2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT)

Can Temperature Be Used as a Predictor of Data Traffic: A Real Network Big Data Analysis

Muhammad Nauman Rafiq*, Hasan Farooq*, Ahmed Zoha¶ and Ali Imran*
*University of Oklahoma, Tulsa, USA 74135,
¶5GIC, University of Surrey, Guildford, United Kingdom GU2 7XH
{nauman.rafiq, hasan.farooq, ali.imran}@ou.edu*, a.zoha@surrey.ac.uk¶

***Abstract*— The proliferation of mobile devices and big data has made it possible to understand the human movements and forecasts of precise and intelligent short and long-term data consumption of services like call, sms, or internet data which has interesting and promising applications in modern cellular networks. Human nature and moods are known to be synonymous with the physical attributes of mother nature such as temperature. The change in those physical features affects the human routines and activities such as cellular data consumptions. The future of telecommunication lies in the exploration of heap of information and data available to companies and inferring the valuable results through extensive analysis. In this paper, we analyze three main traits of cellular activity: sms, call, and internet. This paper investigates whether the relationship between the temperature and the cellular data consumption exits or not. This work introduces a novel approach to identify the strength of relationship between the temperature and cellular activity (sms, call, internet) and discuss the methods to quantify the relationship using correlation method. The real network CDR big data set – Milano Grid data set is used to analyze the behavior of the cellular activity with respect to temperature.**

***Index Terms*— CDR, Big Data, Cellular Activity, Correlation, Data Traffic, Milano Data Set, Temperature**

I. INTRODUCTION

In the world, we live in today; the evolution in technology is heavily dependent on two major phenomena: optimization and efficiency. Whether the optimization is related to cost, resources, usage, etc. and efficiency towards energy consumption, quality or implementation, both play pivotal roles when it comes to designing or planning any aspect of life. In the past few years, the global mobile communication industry has been growing rapidly. Today, virtually every single human being owns a cellular device of some sort or takes part in using mobile communications. This pervasive growth of users resulted in demand of more data resources such as cellular networks with high quality and uninterrupted service, but also lead to high energy consumption and costs for operators or providers. It leads to major issues like global warming and heightened concerns for the environment. This calls for the extensive study and analysis of various factors affecting and deployment of the resources or services with utmost efficiency and with highly optimized infrastructure. In the past two decades, there has been a lot of work done in various capacities such as energy efficiency, cell planning, load balancing, reliability analysis and fault prediction and spatiotemporal mobility prediction to not only improve the cost reduction processes but also to cope with the high demand and load on mobile networks.

The proliferation of mobile devices such as cellphones makes it possible to understand the human movements and predictions of short and long-term data consumption. Precise and intelligent forecast of user behavior for the consumption of services like call, sms, or internet data has interesting and promising applications in modern cellular networks, including resource optimization, infrastructure planning, real-time network service provisioning, new mechanism evaluations and many more. It is very important to understand and analyze the user consumption data to deploy the resources in such a manner that the maximization and quality of the service does not lessen. Investigating the spatiotemporal variations of a mobile user and their associated traffic demands is one of many ways to analyze the performance of the mobile networks.

II. RELATED WORK

Previously, many research works have been done to model the real people mobility. Several methods have been used to identify a mobile device's physical location, such as cell tower look up, Wi-Fi access point look up, triangulation of cell tower and Wi-Fi access point, and Global Positioning System (GPS), with varying accuracy levels and power consumptions. One body of existing modeling work relies on logging locations via these techniques on the phones. They are limited to a small group of participating users and a small geographic area such as a university campus [1], [2]. Other attempts use the Call Data Records (CDRs) collected by cellular network operators to deduce models for large geographic areas with diverse populations. CDRs contain the identity of the cell tower the phone was associated with when a voice call was placed or a text message was received. By correlating the cell IDs with the geographic locations of the towers, the CDR based methods can accurately capture many aspects of human mobility. These efforts perform statistical analysis of CDRs such as distribution fitting for call frequency, call duration, etc. [3], [4], [5], [6], [7], [8]. Some emphasizes on the importance of temporal patterns of communications between people as the evolution of communities and social groups [9], other uses the CDR data to map real population estimates in time and space [10], or explore patterns of international/migrant communities in a global city [11].

Contrary to previous works, this paper aims to highlight the mobile network's performance breakdown by creating a range

978-1-5386-5502-3/18/$31.00 ©2018 IEEE
DOI 10.1109/BDCAT.2018.00028

of spatial – temporal statistical analysis of voice, SMS and data (internet) patterns. The analysis of real mobile network CDRs are performed to investigate the statistical behavior of features available in the data set which, in the future, can be used to develop traffic models that will act as independent modules with SON algorithms evaluations. Once the patterns for different services such as call, SMS and internet are established with respect to the physical attribute such that temperature, the assessment and evaluation of behavior of each CDR type are performed to determine the level of confidence which represents the trend. Then those correlated dependent variables with independent variables like temperature, are used to study the impact of those variables on user behavior. For example, intuitively, if a few areas can be clustered together in same zone depending on the density of the point of interest in the region, the analysis can show that with change in number of point of interests or change in weather can progressively or negatively affect user behavior and data consumption. This study can help to understand how does the cellular network performance can degrade or improve with change in physical factors.

The primary data set used is aggregated spatiotemporal CDRs, open-public "Telecommunications - SMS, Call, Internet – Milano" [13] provided by Telecom Italia Mobile. The need to understand nature of the user's data consumption and change in the pattern with respect to any physical phenomenon such as temperature or point of interest is a backbone of any intelligent system and very important because of its impact on future systems in terms of learning as explained in [12]. In Self-Organized Networks (SON), the information from all different network layers is going to be the first to go to the SON engine. Therefore, it is important to understand and analyses the information in most efficient manner to provide the strong base to the system to achieve the desired goals.

III. Methodology & Results

1. *Milano Data Set*

This dataset provides information about the telecommunication activity Call Detail Records (CDRs) over the city of Milan. There are 10-minute intervals for each of the 235m x 235m grid cells (Figure 1). Each activity record consists of the following entries: cell/square ID, time-stamp of 10-minute time slot, country code, incoming and outgoing SMS activity, and incoming and outgoing call activity and internet activity. Per the data set release information, each activity value corresponds to the level of interaction of all the users in the square with the mobile phone network and it is the aggregated single value of the interaction. They do not provide any information regarding how they aggregate the data.

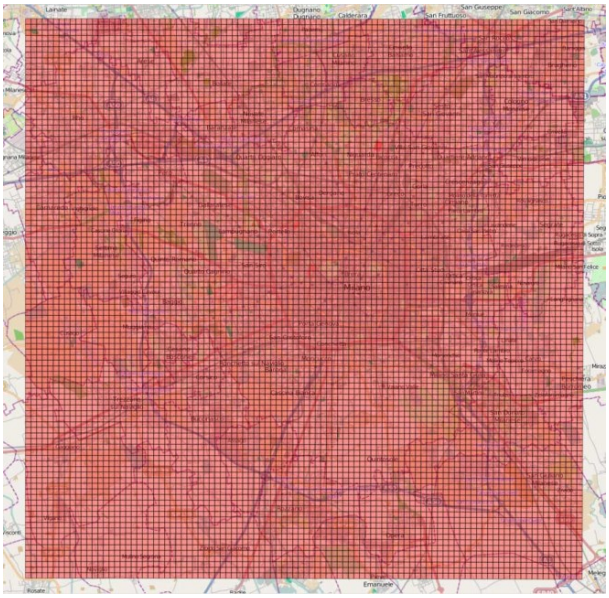


Fig. 1. Milan City Grid 100 x 100

2. *Phase – 01: Data Processing*

Any big data at first glance looks like just a lot of numbers in rows or columns which may not be helpful at all but certainly they can provide most dynamic and crucial information if pre-processed and analyze properly. Therefore, it is vital to understand your data set before jumping to any analysis or extracting any result from it like what type of information there is, what is end goal to achieve from data analysis, how to deploy and implement various techniques to the data to get prolific results etc.

A. *Data Assessment*

Before transforming the data into desired analytical shape, the assessment of data set is important; whether deciding about the missing data or how to replace the missing data. In Milano dataset, there are quite number of missing entries which roughly makes about 5% to 10% of whole data set. It is crucial to understand if activities in dataset are aggregated value of a one complete cell and there is activity happened in previous interval and next time interval, then there is very high possibility and probability that activity must have happened in the cell at missing value time slot. So first step was to replace those missing values with average of previous and next time slot values using:

$$\mathfrak{d}_n^a = \frac{(\mathfrak{d}_{n-1}^a + \mathfrak{d}_{n+1}^a)}{2} \quad (1)$$

where:
$\mathfrak{d}_n^a$ is the missing value of activity 'a' at time slot 'n',
$\mathfrak{d}_{n-1}^a$ is the value of activity 'a' at previous time slot 'n-1',
$\mathfrak{d}_{n+1}^a$ is the value of activity 'a' at next time slot 'n+1',
n is the 10-minute time interval slot,
a is the activity type i.e., call, sms or internet. The missing data was then replaced with averages of previous and next time slot activities using (1).

B. *Data Pre-Processing*

Before implementing the analytical phase, there was need to filter the data out to enhance the output data quality as well as

processing capability. For this outlook, the following filtration steps were considered and implemented to filter the data before heading towards next phase of analysis (i) Replacing the missing data using (1) and (ii) Aggregation by adding together the sms in and sms out, call in and call out columns from dataset. Also, this dataset has separate rows for every activity that occurs to and from each country code. Since our only focus was data pattern and its behavior analysis, for each activity in each cell, therefore the country code column was removed and the data was merged by adding rows to each other to get single value at each time slot in a cell. This gave us 144-time slot values for a single day.

C. *Weather Station Data*

Since, the total area of Milano grid is 23.5 km * 23.5 km or 552.25 km^2, the various weather center locations were identified and weather data was collected from [14]. The acquired data from [14] was only available in 3-hour time slot which means that there was single temperature and precipitation value available for the 3-hour period, for example temperature or precipitation value from 12:00AM – 03AM or 06PM – 09PM etc.

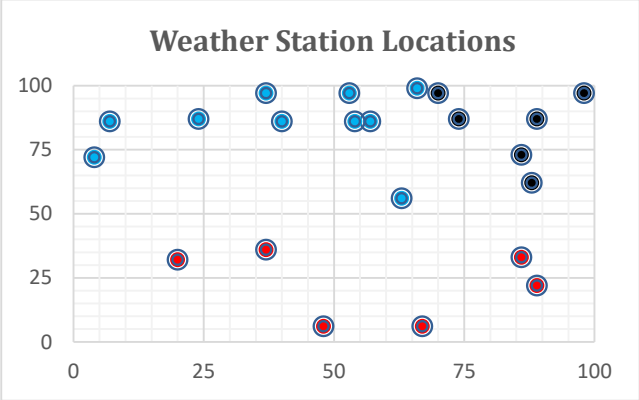


Fig. 2. Weather Center Location Points in Milano Grid

Total of twenty-two weather locations overlapped with Milano grid (Figure 2); each color represents the locations with same temperature and precipitation values for given time frame of the day. It is important to note that due to proximity of these weather locations from each other, the difference in temperature was minimal (±1ºC). Regardless, the average of the temperature data collected was taken. This data was later used to correlate the temperature as independent variable versus activity traffic data as dependent variable.

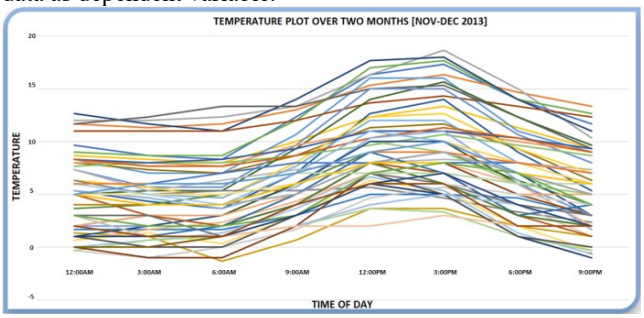


Fig. 3. Two-month (Nov – Dec 2013) Temp. plot over time

D. *Data Post-Processing*

Because we have the temperature data time resolution of 3-hours therefore the filtered Milano data set was further transformed into 3-hour resolution by summing the data traffic of each activity in respective time slots. This eventually shaped the final processed data, ready to perform analysis. The final Milano data now had eight data points against eight temperature time phases for whole day. Figure 3 shows plot of two months temperature data.

3. *Phase – 02: Data Analysis*

Once the processed and filtered data was acquired, in the next phase, the analysis of that filtered data was performed. Initially, the apparent behavior of each telecommunication activity was observed. Figure 4 shows the one-month internet activity of two unique bins in Milano Grid, where top one is internet activity of bin with high number of POIs and bottom one is for grid comprising of mostly rural area. We can clearly see the difference in pattern as urban area activity is not only higher in data traffic usage but also show clear change over the weekend as activity reduces (showed with red dots). On the other hand, the activity usage in rural area is uniform as well as low.

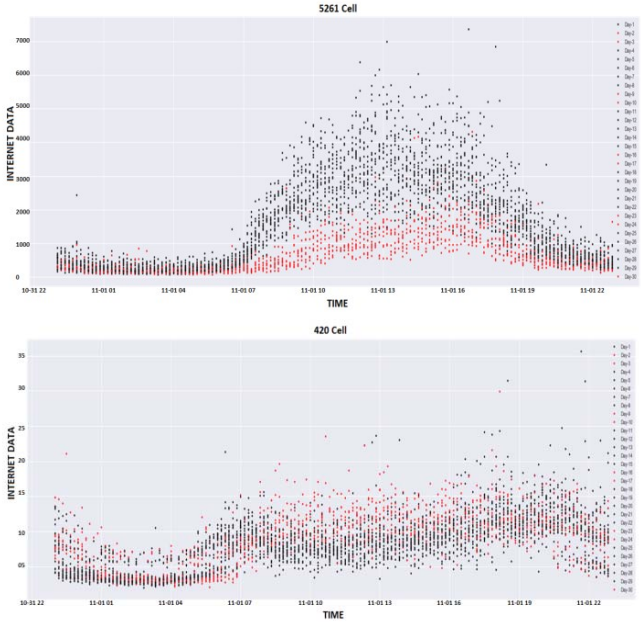


Fig. 4. One Month Internet Activity with top one for Urban and bottom one for Rural area

A. *Zones Defining*

After observing the activity trend and change over the week, it was decided to analyze the data on each day within each time slot basis and observe the correlation of temperature data with the user activity data. To do that, it was important to define an intuitively designed way to cluster the ten thousand bins into zones depending on the nature of areas which attract more activity and changes over the week depending on the location.

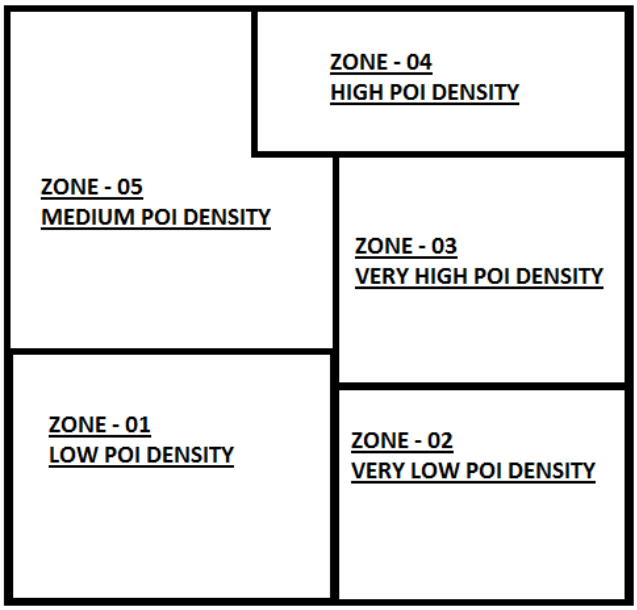


Fig. 5. Zone division based on attractions density in the area

The Milano Grid was further intuitively divided into five zones depending on the density of the Point of Interest (POIs) in the area. The POIs in the area are determined through TripAdvisor website [15]. The Very Low POI Density area refers to the very low number of attractions in the area roughly less than twenty attractions in the area whereas very high POI density area refers to densely populated area in term of attraction roughly more than hundred and fifty attractions in the area. Figure 5 shows the zones from very low to very high POI densities. The main city of Milan which mainly lies in zone – 03 have highest number of POIs.

We have eight-time slots in each day, the correlation between dependent and independent attributes were calculated using spearman correlation method in each time slot. The correlation run between the CDR data and temperature data was performed for complete Milano grid that is 10000 bins and it is well known that each bin has its own behavioral properties depending on the nature of infrastructure in the area for example some bins have only residential locations, some have both residential and commercial buildings and others have only commercial parts which may or may not be considered as attractions. Therefore, the intuitively divided zones contains a mixture of all types of bins ranging from residential to commercial to attraction infested areas for tourist and in those bins the telecommunication activity also changes as per location. It is not feasible and to process and analyze each bin separately to understand the collective behavior of telecommunication usage in the region.

B. Positive and Negative Correlation

As mentioned, due to different urban, suburban and rural setting in the area, the correlation between the temperature and data will be different in each bin for different time slots on different days of week. For example, by looking at some sample bins in different zones, the intuition is clear that if the bin is heavily populated by the residential housing the correlation value between temperature and data is most likely be negatively correlated for some time slots.

This led to designate the zones into three different categories at given time and day of week: Positive Correlated Zone, Negative Correlated Zone and Neutral Correlated Zone. The bins behave differently at different times on different days. If a at given time and day, the number of positively correlated bins in the zone are more than 60% then the zone is designated as PCZ (Positively Correlated Zone) and will be treated as such and vice versa for negative correlation i.e., NCZ (Negatively Correlated Zone). If the percentage of bins with positive and negative correlation lies between 40% - 60%, the zone will be designated as NZ (Neutral Zone) which was influenced by bins with both positive and negative correlation at that given time and day of week.

Figure 6 and Figure 7 show the heat maps which illustrate the change in correlation pattern of Internet activity of complete Milano grid at two-time slots; 09AM-12PM and 03PM-06PM for a whole week and how it transient from positively correlated area to negatively correlated area with change in day of week. From those figures, we can infer that the change from positive correlated zones to negatively correlated zones does happen at different days for these time slots. There are patches with high correlation in both directions at different days. We will discuss how well the temperature is correlated with data in next subsection but for now it is important to identify the areas which can be designated as different zones depending on the user's data consumption behave with increase or decrease in temperature.

Figure 8 and Figure 9 show the heat maps of divided five zones and how those zones change from positively correlated zones to either neutral or negatively correlated zones with change in time slot and day of week. In the heat map, 60% and above heat signatures corresponds to the positive correlated zone at that particular time and day, whereas between 40% and 60% heat signature refers to the neutral zones and for below 40% heat signature, the zone was considered as negatively correlated zone for that specific time and day.

All the figures from Figure 6 to Figure 9 are of same time and days to see the pattern more intuitively and it helps to explain the results in further discussion. The temperature does have correlation with data activity (sms, call, internet), but it may be significant for some or may not be for others depending on the nature of point of view. For example, if an operator wants to see how the trend (either positive, negative or neutral) of data activity in a zone will be with change in temperature on a given day at a particular time slot, this information can give insight with good level of prediction. The resources can then be deployed accordingly.

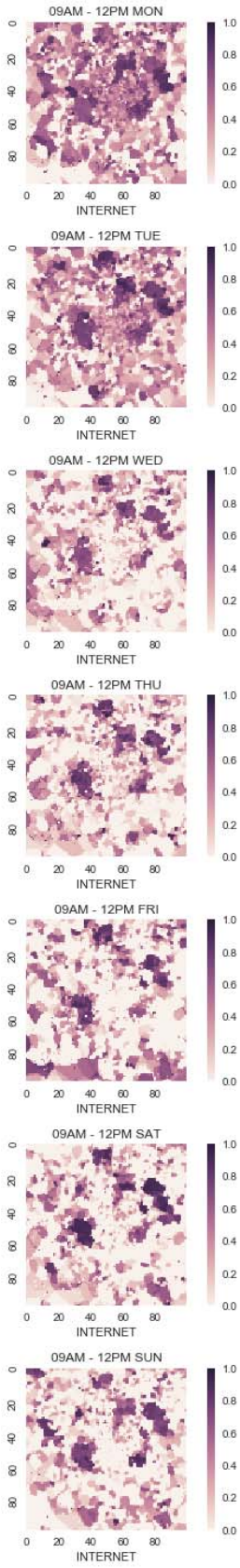
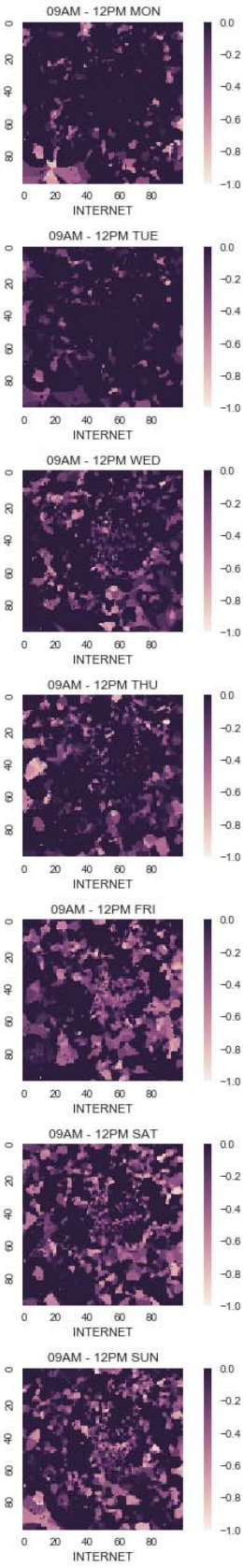
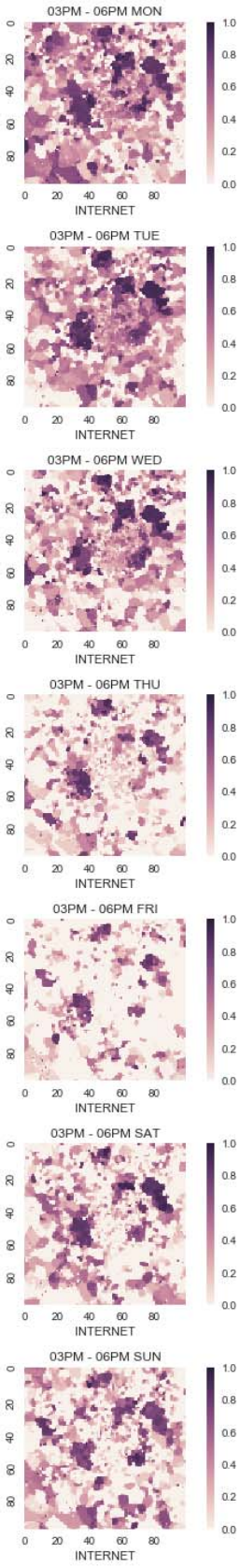


Fig. 6. Milano Grid Heat map: Change in correlation pattern over the week at time slot 09AM – 12PM





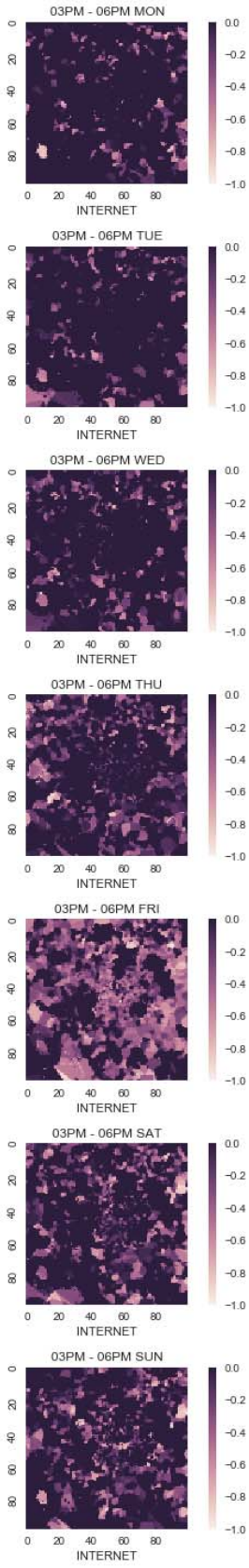


Fig. 7. Milano Grid Heat map: Change in correlation pattern over the week at time slot 03PM – 06PM

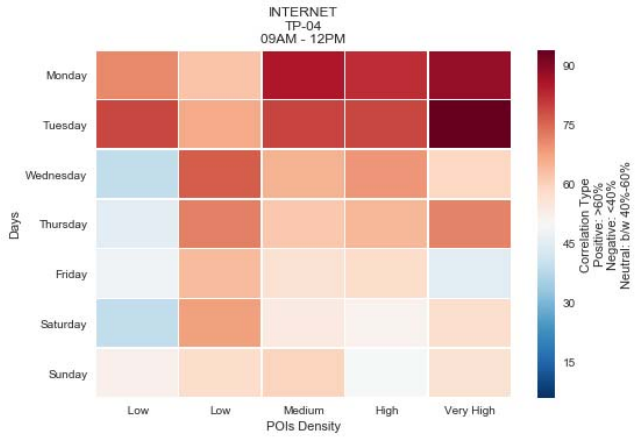


Fig. 8. Zones Heat map: Change in correlation pattern over the week at time slot 09AM – 12PM

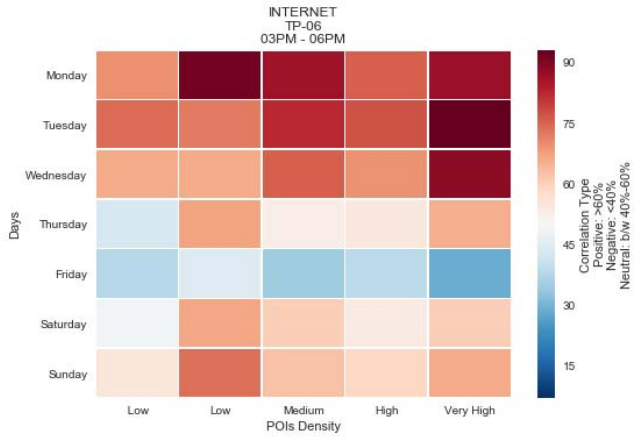


Fig. 9. Zones Heat map: Change in correlation pattern over the week at time slot 03PM – 06PM

4. *Phase – 03: Results*

The Milano grid CDRs data set is two months real data with distinct log entries of data activity for sms, call and internet. We ran spearman correlation for whole data set on all bins in the grid separately for each data activity. In this part, we will discuss the result of correlation between each data activity and temperature and infer some insights which can be valuable for the future work. In many previous works, researchers mostly focused on the call activity or considered sms; call and internet activities in same context but eventually these all have different settings and can be impacted differently when temperature element comes into factor. We will discuss and show results for internet activity and see whether which activity correlate more with temperature and if so then on what times or days and whether this correlation is in positive trend or otherwise.

We will discuss that with change in density of point of interest the correlation between these two dependent and independent variables does get affected and what is the extent of impact over the week at different time slots. In global village world, cellular data activity has most viscous impact and the demand is increasing everyday whether it is in terms of speed, latency or capacity. The activity changes with change in location but never diminishes, whether in public places or residential area. Milan city is famous for its many point of interest around it. Figure 10 shows that for most days in a week, we will find average highest positive correlation of the zone with largest POI density area which in this case is Zone-03 and least is zone-02 with lowest POI density area whereas same zones have opposite results when we look at the negative correlation. The correlation values are the average of all positive and negative correlation values of bins in that zone at that time and day.

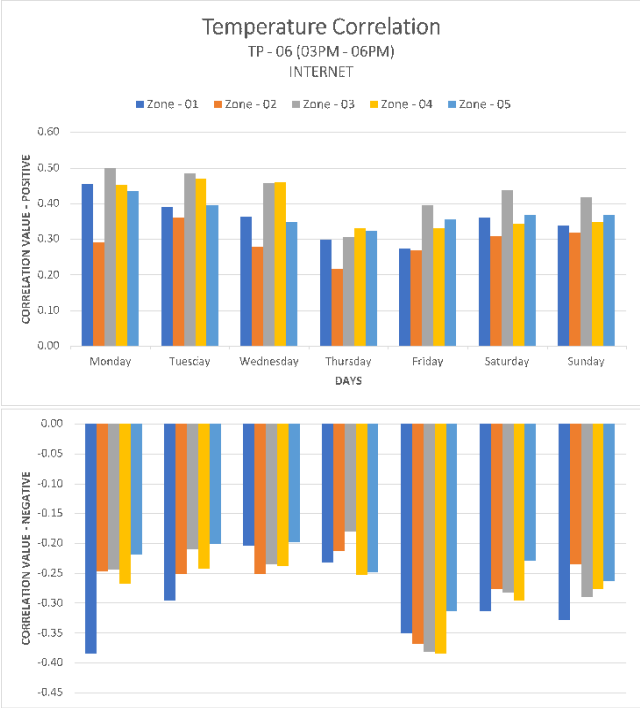


Fig. 10. Day wise average +VE and -VE correlation values of zones over week at time slot 03AM – 06PM

Since, the Milano gird is mashup of both residential and commercial establishment therefore Figure 8 and Figure 9 give us idea which zone has positive or negative correlated trend dominance and Figure 10 gives us the idea of accuracy of correlation for the given day and time. Figure 11 also provide same information but also validate the key important take away that the higher the POI densities in the area mean higher the correlation between the data activity and the temperature. In Figure 11, zone – 03 has the highest number of POIs in the area, the trend line shows that it has highest positive correlation value as well smallest negative correlation value.

Similarly, the negative correlation plot for both figures gave us insight about how the negative correlation decrease with increase in the POI density in the area. As the number of point of interest increased in the area such as zone – 03 results in high correlation between the temperature and the data activities (sms, call, internet) whereas a decrease in negative correlation between both variables.

With this knowledge, an operator not only can decide that what correlation category is dominant in the area/zone at a given time and day but also can also predict the data traffic with level of confidence as shown in figures. Since, the correlation

values in charts are average of bins in zones therefore the level of predictability confidence will vary.

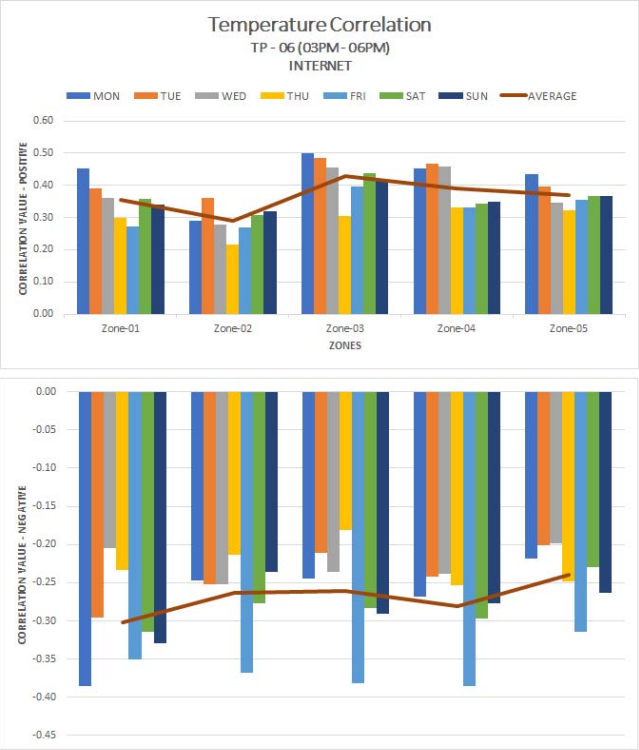


Fig. 11. Zone wise average +VE and -VE correlation values of zones over week at time slot 03AM – 06PM

IV. CONCLUSION

Evolution in the telecommunication industry is eminent and demanding due to vast user activity and the need to connect to the world. This research work aims to highlight the big data analysis and explore new avenues to gain access to the information which, in the future, can be used to develop traffic models that will act as independent modules with SON algorithms, their evaluations, and performances. In this paper, a novel approach is presented to analyze the correlation between Milano Grid CDR (two months data set) and temperature. Analysis show that Temperature being an independent variable and real data set as a dependent variable, the relationship among these two does exist. Milan is a city of mixed geographical settings for both residential and commercial locations. The Milano gird was divided into zones based on its point of interest's density in the area, very high POI density area zone – 03 to very low-density area zone – 02. We observe three categories in terms of correlation: positive, negative and neutral. Intuitively, those bins in the grid, which are mostly commercial point of interest, have a positive correlation at a given time and day, but with residential places, dominant areas have negative correlation with temperature. These correlation categories also change and are dependent on time of day and the day itself for any bin. As the POI density increases, the correlation between the temperature and data activity also increases. We also observed that as POI density change from low to high, it results in high correlation between data activities (sms, call, internet) and temperature.

The impact of data analytics is huge and there are virtually endless possibilities to assess and manipulate the data to gain and infer the information and results tailored to the need of any research. Future work includes zoning or clustering of the grid which can be more insightful for the data analysis approach. More extensively defined clusters based on areas in various settings, such as urban or rural, commercial, or residential, can provide much astute information which can be very helpful in big data empowered SONs.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant Numbers 1619346, 1559483, 1730650 and 1718956. The statements made herein are solely the responsibility of the authors.

REFERENCES

[1] M. Kim, D. Kotz, and S. Kim, "Extracting a mobility model from real user traces," in Proc. IEEE INFOCOM, (Barcelona, Spain), IEEE Computer Society Press, April 2006.
[2] M. A. Bayir, M. Demirbas, and N. Eagle, "Discovering spatiotemporal mobility profiles of cellphone users," in 10th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, WOWMOM 2009, Kos Island, Greece, 15-19 June, 2009, pp. 1-9, 2009.
[3] S. Ščepanovi ´ c, I. Mishkovski, P. Hui, J. K. Nurminen, and A. Ylä-Jääski,"Mobile phone call data as a regional socio-economic proxy indicato," PloS One, 10:e0124160, 2015.
[4] Y. Leo, A. Busson, C. Sarraute and E. Fleury,"Call detail records to characterize usages and mobility events of phone users," Comput. Commun. 95, 43–53 (2016).
[5] F. Gustafson and M. Lindahl, "Evaluation of Statistical Distributions for VoIP Traffic Modeling", Degree Project T-08-117, 2008.
[6] P. Vaz de Melo, L. Akoglu, C. Faloutsos, and A. Loureiro, "Surprising patterns for the call duration distribution of mobile phone users," KDD'10, pp. 354–369, 2010.
[7] H. Zang and J. Bolot, "Mining call and mobility data to improve paging efficiency in cellular networks.," in MOBICOM, pp. 123-134, 2007.
[8] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," Nature, vol. 453, pp. 779-782, June 2008.
[9] F. Botta and C. Genio, "Analysis of the communities of an urban mobile phone network,".PLoS ONE 12(3): e0174198, 2017.
[10] R.W. Douglass, D.A. Meyer, M. Ram, D. Rideout and D. Song, "High resolution population estimates from telecommunications data," EPJ Data Sci 4:4, 2015.
[11] P. Bajardi, M. Delfino and A. Panisson, "Unveiling patterns of international communities in a global city using mobile phone data," EPJ Data Science, Volume 4, Number 1, Page 1, 2015.
[12] A. Imran, A. Zoha and A. Abu-Dayya, "Challenges in 5G: how to empower SON with big data for enabling 5G," IEEE Network, vol.28, no.6, pp.27,33, Nov.-Dec. 2014
[13] https://dandelion.eu/datamine/open-big-data/
[14] https://www.worldweatheronline.com/lombardia-weather/it.aspx
[15] https://www.tripadvisor.com/Attractions-g187849-Activities-Milan_Lombardy.html#ATTRACTION_SORT_WRAPPER