DAZIO: Detecting Activity Zones based on Input/Output call and SMS activity

Documento de Discusión CIUP

DD1620

2016

Miguel Núñez del Prado
Profesor e investigador del CIUP
m.nunezdelpradoc@up.edu.pe

Ana Luna
Profesor e investigador del CIUP

Romain Gauthier
Instersec Labs

DAZIO: Detecting Activity Zones based on Input/Output call and SMS activity

Miguel Nuñez-del-Prado-Cortez
Universidad del Pacífico
Av. Salaverry 2020
Lima - Perú
m.nunezdelpradoc@up.edu.pe

Ana Luna
Universidad del Pacífico
Av. Salaverry 2020
Lima - Perú
ae.lunaa@up.edu.pe

Romain Gauthier
Instersec Labs
París - France
romain.gauthier@intersec.com

Abstract

Mobile telecoms operators possess an enormous quantity of data, which could be used to reduce the cost of installing new infrastructure, to provide a better QoS or to plan their infrastructure. Thus, they are concerned to model, understand and predict SMS and calls activity levels in their infrastructures. Besides, SMS and call activities analysis can open new business opportunities for geomarketing as well as trade area analysis. In the present effort, we detected activity zones with a difference of only 0.5 km from the reference activity areas extracted from Geo-tweets. We also used Markov chains to represent and predict SMS and call activity levels, achieving a prediction success rate between 80% and 90%.

1 INTRODUCTION

Telecoms data is a rich information source for many purposes, ranging from urban planning (Toole et al., 2012), human mobility patterns (Ficek and Kencl, 2012; Gambs et al., 2011), points of interest detection (Vieira et al., 2010), epidemic spread modeling (Lima et al., 2013), community detection (Morales et al., 2013), disaster planning (Pulse, 2013) and social interactions (Eagle et al., 2013).

One common effort for these applications is to determine dense areas where many users stay for a significant amount of time, namely *activity zones*. Another task is to identify contiguous zones relaying *activity zones* (*i.e., transit zones*). Thus, in our context, the detection of *activity* and *transit zones* as well as the interaction between identified *activity zones* are crucial tasks for reducing the cost of installing new infrastructure, to provide a better QoS or to plan their infrastructure.

Therefore, in the present paper, we will identify *activity* and *transit* zones to monitor and to predict the activity levels in the telecoms operators network. These monitoring and prediction are based on the SMS and calls input/output activity levels issued from the Telecoms Italia Big Data Challenge[1]. The results of the present study is directly applied for: (1) targeting advertisement to *activity zones*; (2) proposing a suitable place to open a new store in a city or (3) planning where to add cell towers to improve QoS.

In the present effort, we describe a methodology to detect activity and transit zones. More precisely, the contribution of this work is twofold. On one hand, we present an Activity Markov chain model to represent activity levels. On the other hand, we predict future activity levels using the aforementioned model. The rest of the paper is organized as follows. First, Section 2 describes the related works on *activity zones* detection. Then, Section 3 presents the datasets we use for experiments. Next, Section 4 introduces our technique to detect and model *activity zones* as well as the approach to forecast activity levels. Section 5 shows correlation measurements of activity levels versus pollution and weather conditions. Finally, Section 6 concludes the paper and depicts some future directions.

2 RELATED WORK

Dense areas detection has been studied from a human mobility point of view, using fine grain and coarse-grained location data. As an example of fine-grained location, the work of (Gambs et al., 2011) use mobility traces of 172 Yellow Cabs

[1]Telecoms Italia Big Data Challenge website: www.telecomitalia.com/tit/en/bigdatachallenge.html

Taxis, issued from GPS, in San Francisco Bay (Piorkowski et al., 2009) to detect taxi's point of interests (POI). These POIs are equivalent to *activity zones*, which tend to be zones with high pedestrian presence. The authors rely on the begin-end heuristic (Gambs et al., 2010) and clustering algorithms, such as Density Joinable (Zhou et al., 2004), Density Time (Hariharan and Toyama, 2004) and Time Density clustering (Gambs et al., 2010) to detect POIs in San Francisco city. Another fine-grain data used to detect activity areas are issued from Geo-social networks.

The work of (Qu and Zhang, 2013) uses Foursquare's check-ins from 446 users during ten months for identifying trade areas. They rely on four different techniques like *Center of Mass* location, the *most commonly checked-in location*, the *place with the highest check-in density* and the *center of mass of the most frequently visited location cluster*. The algorithm use for clustering is DBSCAN (Ester et al., 1996). Once they have identified the activity centers, they mark the boundary of the area using *drive-time/distance polygon* (Kures and Pinkovitz, 2011). This technique consists of computing the decay distance from a given store to home or work (authors assume that the two most checked-in places are home and work). The drawback of this approach is that the selected users are conditioned to check-in in the store under study. Thus, the dataset is biased.

Other works use coarse grain location from Call Data Records (CDR). For instance, the work of (Isaacman et al., 2011) uses Hartigan's leader clustering algorithm (Hartigan, 1975) to identify dense areas. First, authors sort antennas by the amount of time that phones contact the antenna. Once data is sorted, the clustering algorithm takes the first antenna as the centroid of a cluster. Then, it verifies if the next antenna is within a distance d from the centroid. If it is not the case, the antenna becomes the centroid of a new cluster. In the case the antenna is within the distance, the algorithm computes the new centroid as the weighted average. They repeat the process until all antennas belong to a cluster. Researchers use CDR locations of 97 and 71 thousand unique users in Los Angeles and New York cities collected over 2 and a half months as well as 19 volunteers as the ground truth to validate their results. They were able to estimate dense areas with an error of 3 miles compared to the ground truth.

Another, more refined technique to identify dense areas respecting natural tessellation is presented by (Vieira et al., 2010). Authors use CDR locations from calls of one million users during four months over an area of 80 000 km^2. They propose a method composed of three phases: the first step is the *graph construction*, which relies on Delaunay triangulation (Dobkin and Laszlo, 1987). The triangulation algorithm makes connexions (edges) between near antennas (vertex) maximizing the size of the angles of the triangles. Once the graph is built, all edges are weighted by the total activity and by the number of users of both connected antennas. The second phase is the *computation of dense areas* based on a maximum spanning tree build using the Kruskal algorithm (Kruskal, 1956). Taking as input the weighted graph G, the idea behind this algorithm is to find a subgraph of G, which maximizes the density and does not contain any cycle. At last, the *post-processing* phase uses (Shiloach and Vishkin, 1982) algorithm to establish groups of antennas representing dense areas. Thus, the algorithm groups adjacent vertex from previously computed sub-graph to find a set of close vertex (a set of antennas). The authors validate their results empirically based on the subway structure of the region under study. Inspired by these works, we propose a novel ad-hoc methodology to find activity (dense) zones as well as to model and forecast their activity levels using the data provided by the Telecom Italia Big Data Challenge (TIM challenge). We describe this dataset in the next section.

3 DATASET

Datasets provided by the TIM challenge were collected in the cities of Milan and Trento over November and December 2013. Our study only takes into account the dataset gathered from Milan city. Nevertheless, all the analysis and the methodology could be generalized for any city. In the next subsections, we will describe the datasets provided for Milan city. These datasets were used to detect and model the *activity zones* (primary datasets) and to find some correlation between activity and other measures like air quality and weather conditions (auxiliary datasets). It should be noted that space was discretized in a grid, and all measures are normalized to correspond in one square of the grid.

3.1 Main datasets

The datasets of Milan city we use to detect and model the *activity zones* are:

Milan Grid is a geographical segmentation over the city to aggregate the measurements of the other datasets. The area of each square is 55 225 m^2, and it has 10 000 squares in the form of a point (x, y) and the latitude and longitude belonging to this x, y position. An example of the described data as well as the grid over Milan are introduced in Table 1 and Figure 1, respectively.

Point	Latitude	Longitude
x_1, y_1	9.011	45.568

Table 1: Example of the Milan grid data

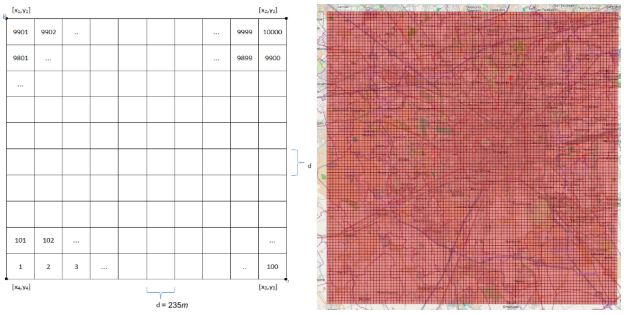

Figure 1: Milan grid

Telecommunications (SMS, Call and Internet) provides information about the activity of a square concerning received and sent SMS, incoming and outcoming calls as well as internet usage. This data is temporal aggregated in timeslots of ten minutes and provides the measure of the activity of a given event as well as the square id (*c.f.* Table 2). This kind of information is organized in a way that SMS-in and SMS-out activity scale are given in arbitrary units, and their values range from 0 to 1.

Id	Time	Country	SMS-in
1	1383265200	39	0.24
SMS-out	Call-in	Call-out	Data
0.16	0.108	0.026	6.83

Table 2: Example of activity data in Milan

Private Transportation (Cobra Telematics) gives information about the private mobility in Milan city by measuring the speed, the number of vehicles that belongs to Milan, the number of vehicles that is not from Milan, the number vehicles with engine ignition systems, in movement and stopped (*c.f.* Table 3).

Id	Time	Direction
60	17/12/13 18:00	WEST
Avg speed	Std speed	Mi plates
24	95	21
Non-Mi plates	Ignition	Mov/Stopped
62	2	2/0

Table 3: Example of private mobility data in Milan

3.2 Auxiliary datasets

We also used additional datasets to analyze *activity zones*, dynamics, and correlations, as we show in the following paragraphs.

Telecommunications - MI to MI provides information regarding the directional interaction strength, between the city of Milan and different areas based on the calls exchanged between Telecom Italia Mobile users. More precisely, this dataset contains the origin and destination Id squares, the time and the directional interaction strength *i.e.,* Activity (*c.f.* Table 4)

Id 1	Id 2	Time	Activity
1	3	1383345474	0.24

Table 4: Example of activity data between Milan zones

Precipitation describes the intensity and the precipitation type over the city of Milan. In more detail, the dataset uses a coarse spatial aggregation by dividing Milan city into four quadrants (northeast, northwest, southeast and southwest). The intensity value of the phenomenon is between 0 and 3, the percent of coverage of a given quadrant and the precipitation type between 0 and 2, where 0 means absence of precipitation, 1 is rain and 3 is snow (*c.f.* Table 5).

Time	Id	Intensity	Coverage	Type
201311060220	1	1	45	1
201311060220	2	0	0	0
201311060220	3	0	0	0
201311060220	4	2	78	1

Table 5: Example of activity data between Milan zones

Air Quality describes the air pollution monitoring system of Milan city obtained by using various types of sensors located within the city limits. This environmental dataset measures a different kind of contamination agents, such as Ammonia, Nitrogen Dioxide, Total Nitrogen, Particulate Matter 2.5 μm (PM2.5), Particulate Matter 10 μm (PM10), Benzene, Sulphur Dioxide, Black Carbon, Carbon Monoxide and Ozone. An example of pollution measure is given in Table 6, where the characteristics of this particular sensor are in Table 7.

Sensor id	Time	Measure
5823	2013/12/30 04:00	1.9

Table 6: Example of air quality measure

Sensor id	Lat/Lon	Pollution
5823	45.24/9.27	Carbon Monoxide

Table 7: Description of the sensor 5823

Social Pulse contains data derived from an analysis of geolocalized tweets originated in Milan. This dataset provides a user id, DBPedia entity, tweets language, municipality, time, timestamp and location (*c.f.* Table 8).

User	Entities	Language
5fa4b1cc71	Halloween	En
Municipality	Timestamp	Lat, Lon
Milan	1383260474	9.21, 45.49

Table 8: Example of Social Pulse data

Based on the aforementioned datasets, we have implemented our experiments using the main and the auxiliary datasets. These experiments are detailed in sections 4 and 5, respectively.

4 EXPERIMENTS

In the present section, we describe our methodology to discover, model and predict the behavior of a zone. We distinguish two different areas, the *activity zone*, where people stay on a regular basis for a significant amount of time and the *transit zone* which is the area used by individuals to go from one activity zone to another. In the next subsections, we describe how to recognize an activity zone from a transit zone (Subsection 4.1), how to model activity levels (Subsection 4.2) and finally how to predict them (Subsection 4.3).

4.1 Detecting activity levels

The basic idea behind this method is to have a good representation of the activity variation levels over the time. Activity levels could be classified in three different degrees, low, medium and high. Cumulative distribution of incoming/outcoming SMS and call activity levels illustrated by a Heat map over Milan city, as shown in Figure 2, were used to analyze data. The objective is to, empirically, find a suitable threshold to distinguish a square with high activity level from a square with medium or low activity levels represented as green and red in Figure 3, respectively.

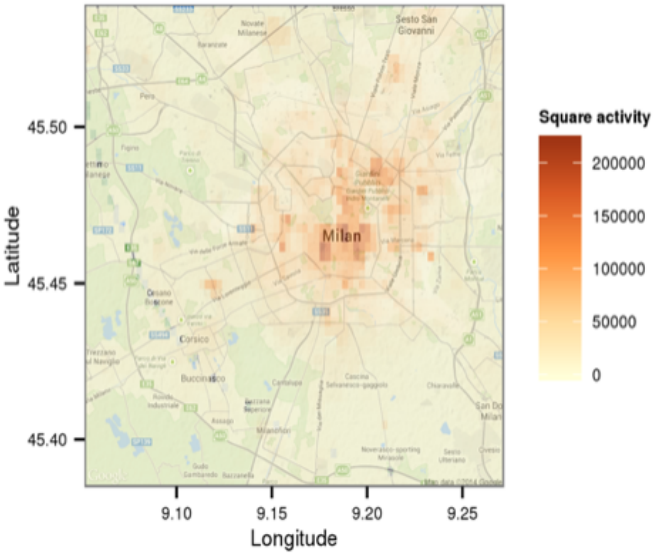

Figure 2: Milan activity heat map

Figure 3 depicts the cumulative distribution of the aggregated incoming and outcoming SMS and call activity of the telecommunications dataset (*c.f.* Subsection 3.1). The Heat map is built for an activity threshold of 25 units. Based on this visualization technique, that amount of units seems to be a good trade-off between compact and well-separated *activity zones*. Heat maps were used to represent represent tourist activity as shown by Olteanu et *al.* (Olteanu et al., 2011).

In order to detect groups of squares representing an activity zone, we can use a high activity threshold (*c.f.* Subsection 4.2). In addition, we study, in detail, the activity over work hours to analyze the difference between busy and idle squares. Figure 4 shows the difference in activity levels between activity and transit zones. From 8 AM to 8 PM the activity is considerably high, that is why that area is composed of a vast number of squares during the day. On the other hand, transit zones display a

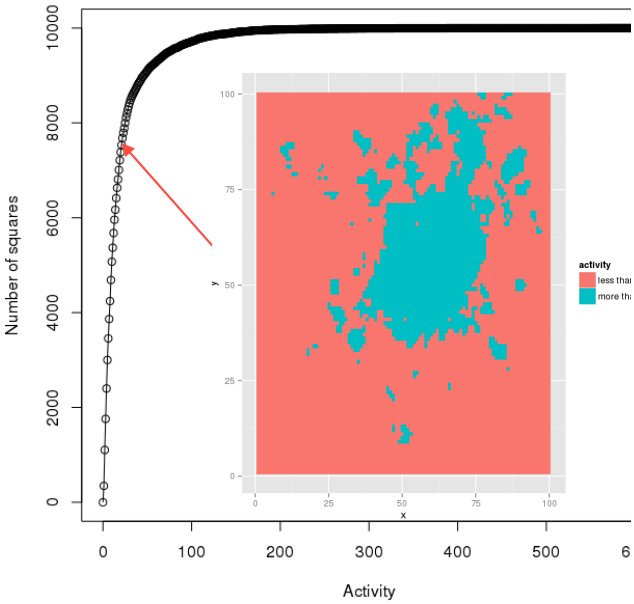

Figure 3: Cumulative distribution of the sms/call activity with heat maps

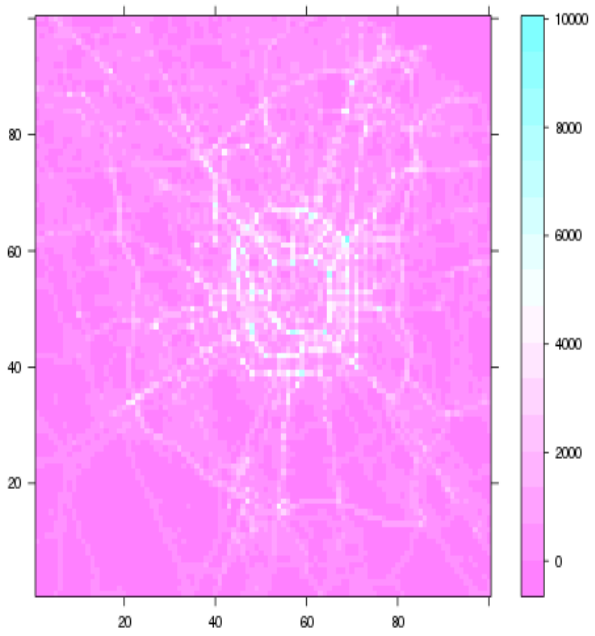

Figure 5: Vehicles speed heat map based on Cobra dataset

much lower activity level and a higher fluctuation throughout the day.

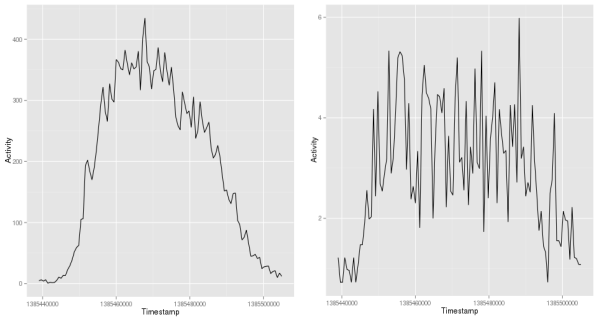

Figure 4: Activity over time in workdays between 8AM-8PM. Activity zone (left) and transit zone (right)

Taking into account the elements as mentioned earlier, we use the Heat map and activity threshold presented in Figure 3 for detecting high *activity zones*. Nevertheless, we need to define the borders of these *activity zones*. From Figure 4, we can infer that this irregularity of transit zone represents movement. Thus, the Cobra dataset gathers the information about the movement of vehicles and we depicted this information in a Heat map over Milan city in Figure 5. The speed combined with the activity level allowed us to detect the *activity zones*, as well as their borders. As the result of the combination of this two variables, we obtained 28 activity areas and the centroids of these *activity zones* which are shown in Figure 6 in blue color.
Since we do not have the ground truth, we used the

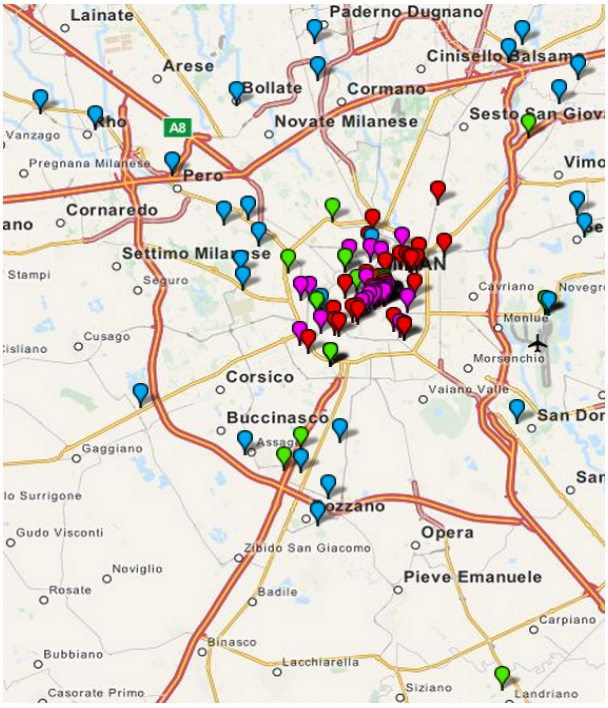

Figure 6: Centroids of the activity zones in Milan city (blue), centroid of the clusters from Geo tweets (green) and commodities issued from Foursquares check-ins shops (red) and restaurants (purple).

geolocalized tweets dataset to verify the accuracy of our methodology. Applying DBSCAN (Ester et al., 1996) clustering algorithm with at least 5 points per cluster within a radius of 3 km over 8 282 users (*i.e.,*109 762 geolocalized tweets). We obtained 24 clusters depicted as green points in

Figure 6. Thus, some groups are close to the identified *activity zones* in the northeast and south. In Northwestern, *activity zones* are represented by only one cluster due to the proximity of geolocalized tweets and the approach of DBSCAN to build clusters. The Downtown area has many clusters due to commodities concentration. To verify this fact, we have included two categories of check-ins from Foursquare, like shops (red) and restaurants (purple). One thing that surprised us was the detection of a large group of geolocated tweets in the southern outskirts of Milans grid. Using these data we found that the distance between the centroids of the clusters and *activity zones* is 0.5 km closer. These results validated the accuracy of our heuristic method to find *activity zones*. Finally, we present the centroid of the detected *activity zones* in Figure 7. These regions are used in the next subsection for predicting purpose.

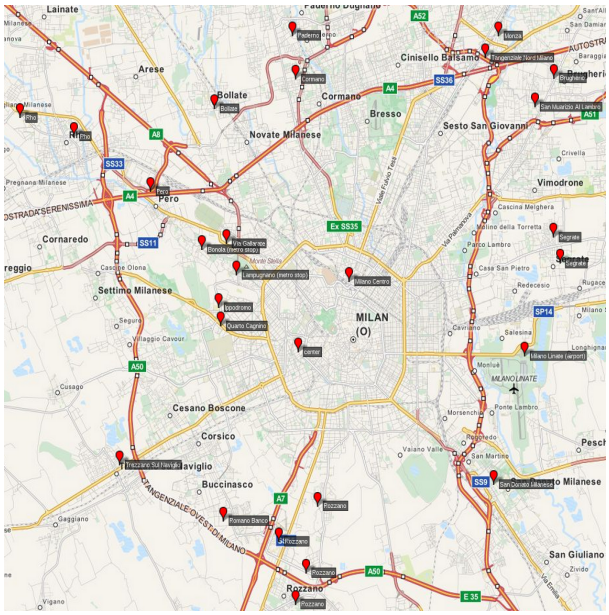

Figure 7: Centroids of the activity zones in Milan.

Up to this point, we are able to identify *activity zones*. Thus, the next tasks are to model the behavior activity levels in the detected regions as well as to predict the activity levels in the identified *activity zones*.

4.2 Modeling activity levels

Relying on the thresholds mentioned above, we can build an *Activity Markov chains* model to represent the change of activity levels over time. An *Activity Markov chain* model is a stochastic process where the changes of states are related to a probability associated with various state changes (transition probabilities). In our case, an *Activity Markov chain* is a probabilistic automaton (PA) model that represents, in a compact way, the occupation (activity) of a square or activity zone. The nodes symbolize the state (low, medium or high) of the squares or zones ([activity level]_[zone code], ex: L_A) and edges, weighted with a probability, represent the transition from one state to another over time windows. This model could be expressed in the form of a graph or the form of a transition matrix (*c.f.* Figure 9).

The process for building an Activity Markov model is divided into two stages. The first one is basically to order the events in a chronological way. Then we classify them as *low*, if they have less than 15 activity units, as *medium* if activity units are between 15 and 25 and as *high* if there are more than 25 activity units. Then the transition matrix is built by counting the variations from one level to another, taking care to avoid loops. When events are not recorded anymore, the matrix is normalized to obtain the transition probabilities.

As shown in Figure 8, we have divided the time into 4 different windows depending on the range of time studied, each one has 6 hours and starts at 6:00 am; for both weekdays and weekends; giving a total number of eight windows. Furthermore, in each time window interactions between different levels of activity are also modeled. Moreover, we matched, after a model processing, an *activity zone* of a time window with another (blue arrows). We finally conclude, from the stationary vector of Markov chains, that activity areas are occupied in only 11% and free in 71%. It is important to point out that, for improving the accuracy of the analysis, it would be better to divide the time slots in less intervals instead of 6 hours.

Until this point we showed that we are able to identify high activity squares, activity and transit zones. In the next subsections, we detail how to predict, in an unprecedented way and with a very acceptable rate of success, not only activity levels but also their possible changes.

4.3 Prediction of activity levels

Anticipating high *activity levels* within an "activity area" allows Telecoms operators to plan or avoid unnecessary investment in infrastructure, as well as ensure the QoS or to start a new entrepreneurship. In this paper, our prediction, inspired from the work of Gambs *et al.* (Gambs et

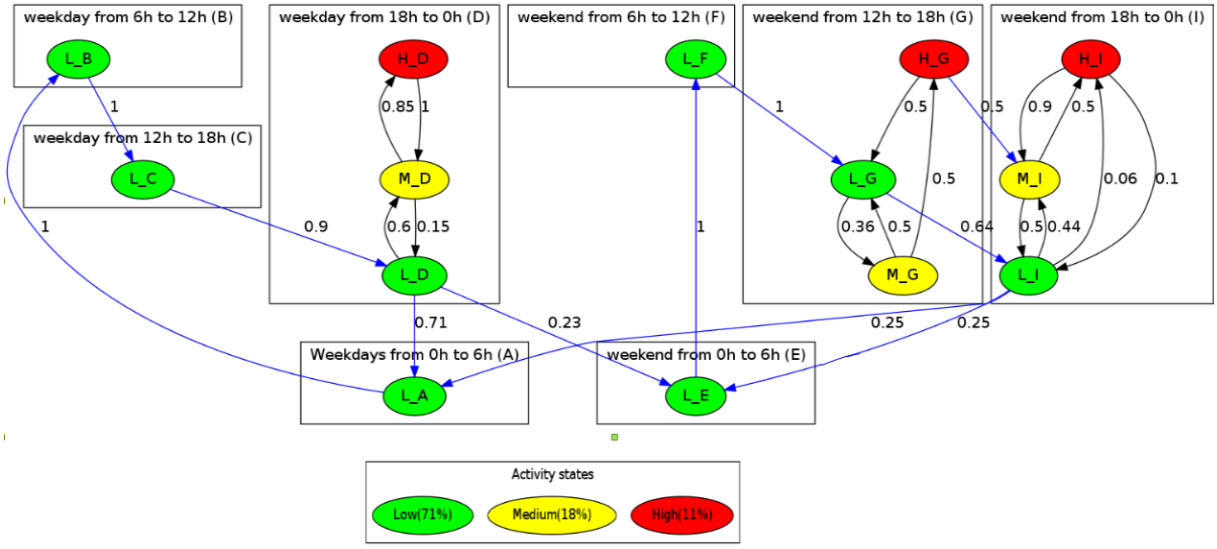

Figure 8: Example of an activity Markov chain (activity over time windows)

al., 2012), is performed using the transition matrix and allows us to obtain changes of activities between temporary windows and within them. Then, predictions of activity changes within the same window and estimations can be made. For example, we wonder what is the probability of passing from a low activity state to a high one or what is the likelihood for a given activity to remain in the same state when we consider the next window and the same range of time. To answer these questions, we rely on the algorithm presented in Algorithm 1.

Algorithm 1: Prediction algorithm

Data: TransitionMatrix, indexRow, inTimeWindows
Result: indexColumn

```
1 if inTimeWindows then
2     //predict next state in the time windows
3     i=maxOutgoingProbOnWin(indexRow)
4     indexColumn = i
5 else
6     //predict the next state when the time
      window change
7     i=maxOutgoingProbOnNextWin(indexRow)
8     indexColumn = i
9 return indexColumn
```

More precisely, Algorithm 1 takes as input a transition matrix (*transitionMatrix*). Where the index of the row in the transition matrix corresponds to the actual state of the system (*indexRow*) and a boolean value is used to indicate whether the prediction is local (*inTimeWindows*). Based on these inputs, the algorithm returns the maximal outgoing probability from the transition matrix taking into account only columns corresponding to the same time windows of the index row (local transition from line 1 to 4 of the Algorithm 1). For instance, in Figure 8, if the actual state is *medium* on the time windows from 18:00h to 0:00h on weekdays, the prediction algorithm will give an output in the *high* level in the same time window. Another kind of prediction is to take into account others columns instead of those that belong to the same time window (inter time windows transition from line 6 to 8 of the Algorithm 1). Given the *low* in the time windows from 18h to 0h on weekdays in Figure 8, the algorithm will output the *low* state on time windows from 0h to 6h on weekends. In the case of the output we have the same probability, ties are break randomly.

	HW0_6	MW0_6	LW0_6	...	HWe18_0	MWe18_0	LWe18_0
HW0_6	0.00	0.50	0.50	...	0.00	0.00	0.00
MW0_6	0.09	0.00	0.91	...	0.00	0.00	0.00
LW0_6	0.03	0.32	0.00	...	0.00	0.00	0.00
HW6_12	0.00	0.00	0.00	...	0.00	0.00	0.00
...	...	...	...	...	...	...	...
HWe12_18	0.00	0.00	0.00	...	1.00	0.00	0.00
HWe18_0	0.00	0.00	0.00	...	0.00	0.93	0.07
MWe18_0	0.00	0.00	0.06	...	0.56	0.00	0.28
LWe18_0	0.00	0.00	0.18	...	0.00	0.73	0.00

Figure 9: Transition matrix example. Where H=high, M=medium, L=low, W=Weekday, We=Weekend and begin_end hours. As a way of example we show two temporary windows framed with purple numbers inside.

To validate the accuracy of the predictions, we used data from the whole month of November as training set and the first 16 days of December as testing set (we did not take into account New Year

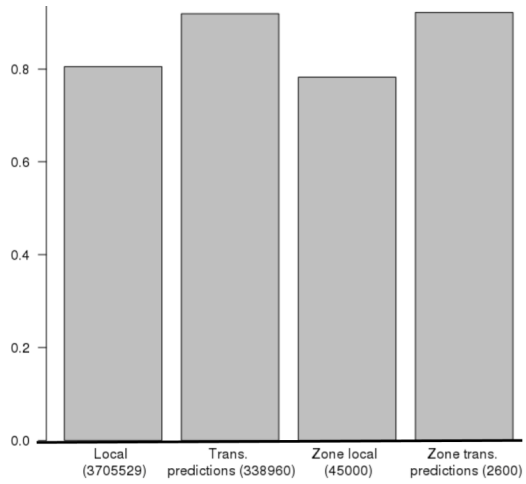

Figure 10: Success rate of prediction. Where local refers to prediction inside a time windows and transition (trans) refers to prediction over time windows.(#) means prediction value.

	square	zone
Local	0.86	0.7
Trans.	0.86	0.89

Table 9: Table summarizing results of Figure 10

celebration to avoid special dates that separate our study from normal behavior). The results are depicted in Figure 10, where the success rate (Equation 1) is the ratio between the correct prediction and the total number of predictions and the number of overall forecast. Note that the number of predictions is indicated in parentheses at the bottom part of each bar.

$$success\ rate = \frac{\#goodprediction}{\#predictions} \quad (1)$$

We observe, from Table 9, the success rate for both kinds of predictions, namely (1) within a time window (*local*) and (2) in different time windows (*trans.*). It is important to note that there are two distinct scenarios; the first one considers both *local* and *trans.* predictions based on *Activity Markov chains* models which were built from *activity levels* of squares; while the second scenario takes into account activity levels from detected *activity zones* to forecast future values. We did not performed k-fold cross validation since the training set of a month is representative of the mobility pattern. Thus, adding more mobility traces to the training test does not contribute to increase the success rate.

So far, we are able to model, identify and predict activity levels in *activity zones*. In the next section, we are going to use the auxiliary datasets to analyze *activity zones* interaction and to study possible correlation with the levels or evaluate how the weather impact the utilization of the Telecoms service.

5 PLAYING WITH OTHER DATASETS

In the present section, we will study the interaction between detected *activity zones* (Subsection 5.1); the correlation between activity levels versus pollution measures (Subsection 5.2) and the influence of the weather on the activity levels in the Telecoms operator infrastructure (Subsection 5.3).

5.1 Interaction between activity zones

Using the directional interaction activity dataset between zones in the area of Milan, we plotted a graph to visualize the communication exchange, as well as various activity levels as we can appreciate in Figure 11, where the width of the edges accounts for the logarithm of the aggregated activity for the whole month of November. To extend the semantic of this graph, we modulated the size of the nodes according to the amount of tweets emitted from the corresponding zone taking into account global pulses dataset (*c.f.* Subsection 3.2).

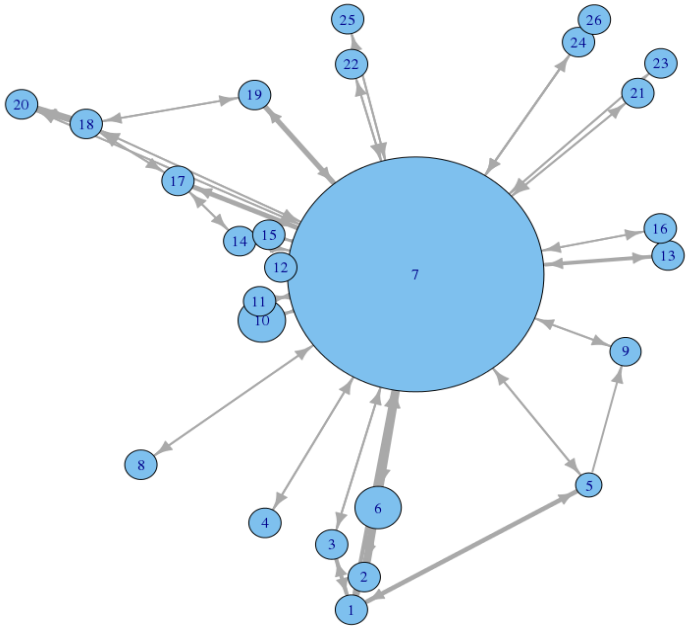

Figure 11: Interaction graph of activity zones.

We observe that Milan city has a star topology, where there is a central node that communicates with the other peripheral nodes. Another interesting fact is that small nodes tend to communicate

to the central node. Nevertheless, there are a few exchanges between small contiguous nodes.

5.2 Forecasting pollution through activity

In this subsection, we study the correlation between the activity level presented on the telecommunication dataset and air quality measurements (both described in Subsection 3.2) to forecast the pollution level of the Milan city based on the activity of the telecom operator antennas. Figure 12 shows the results of the correlation of the activity with respect to different polluting gasses as well as the number of vehicles in movement, ignition or stopped.

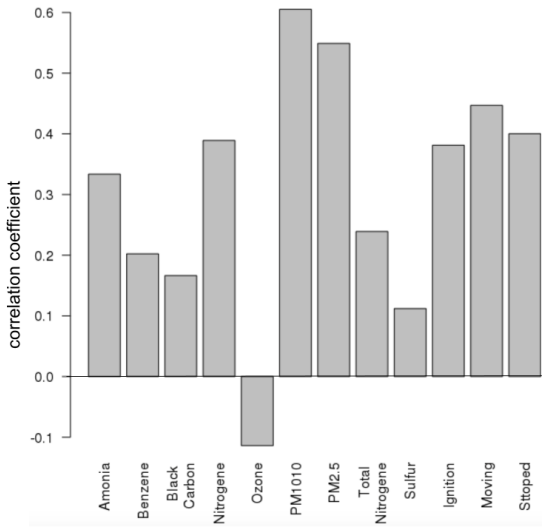

Figure 12: Correlation of the activity with pollution measures and private mobility (Cobra).

We found out that the activity has a positive correlation with PM10 (particulates matter with a diameter of 10 microns or less) and PM2.5 pollution measures (fine particles of 2.5 micrometers of diameter or less). From Figure 13 we can visualize that the activity levels has a positive correlation with the radiation measures and a negative correlation with the relative humidity.

5.3 Influence of weather on the activity

We compare the outgoing SMS and call activity in the presence of different weather phenomena's scenarios, like rain, snow or the absence of both. For this purpose, we used the Precipitation dataset (*c.f.* Subsection 3.2).

From Figure 14, we can observe that people send more SMS and give more calls in presence of rainy weather, even if rain is slight (blue, red, yellow and green bars) than in normal conditions (dark,

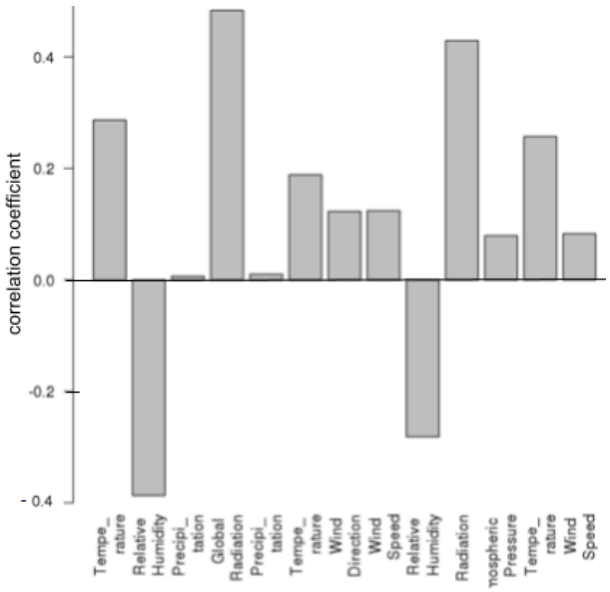

Figure 13: Correlation of the activity of a square with radiation sensors.

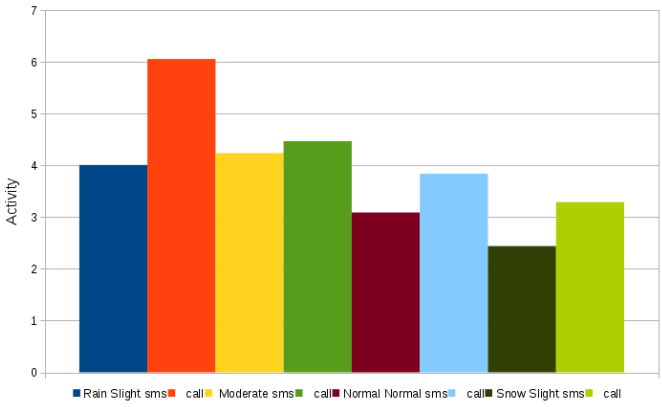

Figure 14: Activity level of outgoing SMS and call.

red and light blue bars). Nevertheless, people tend to call or send less SMS when it is snowing (dark and light green bars).

6 CONCLUSION AND FUTURE DIRECTIONS

Our purpose in this research is to understand the activity of the telecommunication network by analyzing several aspects of Milan's phone traffic flows. We were interested in the definition and morphology of the activity and transit zones; the prediction of activity levels over different regions with a success rate between 80% and 90%; the interactions between the different *activity zones* and the influence of the weather and the pollution on that activity. Thus, our results offer a new way of looking at the telecommunication traffic data by examining the various connections between appar-

ently uncorrelated datasets, providing insights to manage and to optimize the whole network. For business opportunities, this means (1) new geomarketing opportunities through a better understanding of users communication patterns, (2) new trade area analysis, (3) cheaper network load balancing as well as (4) improved QoS. In the future, we would like to study and analyze in detail the opinions (sentiment analysis) discussed by users and generators of tweets and identifying the geolocation of these activity areas. Another line of investigation would include coarse mobility Call Data Record (CDR) to take into account as an additional element for detecting activity and traffic areas.

References

D. P. Dobkin and M. J. Laszlo. 1987. Primitives for the manipulation of three-dimensional subdivisions. In *Proceedings of the Third Annual Symposium on Computational Geometry*, pages 86–99, New York, NY, USA.

Nathan Eagle, Alex S. Pentland, and David Lazer. 2013. Inferring social network structure using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278., September.

Martin Ester, Hans-Peter Kriegel, Joerg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231.

M. Ficek and L. Kencl. 2012. Inter-call mobility model: A spatio-temporal refinement of call data records using a gaussian mixture model. In *Conference on Computer Communications*, pages 469–477, March.

S. Gambs, M.-O. Nuñez Killijian, and M.. del Prado Cortez. 2010. Gepeto: A geoprivacy-enhancing toolkit. In *Advanced Information Networking and Applications Workshops (WAINA), 2010 IEEE 24th International Conference on*, pages 1071–1076, April.

Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núòez del Prado Cortez. 2011. Show me how you move and i will tell you who you are. *Transition on Data Privacy*, 4(2):103–126, August.

Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. 2012. Next place prediction using mobility markov chains. In *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, page 3. ACM.

Ramaswamy Hariharan and Kentaro Toyama. 2004. Project lachesis: parsing and modeling location histories. In *In Geographic Information Science*, pages 106–124.

John A. Hartigan. 1975. *Clustering Algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 99th edition.

Sibren Isaacman, Richard Becke, Ramn Cceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. 2011. Identifying important places in peoples lives from cellular network data. In *Pervasive Computing*, June.

J. B. Kruskal. 1956. On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. In *Proceedings of the American Mathematical Society, 7*.

M Kures and Ryan Pinkovitz. 2011. Downtown and business district market analysis. `http://fyi.uwex.edu/downtown-market-analysis/`.

Antonio Lima, Manlio De Domenico, Veljko Pejovic, and Mirco Musolesi. 2013. Exploiting cellular data for disease containment and information campaigns strategies in countrywide epidemics. *Computing Research Repository*, abs/1306(4534):1–7, Jun.

J. Morales, W. Creixell, J. Borondo , J. C. Losada , and R. M. Benito. 2013. Understanding ethnical interactions on ivory coast. In *Data for Development (D4D) challenge*, pages 115–120, May.

Ana-Maria Olteanu, Roberto Trasarti, T Couronn, Fosca Giannotti, Mirco Nanni, Zbigniew Smoreda, and Cezary Ziemlicki. 2011. Gsm data analysis for tourism application. In *Proceedings of the 7th International Symposium on Spatial Data Quality (ISSDQ)*.

Michal Piorkowski, Natasa Sarafijanovic-Djukic, and Matthias Grossglauser. 2009. CRAWDAD data set epfl/mobility (v. 2009-02-24). Downloaded from http://crawdad.org/epfl/mobility/, February.

United Nations Global Pulse. 2013. Mobile phone network data for development. `http://www.unglobalpulse.org/research`, October.

Yan Qu and Jun Zhang. 2013. Trade area analysis using user generated mobile location data. In *International Conference on World Wide Web*, pages 1053–1064.

Yossi Shiloach and Uzi Vishkin. 1982. An o(log n) parallel connectivity algorithm. *Journal of Algorithms*, 3(1):57–67.

Jameson L. Toole, Michael Ulm, Marta C. González, and Dietmar Bauer. 2012. Inferring land use from mobile phone activity. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, pages 1–8, New York, NY, USA.

M.R. Vieira, V. Frias-Martinez, N. Oliver, and E. Frias-Martinez. 2010. Characterizing dense urban areas from mobile phone-call data: Discovery and social dynamics. In *Second International Conference on Social Computing*, pages 241–248, August.

Changqing Zhou, Dan Frankowski, Pamela Ludford, Shashi Shekhar, and Loren Terveen. 2004. Discovering personal gazetteers: An interactive clustering approach. In *International Workshop on Geographic Information Systems*, pages 266–273, November.