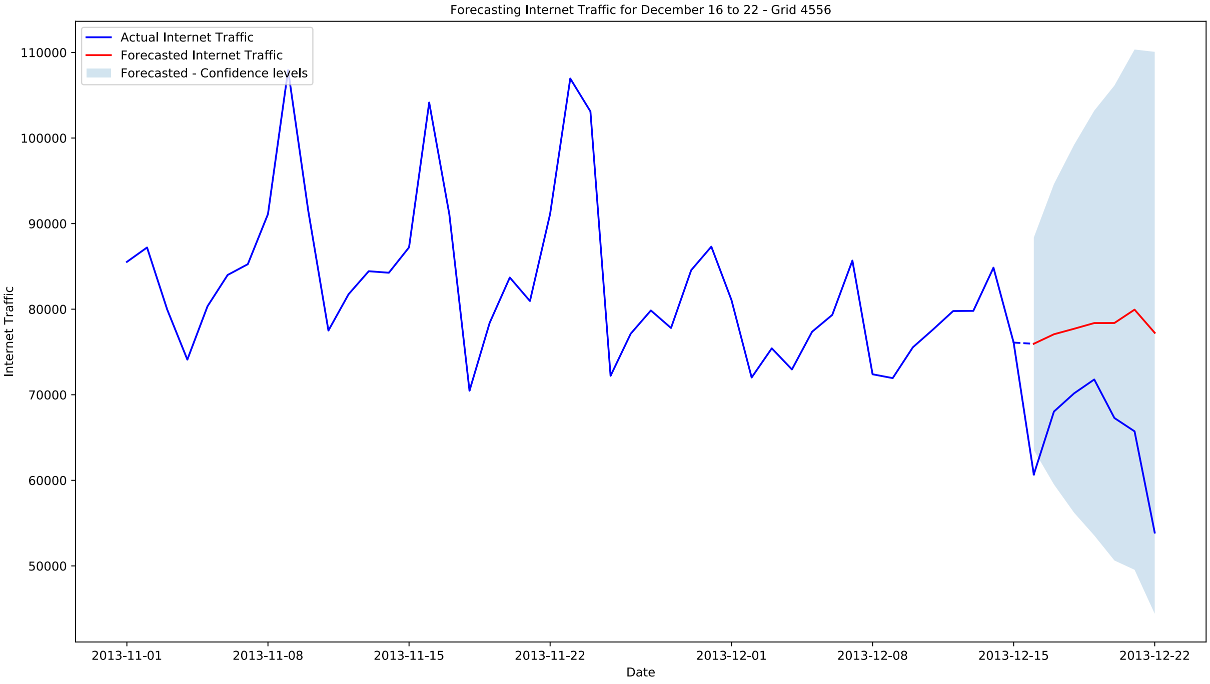


Figure 30: Forecasted data for Grid 4556

5.4 Results

The results obtained are reported in Table 1. The *auto_ arima* provides the tuning statistics which includes the execution time. Some suggestions on how the forecasting can be improved has been brief in section 5.1.2.

Grid	**MAE**	**MAPE**	**Fitting Time (s)**	**Net Execution Time (s)**
5161	60053.980	30.254%	2.108	2.154
4159	5853.548	16.809%	3.392	3.484
4556	12445.145	19.978%	2.192	2.248

Table 1: Forecasted Statistics for three grids

Net Execution Time = Fitting Time + Inference Time

References

[1] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, January 2008.

[2] Alan F. Gates, Olga Natkovich, Shubham Chopra, Pradeep Kamath, Shravan M. Narayanamurthy, Christopher Olston, Benjamin Reed, Santhosh Srinivasan, and Utkarsh Srivastava. Building a high-level dataflow system on top of map-reduce: The pig experience. *Proc. VLDB Endow.*, 2(2):1414–1425, August 2009.

[3] Telecom Italia. Telecommunications - SMS, Call, Internet - MI. *https://doi.org/10.7910/DVN/EGZHFV*, 2015.

[4] Telecom Italia. Milano Grid. *https://doi.org/10.7910/DVN/QJWLFU*, 2015.