Modeling of Telecommunication CDRs to identify network usage pattern

Capstone Project 1 - Milestone Report

1. PROBLEM STATEMENT

The exponential increase in the use of internet services and mobile phones is generating large amount of data that can be used to provide useful insights about the network usage pattern.

Call Detail Record [CDR] describes a specific instance of a telecommunication transaction that passes through a network element. Every time a user performs a telecom activity such as send/receive SMS and calls, a CDR is generated. It contains information about the caller/sender ID, location, time, data used, etc. Millions and millions of such records are generated and is mainly used for billing purposes by the telecom company.

Analysis and modeling of this time series data helps to identify geographical boundaries of various usage patterns. This helps in decision making of resource allocation by telecommunication companies who own the network elements, inspecting quantitatively different aspects of human behavior such as socio-economic status of geographical regions and people's mobility. CDRs collected for a span of time can also be used in forecasting future volumes for a network.

2. DATASET

This dataset 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.

For this project we will use telecommunication data from the city of Milan, Italy. It is available as .txt files with tab-delimited values (TSV) from the link, https://doi.org/10.7910/DVN/EGZHFV.
There are 62 files consisting of CDRs collected from Nov 1,2013 to Jan 1, 2014, one file for each day.
File size: ~ 200MB – 300MB with ~4 million rows each file
Files: sms-call-internet-mi-2013-11-01.txt to sms-call-internet-mi-2014-01-01.txt
There are 8 columns, with no headers. Each column represents,

- Grid id: identification string of a given square of Milan GRID. The geographical region of the city is spatially divided into 1000 square grids.
- Time Interval: start interval time expressed in milliseconds. The end interval time can be obtained by adding 600,000 milliseconds (10 min) to this value;
- SMS-in activity: activity proportional to the amount of received SMSs inside a given Square id and during a given Time interval. The SMSs are sent from the nation identified by the Country code;
- SMS-out activity: activity proportional to the amount of sent SMSs inside a given Square id during a given Time interval. The SMSs are received in the nation identified by the Country code;

- Call-in activity: activity proportional to the amount of received calls inside the Square id during a given Time interval. The calls are issued from the nation identified by the Country code;
- Call-out activity: activity proportional to the amount of issued calls inside a given Square id during a given Time interval. The calls are received in the nation identified by the Country code;
- Internet traffic activity: number of CDRs generated inside a given Square id during a given Time interval. The Internet traffic is initiated from the nation identified by the Country code;
- Country code: the phone country code of the nation.

As the original bundle of dataset comes from various companies (telecommunications, weather, news, social networks and electricity) with different standards, in order to ease the comparisons of different geographical areas, the city of Milan's spatial distribution is aggregated in a grid with square cells. The area is composed of a grid overlay of 10,000 squares with size of about 235×235 meters. This dataset provides the geographical reference of each square which composes the grids in the reference system: WGS 84—EPSG:4326. It is downloaded from the link, https://doi.org/10.7910/DVN/QJWLFU.

File type: geojson

File size: 3MB

File: milano-grid.geojson

Columns:

- *square id*: identification string of a given square of the Milan GRID;
- *Time Interval*: The cell geometry expressed as geoJSON and projected in WGS84 (EPSG:4326).

3. DATA WRANGLING

Reading data into Pandas Dataframe

Reading 20GB of data from 62 files iteratively and merging them into a single dataframe takes about 30 min and utilizes very high system memory. Alternatively, reading a file into a dataframe, perform operations like sampling [daily, hourly], grouping & indexing that reduces the number of rows and combining these individual data frames into a single dataframe results in a faster and efficient loading operation.

Time Interval column

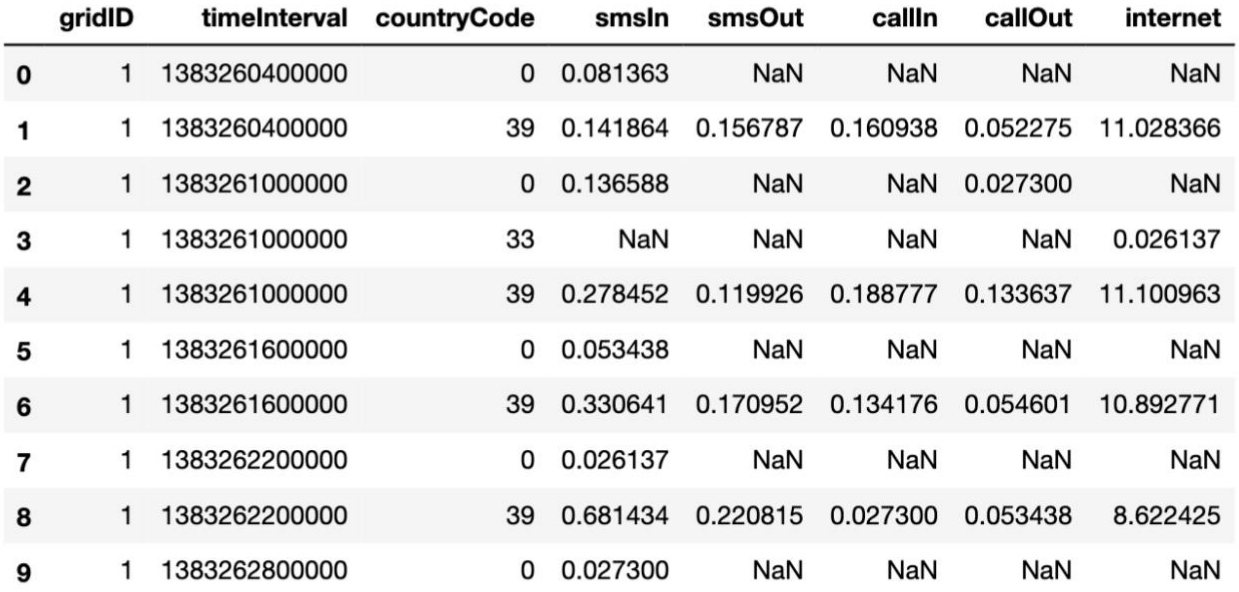
Time Interval column is represented in milliseconds, as epoch/Unix timestamps. It is the number of milliseconds passed since 00:00:00 UTC Thursday, 1 January 1970. This column values are converted to pandas Datetime object that can be used with Pandas in built functions. This value is stored in a new column, startTime.

Dropping unwanted columns

Time Interval column now has redundant values and Country code column will not be used in this project. They are both dropped from the data frame.

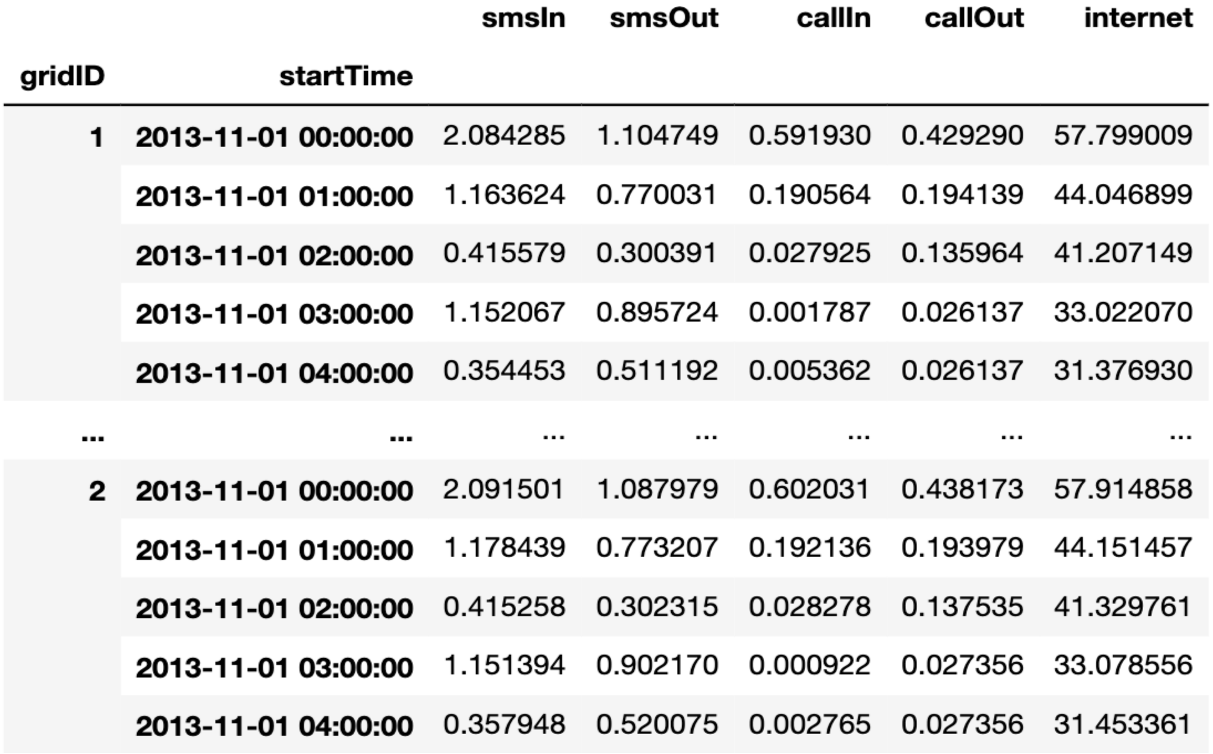
Resampling, Grouping & Indexing

Several rows of data, with 10 min time interval are aggregated into daily (24 hour) and hourly time intervals. They are grouped and indexed by Grid ID & startTime columns. Total volume of each activity over the 2 months for individual grids is calculated.



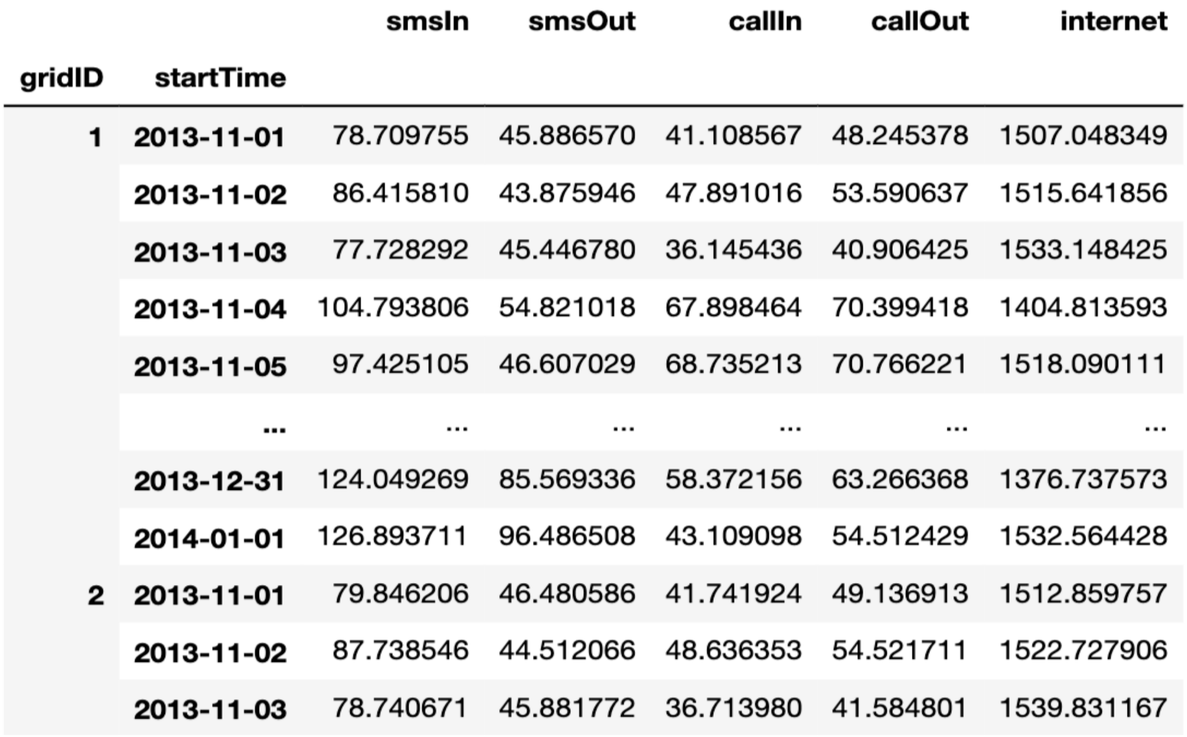
	gridID	timeInterval	countryCode	smsIn	smsOut	callIn	callOut	internet
0	1	1383260400000	0	0.081363	NaN	NaN	NaN	NaN
1	1	1383260400000	39	0.141864	0.156787	0.160938	0.052275	11.028366
2	1	1383261000000	0	0.136588	NaN	NaN	0.027300	NaN
3	1	1383261000000	33	NaN	NaN	NaN	NaN	0.026137
4	1	1383261000000	39	0.278452	0.119926	0.188777	0.133637	11.100963
5	1	1383261600000	0	0.053438	NaN	NaN	NaN	NaN
6	1	1383261600000	39	0.330641	0.170952	0.134176	0.054601	10.892771
7	1	1383262200000	0	0.026137	NaN	NaN	NaN	NaN
8	1	1383262200000	39	0.681434	0.220815	0.027300	0.053438	8.622425
9	1	1383262800000	0	0.027300	NaN	NaN	NaN	NaN

Fig 1: Data as read from the files



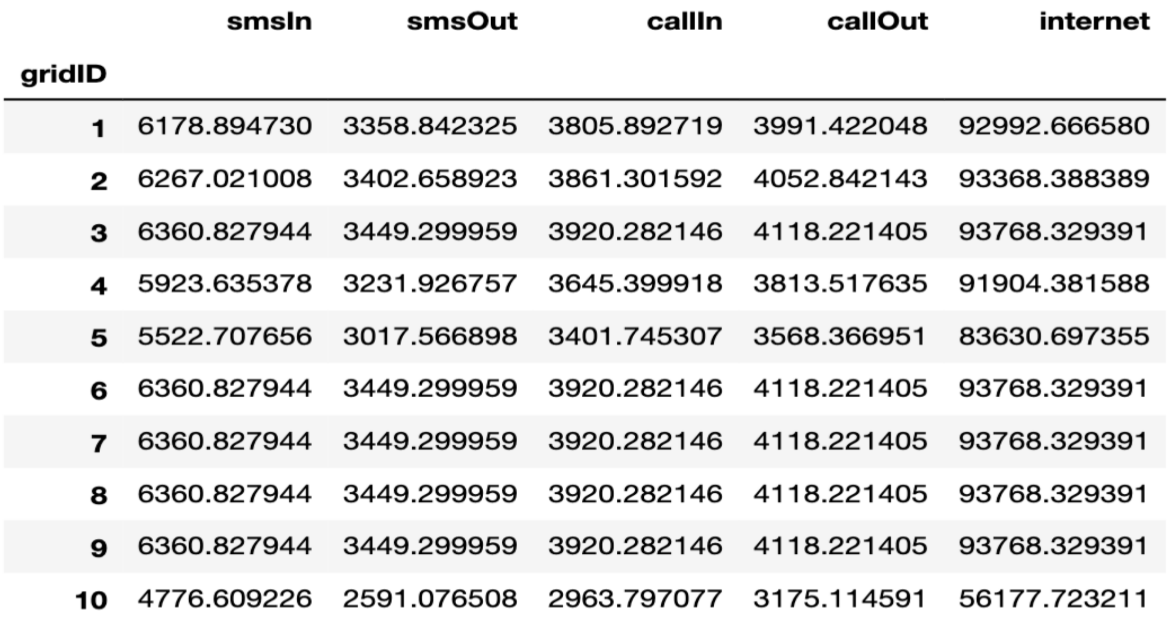
		smsIn	smsOut	callIn	callOut	internet
gridID	**startTime**					
1	**2013-11-01 00:00:00**	2.084285	1.104749	0.591930	0.429290	57.799009
	2013-11-01 01:00:00	1.163624	0.770031	0.190564	0.194139	44.046899
	2013-11-01 02:00:00	0.415579	0.300391	0.027925	0.135964	41.207149
	2013-11-01 03:00:00	1.152067	0.895724	0.001787	0.026137	33.022070
	2013-11-01 04:00:00	0.354453	0.511192	0.005362	0.026137	31.376930
...	**...**	...	...	...	...	...
2	**2013-11-01 00:00:00**	2.091501	1.087979	0.602031	0.438173	57.914858
	2013-11-01 01:00:00	1.178439	0.773207	0.192136	0.193979	44.151457
	2013-11-01 02:00:00	0.415258	0.302315	0.028278	0.137535	41.329761
	2013-11-01 03:00:00	1.151394	0.902170	0.000922	0.027356	33.078556
	2013-11-01 04:00:00	0.357948	0.520075	0.002765	0.027356	31.453361

Fig 2: Grid-wise hourly aggregation of telecommunication activities



		smsIn	smsOut	callIn	callOut	internet
gridID	**startTime**					
1	**2013-11-01**	78.709755	45.886570	41.108567	48.245378	1507.048349
	2013-11-02	86.415810	43.875946	47.891016	53.590637	1515.641856
	2013-11-03	77.728292	45.446780	36.145436	40.906425	1533.148425
	2013-11-04	104.793806	54.821018	67.898464	70.399418	1404.813593
	2013-11-05	97.425105	46.607029	68.735213	70.766221	1518.090111
	...	...	...	...	...	...
	2013-12-31	124.049269	85.569336	58.372156	63.266368	1376.737573
	2014-01-01	126.893711	96.486508	43.109098	54.512429	1532.564428
2	**2013-11-01**	79.846206	46.480586	41.741924	49.136913	1512.859757
	2013-11-02	87.738546	44.512066	48.636353	54.521711	1522.727906
	2013-11-03	78.740671	45.881772	36.713980	41.584801	1539.831167

Fig 3: Grid-wise daily aggregation of telecommunication activities



	smsIn	smsOut	callIn	callOut	internet
gridID					
1	6178.894730	3358.842325	3805.892719	3991.422048	92992.666580
2	6267.021008	3402.658923	3861.301592	4052.842143	93368.388389
3	6360.827944	3449.299959	3920.282146	4118.221405	93768.329391
4	5923.635378	3231.926757	3645.399918	3813.517635	91904.381588
5	5522.707656	3017.566898	3401.745307	3568.366951	83630.697355
6	6360.827944	3449.299959	3920.282146	4118.221405	93768.329391
7	6360.827944	3449.299959	3920.282146	4118.221405	93768.329391
8	6360.827944	3449.299959	3920.282146	4118.221405	93768.329391
9	6360.827944	3449.299959	3920.282146	4118.221405	93768.329391
10	4776.609226	2591.076508	2963.797077	3175.114591	56177.723211

Fig 4: Grid-wise total volume of telecommunication activities over the 2 months

Visualization of Grids

milano-grid.geojson file is loaded using packages geopandas & geojsonio that shows the overlay of 10000 grids over the city of Milan's map.

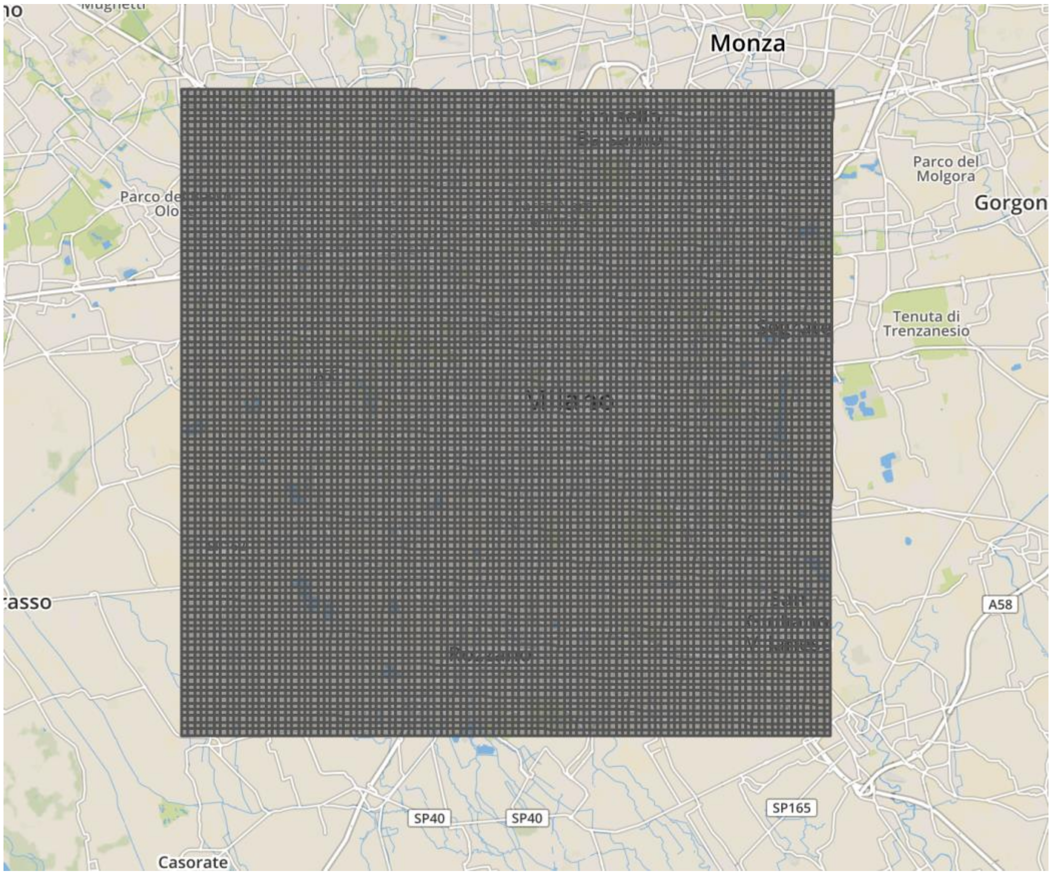


Fig 5: Spatial aggregation of 10,000 grids over the city of Milan

4. EXPLORATORY DATA ANALYSIS

In the network, SMS-In and SMS-Out utilizes the control channel, Call-In and Call-Out utilizes transmitted over voice channel and the internet is transmitted over broadband frequencies. Thus, we will use SMS (sum of SMS-In & SMS-Out), Call (sum of Call-In & Call-Out) and Internet activity for the analysis.

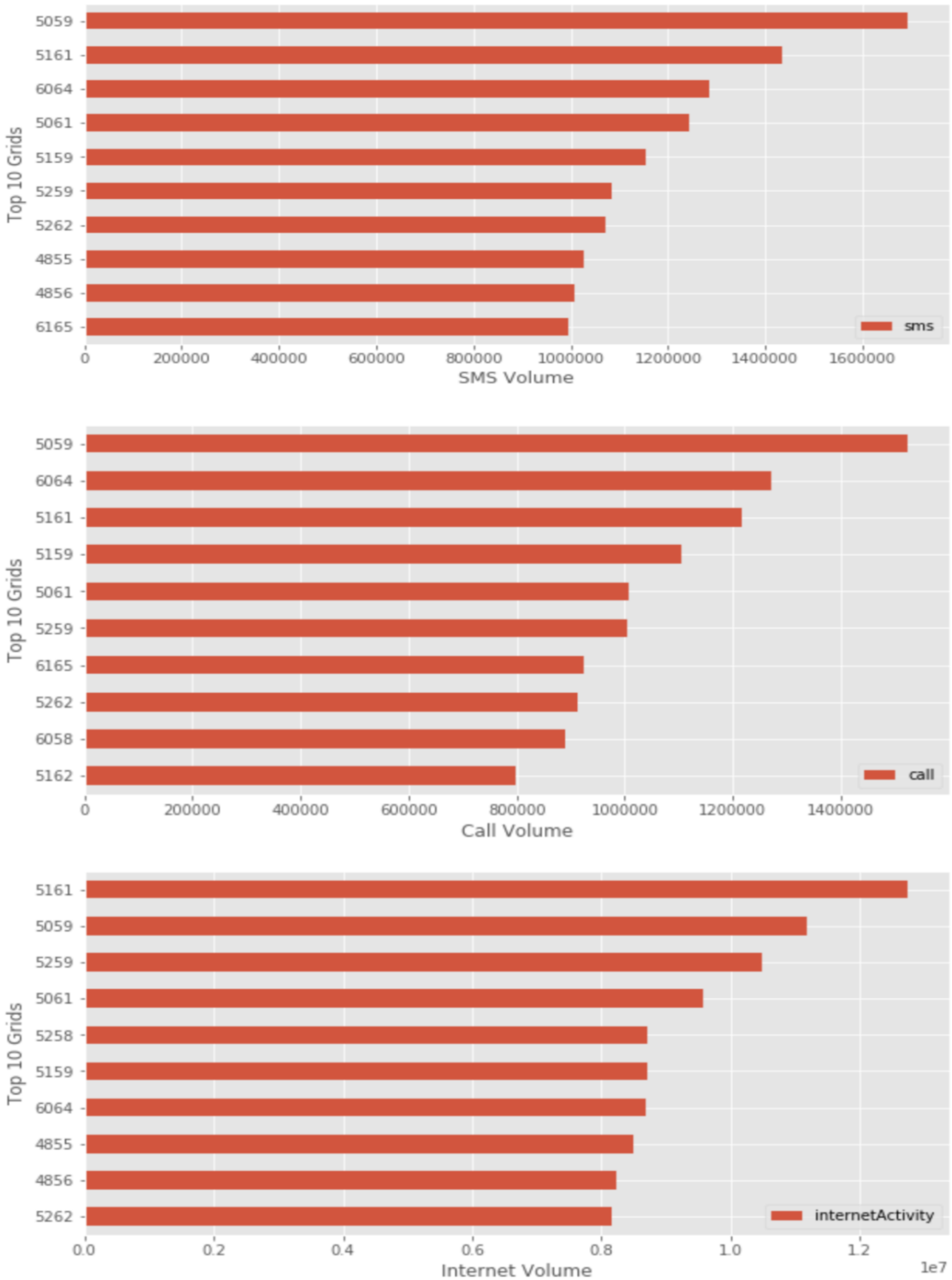


Fig 6: Horizontal bar plots showing top 10 grids with high total volumes in each telecommunication activity

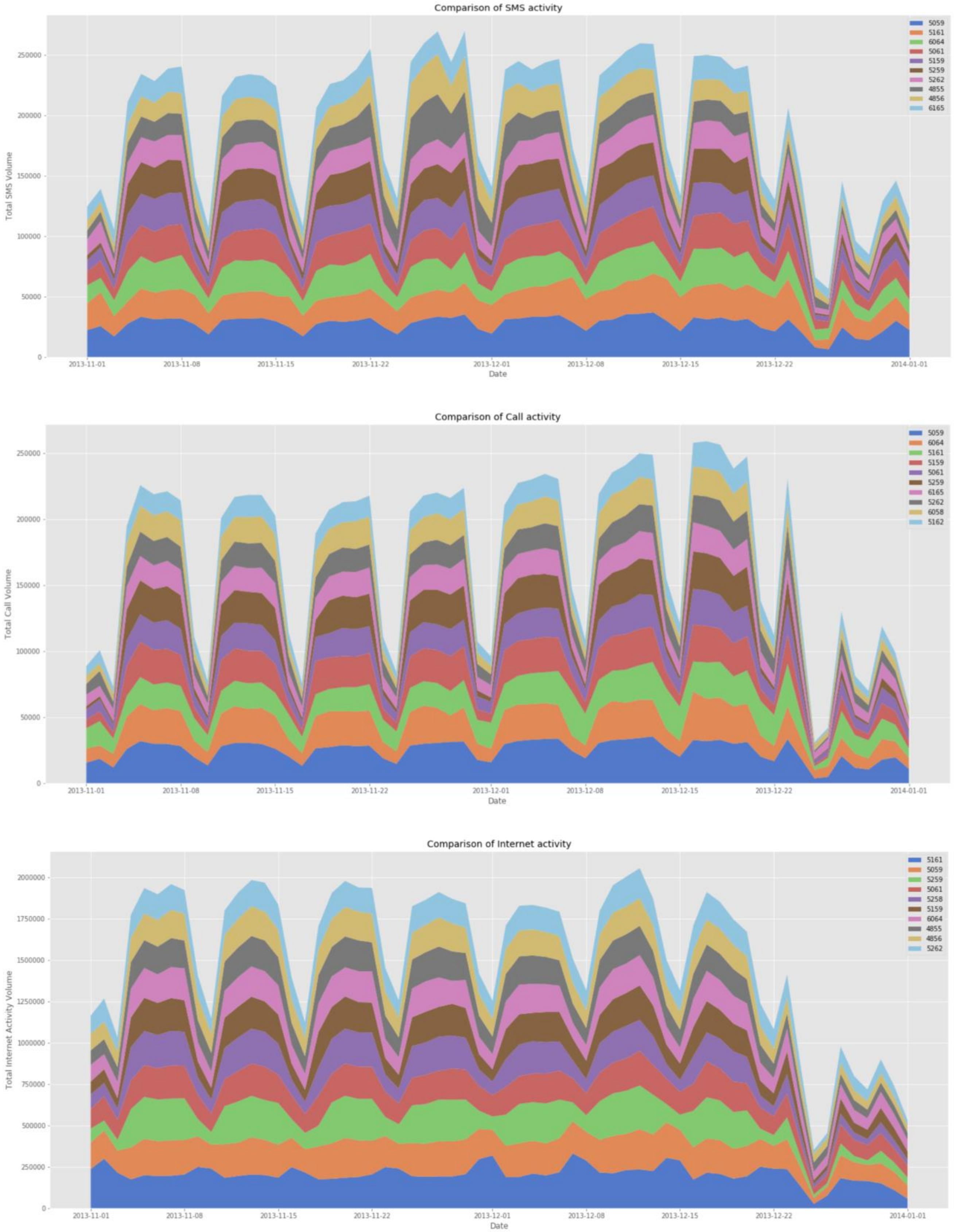
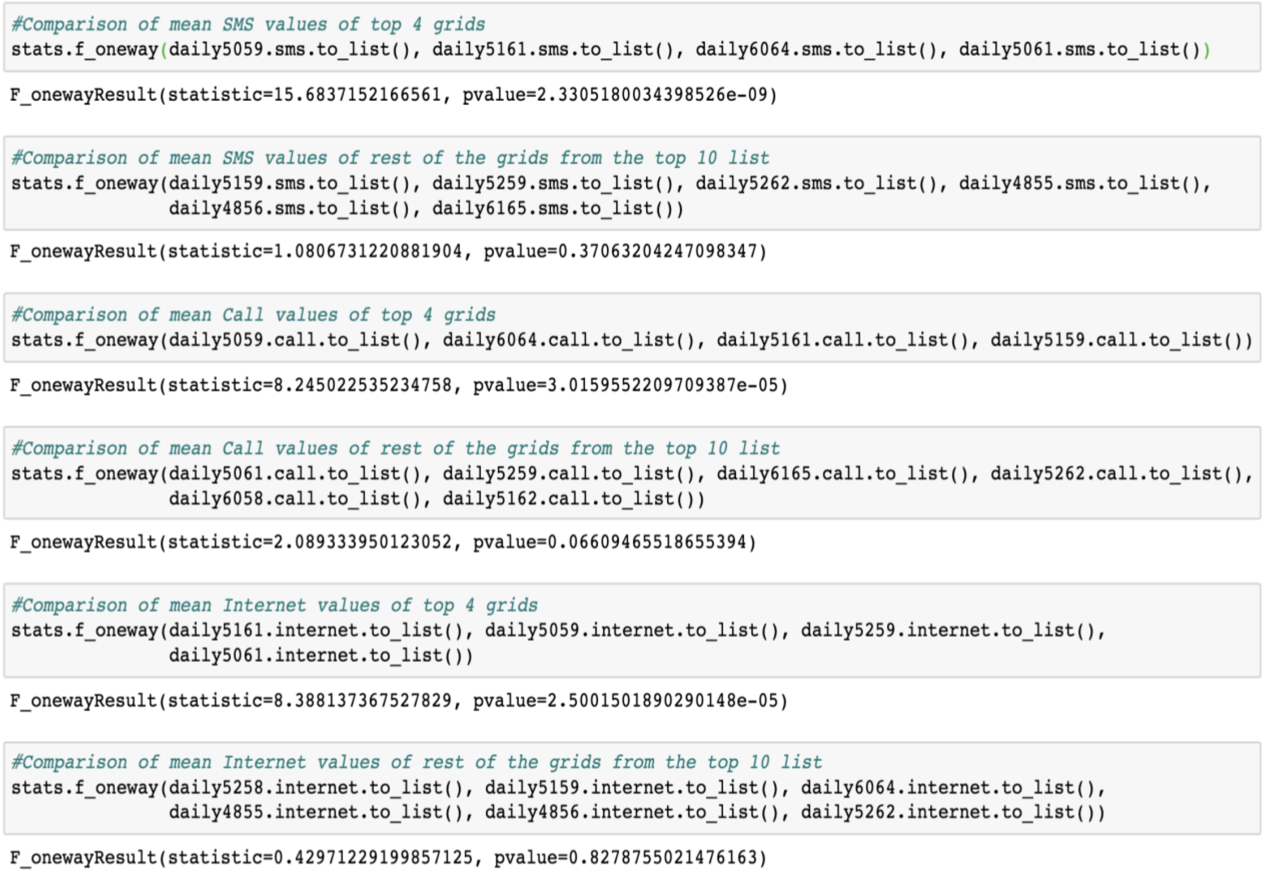


Fig 7: Stacked area plots showing comparison of top 10 grids daily pattern

Top 10 grids that experience high volumes for each of these activities for the 2 months are identified. From the above plots, we see that the top four grids have highly varying total volume and rest of the grids have almost same total volume for each activity. This is further verified by performing a set of one-way ANOVA tests as shown in Fig 7. P-value < 0.05 indicates mean values of the grids are not equal, P-value > 0.05 indicates means values of the grids are equal.



```
#Comparison of mean SMS values of top 4 grids
stats.f_oneway(daily5059.sms.to_list(), daily5161.sms.to_list(), daily6064.sms.to_list(), daily5061.sms.to_list())

F_onewayResult(statistic=15.6837152166561, pvalue=2.3305180034398526e-09)

#Comparison of mean SMS values of rest of the grids from the top 10 list
stats.f_oneway(daily5159.sms.to_list(), daily5259.sms.to_list(), daily5262.sms.to_list(), daily4855.sms.to_list(),
               daily4856.sms.to_list(), daily6165.sms.to_list())

F_onewayResult(statistic=1.0806731220881904, pvalue=0.37063204247098347)

#Comparison of mean Call values of top 4 grids
stats.f_oneway(daily5059.call.to_list(), daily6064.call.to_list(), daily5161.call.to_list(), daily5159.call.to_list())

F_onewayResult(statistic=8.245022535234758, pvalue=3.0159552209709387e-05)

#Comparison of mean Call values of rest of the grids from the top 10 list
stats.f_oneway(daily5061.call.to_list(), daily5259.call.to_list(), daily6165.call.to_list(), daily5262.call.to_list(),
               daily6058.call.to_list(), daily5162.call.to_list())

F_onewayResult(statistic=2.089333950123052, pvalue=0.06609465518655394)

#Comparison of mean Internet values of top 4 grids
stats.f_oneway(daily5161.internet.to_list(), daily5059.internet.to_list(), daily5259.internet.to_list(),
               daily5061.internet.to_list())

F_onewayResult(statistic=8.388137367527829, pvalue=2.5001501890290148e-05)

#Comparison of mean Internet values of rest of the grids from the top 10 list
stats.f_oneway(daily5258.internet.to_list(), daily5159.internet.to_list(), daily6064.internet.to_list(),
               daily4855.internet.to_list(), daily4856.internet.to_list(), daily5262.internet.to_list())

F_onewayResult(statistic=0.42971229199857125, pvalue=0.8278755021476163)
```

Fig 8: One-Way ANOVA tests verifying that, except the top 4 grids, rest have similar mean volumes

Location of these grids in the map shows that they are all from Duomo & Milano Centrale region.

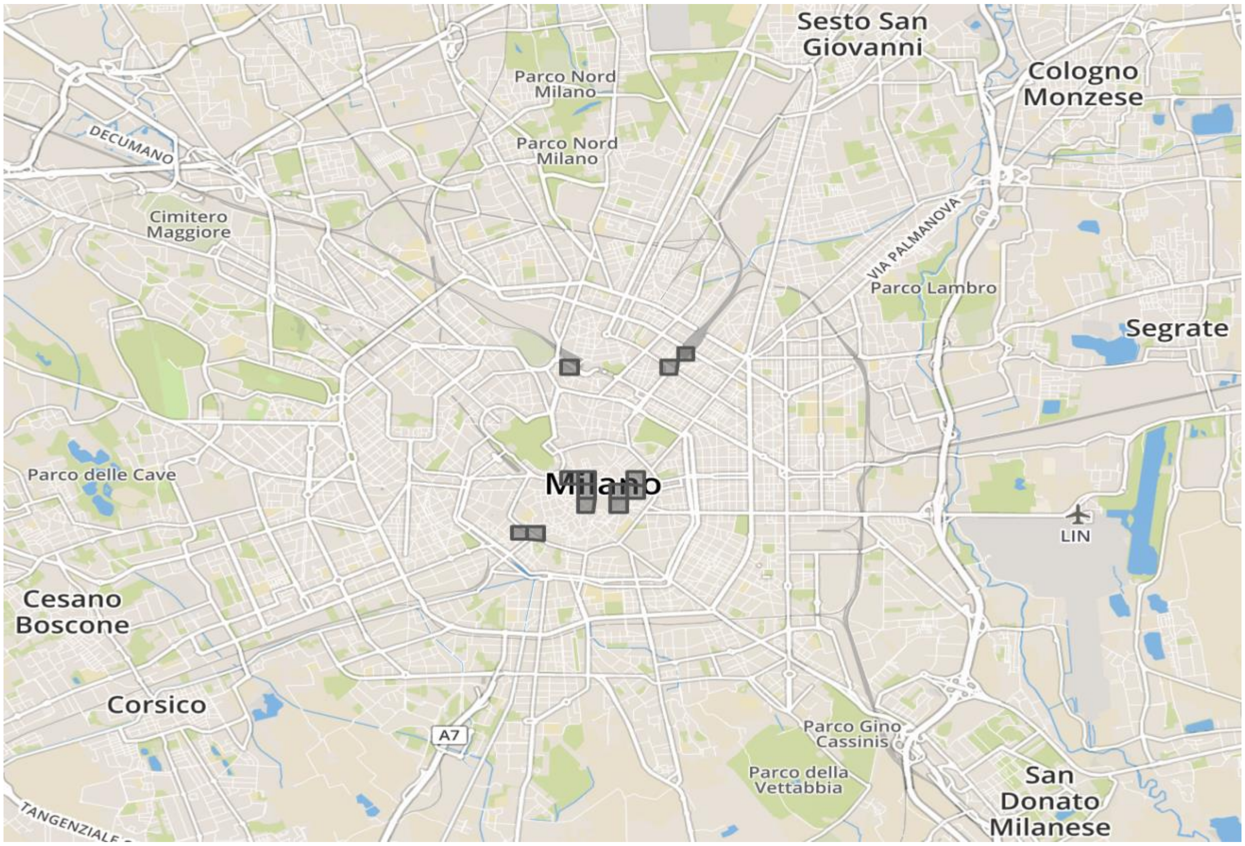


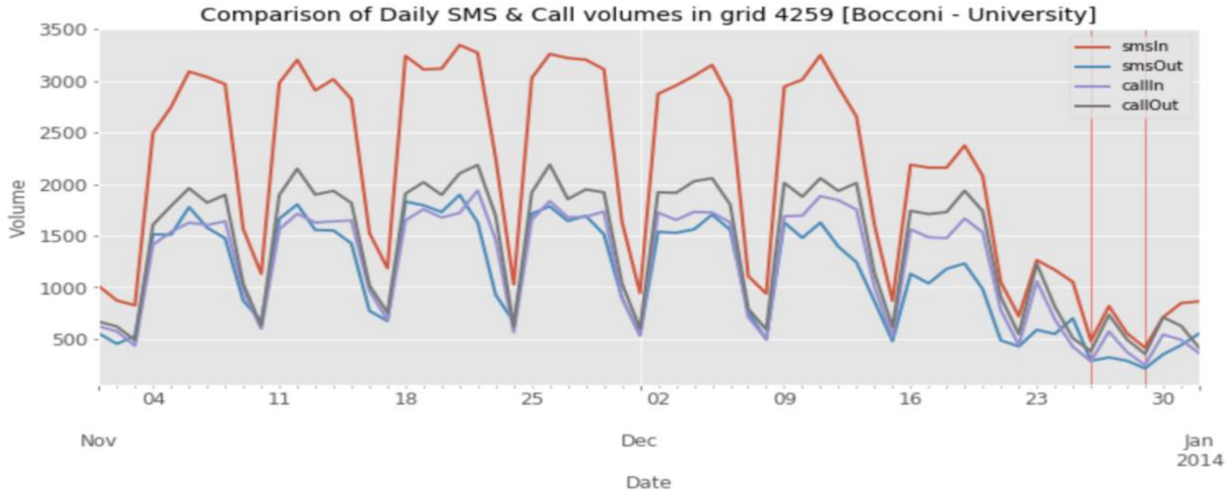
Fig 9: Location of top 10 high-volume grids in the city map

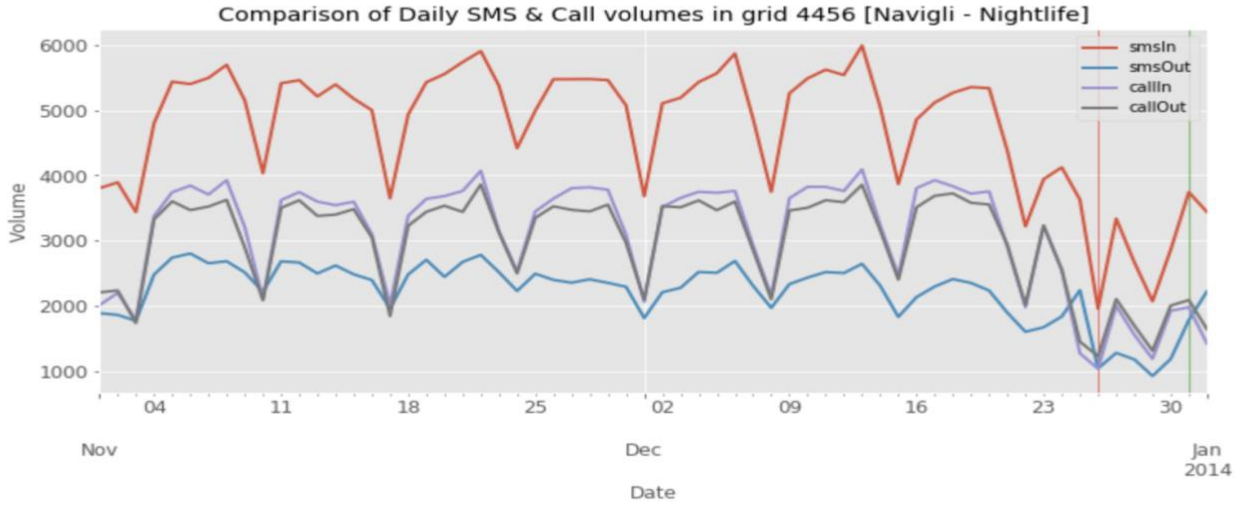
In Detail Analysis of grids from different sectors

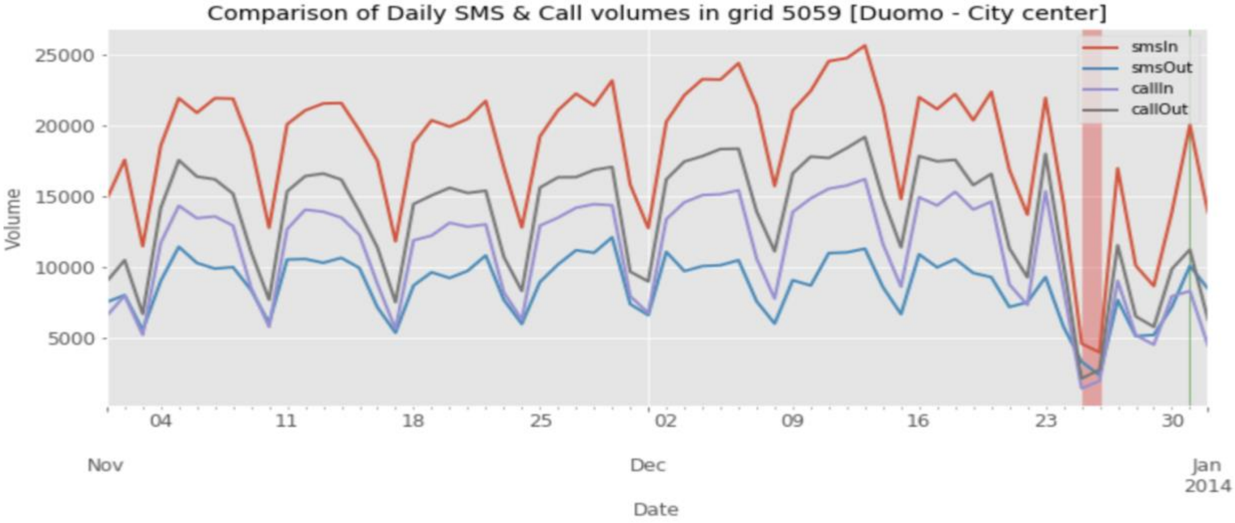
All the top contributing grids are from the city’s center and mostly near transport hubs, they are expected to show similar behavior and most of these grids have approximately same mean values. In order to capture variations in the city’s telecommunication activities, we will examine the following four grids that has markedly different behavioral signatures,

4259 - Bocconi, one of the most famous Universities in Milan
4456 - Navigli district, one of the most famous nightlife places in Milan
5059 - Duomo, the city center of Milan
5346 - Fiera, residential neighborhood of Milan

Daily Activity Plots:







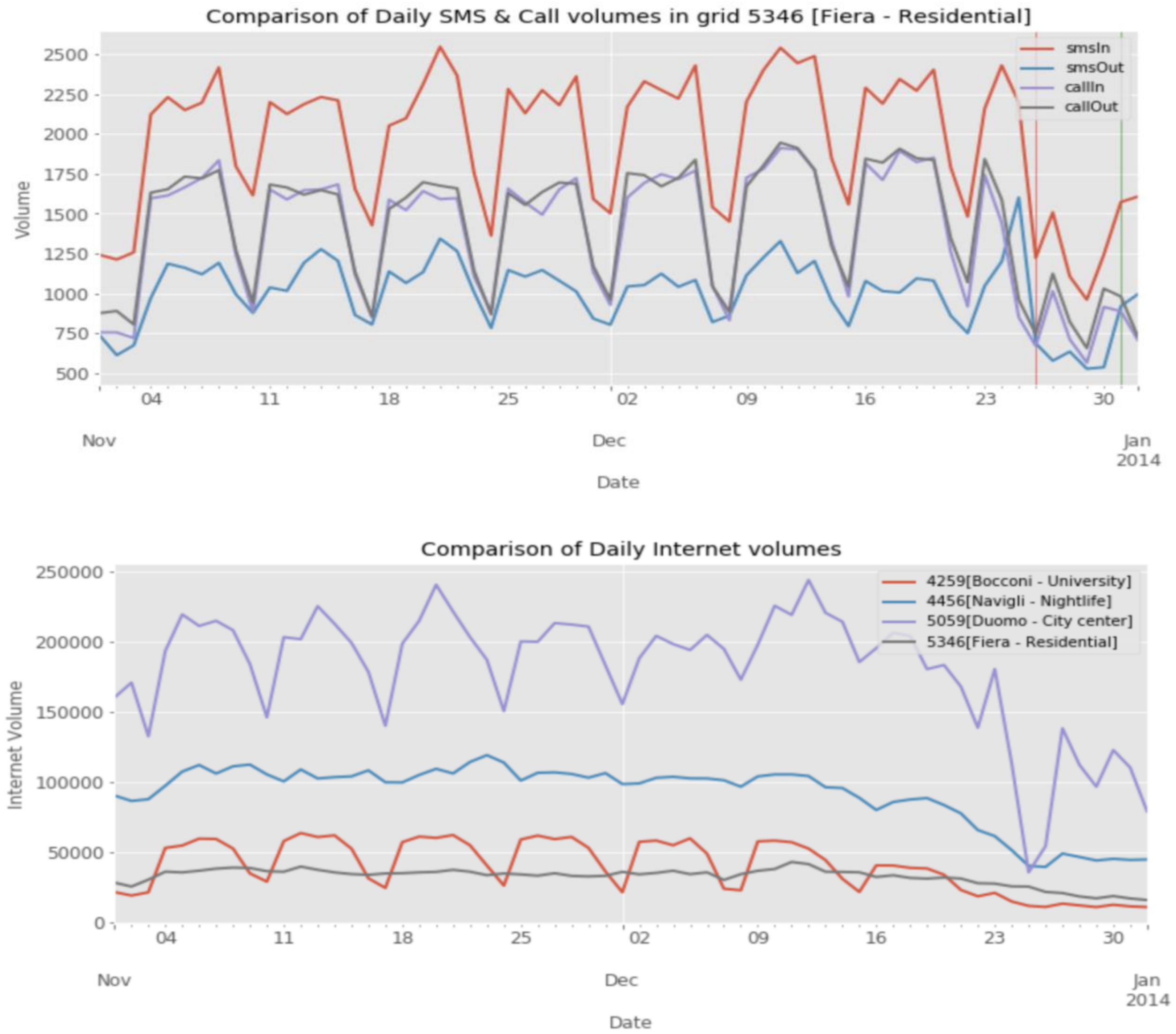


Fig 10: Time-series plot of daily telecommunication activities of the four grids

- All four grids have received high volumes of incoming SMS compared to other activities. Outgoing SMS has the least volume, almost equal amounts of Calls are made and received.
- Duomo has highest volume of all activities, followed by Navigli, then Bocconi and Fiera in the end. We can order the grids based on total volumes as,
  Duomo [city center] > Navigli [nightlife] > Bocconi [university] > Fiera [residential]
- All four grids exhibit seasonality in SMS & Call activities. In internet activity Navigli & Fiera doesn’t show any seasonality. This may because of IoT, with many devices always being connected to the network.
- There is a drop in the volumes towards December end in all the plots [holiday season].

Hourly Activity Plots:

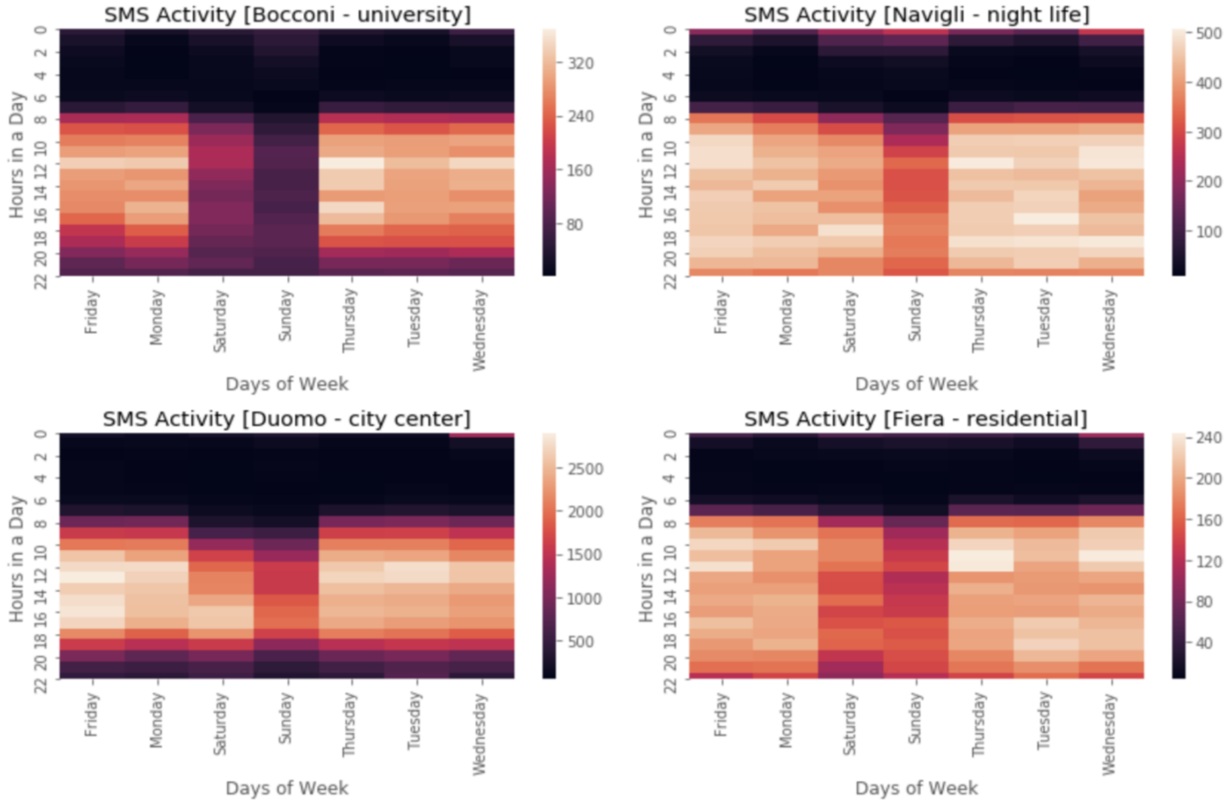


Fig 11: Heat map of hourly SMS activity of the four grids

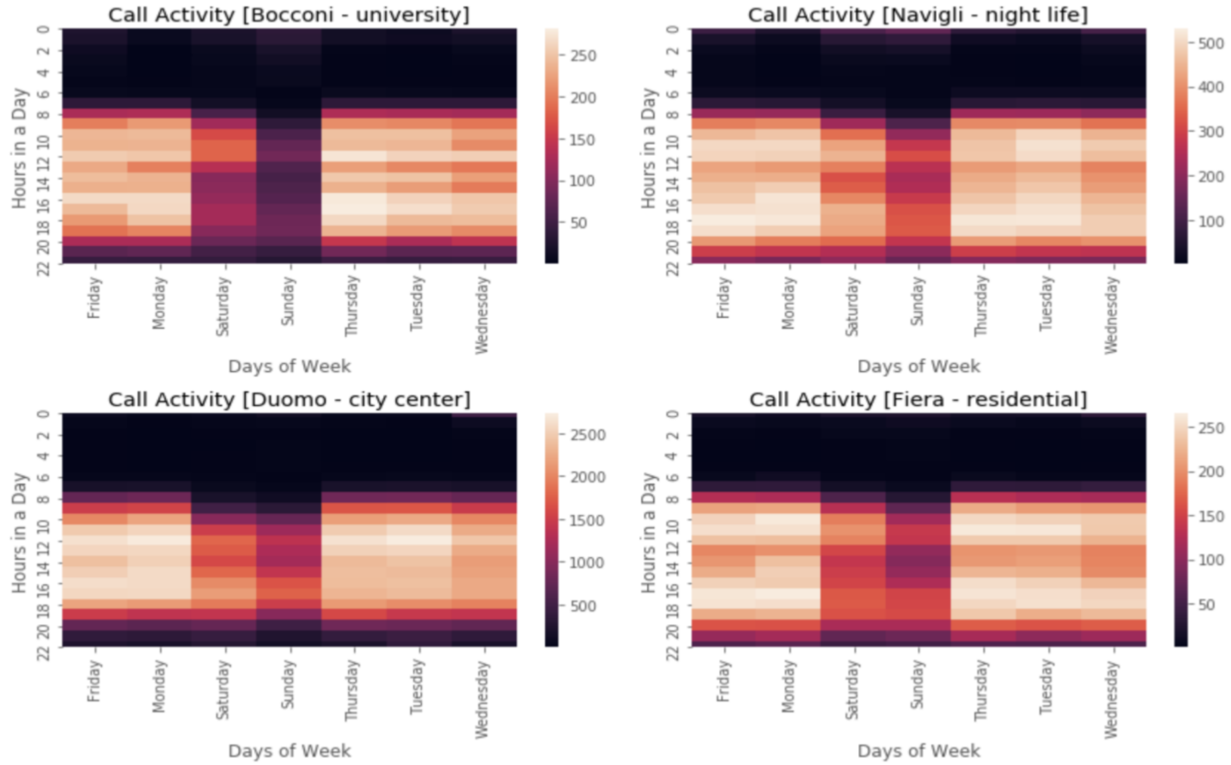


Fig 12: Heat map of hourly Call activity of the four grids

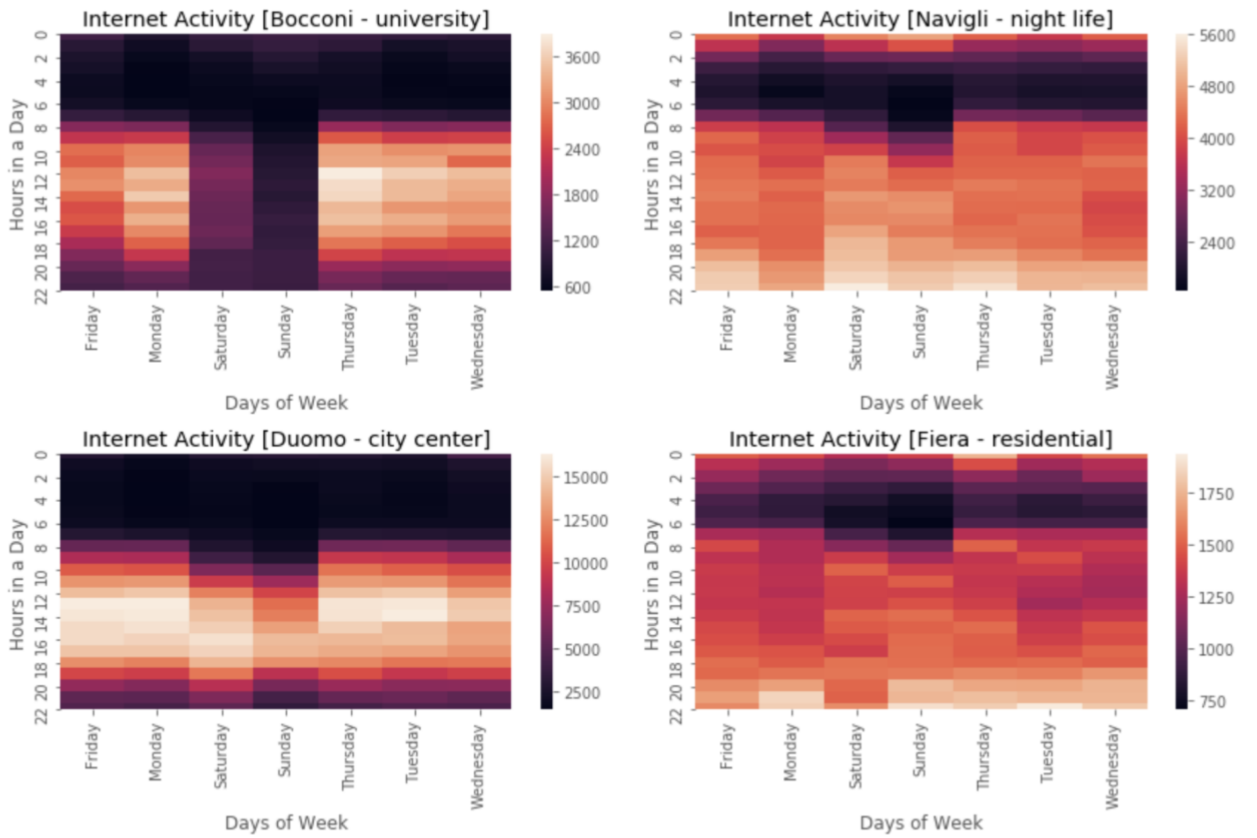


Fig 13: Heat map of hourly telecommunication activities of the four grids

Heat maps shows significant differences in behavior of the four grids,

SMS Activity:

- In general, there is less SMS activities during the weekends (Saturday & Sunday).
- Navigli, night life region shows SMS activities until 2 am in the night on weekend.
- Fiera & Navigli regions are very active from 7am till 10pm on all days.
- Duomo, city center has very less SMS activity from 10pm until 7am in the morning on weekdays and 9 am on weekends.
- Bocconi, university shows less SMS activity compared to others.
- Navigli, Duomo & Fiera shows a sudden bright region on Wednesday 12am, this must be due to New Eve falling on Tuesday.

Call Activity:

- Call activity has similar pattern as SMS, but lesser volumes.
- Surprisingly, there is no significant call activity on New Year eve. This shows how people are more connected via SMS and internet these days. Another possibility is that calls may have been made via internet.

Internet:

- Navigli & Fiera has internet activities almost all through the night. Even Bocconi, university shows some sparse activity after midnight on weekends.
- Duomo, city center although has the highest internet volumes, shows a steady pattern for all activities, 8am – 10pm on weekdays and 10am to 10pm on weekends.

5. CITATIONS

- Barlacchi, G. *et al.* A multi-source dataset of urban life in the city of Milan and the Province of Trentino. *Sci. Data*2:150055 doi: 10.1038/sdata.2015.55 (2015).
- Telecom Italia, 2015, "Telecommunications - SMS, Call, Internet - MI", https://doi.org/10.7910/DVN/EGZHFV, Harvard Dataverse, V1
- Telecom Italia, 2015, "Milano Grid", https://doi.org/10.7910/DVN/QJWLFU, Harvard Dataverse, V1