The effect of weather on user-generated big geo data in mobile phone networks

Carolina Arias Muñoz
Politecnico di Milano
Via Valleggio 11, 22100 Como Italy
carolina.arias@polimi.it

Maria Antonia Brovelli
Politecnico di Milano
Via Valleggio 11, 22100 Como Italy
maria.brovelli@polimi.it

ABSTRACT
User-generated data in mobile networks are normally used as proxies for human activity and mobility, and it can be used in an extensive variety of research problems including mobility, city planning, tourism, event detection, urban well-being and many others. In not so many studies, the relationship between environmental variables and mobile usage data has been explored. This work tests this possible relationship in the city of Milan, as one first approach towards a predictive model for weather conditions/environmental stress from mobile usage data. Used data corresponded to two months (November and December 2013) of Call Detail Records (CDRs) of sent and received SMS, incoming and outgoing calls, and internet traffic; as well as the precipitation data from the meteorological network. According to the Kruskal-Wallis test results, we can conclude that for a confidence interval (95%) the null hypothesis of equality of medians can be rejected: there is a significant relationship between telecommunications data and precipitation intensity levels.

CCS Concepts
• **Applied computing** • *Applied computing~Internet telephony* • ***Applied computing~Mathematics and statistics*** • ***Information systems~Open source software***

Keywords
Mobile phone data; spatial data mining; urban planning; Big Geo Data; Kruskal-Wallis; Extraction of spatial relations in Big Data.

1. INTRODUCTION
User-generated data in mobile networks are normally used as proxies for human activity and mobility and it can be used in an extensive variety of research problems including mobility, city planning, tourism, event detection, urban well-being and many others. In not so many studies, the relationship between environmental variables and mobile usage data have been explored.

There are two studies worth to be mentioned: First important work was done by [1] were the authors explored the influence of weather on mobile phone usage and (indirectly) human behavior. They used factor analysis to reduce dimensionality and redundancy in some meteorological variables, and then spectral analysis to unveil significant periodical components in the time series of the remaining factors (output of the factor analysis) and mobile telecom traffic intensity. Later on [2] reproduced [1] experience, both to validate the datasets they had available and to confirm the authors' results. Their focus was on the creation of a predictive model of weather conditions and to evaluate an autoregressive integrated moving average (ARIMA) model and its forecast.

On both studies, the correlation between temperature (or variables related to Thermal perception) and telecommunications data is high, but the results related to precipitation do not match, probably due to a series of factors including the data quality and the different human behavior of the two regions used as case study.

Taking into account these previous works, we wanted to test if there is any relationship specifically between precipitation and mobile outgoing calls in the city of Milan, as a first approach towards a predictive model for weather conditions/environmental stress from mobile usage data. In this work we consider:

- The use of only Free and Open Source tools
- The of use of Kruskal-Wallis test instead of factor analysis.

2. DATA AND METHODS
This research is based on two types of data, namely the user-generated traffic in mobile networks and precipitation data. The latter is considered, to our purposes, an indicator of the weather conditions.

2.1 User-generated big geo data in mobile phone networks
Telecom Italia together with, SpazioDati, MIT Media Lab, EIT ICT Labs, Polytechnic University of Milan, Northeastern University, University of Trento, Fondazione Bruno Kessler and Trento RISE have been organizing the Telecom Italia Big Data Challenge, providing various geo-referenced and anonymized datasets. For the 2014 edition, they provided data for two Italian areas: the city of Milan and the Province of Trentino [3]. These data are available[1] to the public under the Open Database License (ODbL) .

From all the Telecom open data available, used data corresponds to two months (November and December 2013) of mobile telecommunications of the city of Milan. The datasets are user-generated telecommunication traffic, corresponding to the result of computation over the Call Detail Records (CDRs) of sent and received SMS, incoming and outgoing calls, and Internet traffic.

2017, Copyright is with the authors. Published in the Workshop Proceedings of the EDBT/ICDT 2017 Joint Conference (March 21, 2017, Venice, Italy) on CEUR-WS.org (ISSN 1613-0073). Distribution of this paper is permitted under the terms of the Creative Commons license CC-by-nc-nd 4.0

[1] https://dandelion.eu/datamine/open-big-data/

All Datasets have a temporal aggregation of ten minutes, being in total 14.877.485 records. Data are provided in a series of CSV files, each containing one day of records, organized according to the following schema:

- *Square id*: the id of the square that is part of the Milano grid (see Figure 1)
- *Time interval*: the beginning of the time interval.
- *Country code*: the phone country code of a nation.
- *Mobile network activity*: the activity (a number) inside the Square id, in terms of outgoing and received SMS/calls and internet connection (one column for each variable) during the Time interval and sent from the nation identified by the Country Code.

The CDRs records provided by Telecom Italia are not the real records; they are proportional values of the actual records, to provide anonymized data. To understand how these values were calculated, please refer to [3].

2.1.1 Spatial distribution of the mobile network data

Since the data come from different companies which have adopted different standards, their spatial distribution irregularity is aggregated in a square grid (100 columns by 100 rows) covering the city of Milan, with a square cell size of 235 meters and WGS84 projection (EPSG:4326). The Milano Grid is available in GeoJSON format (see Figure 1).

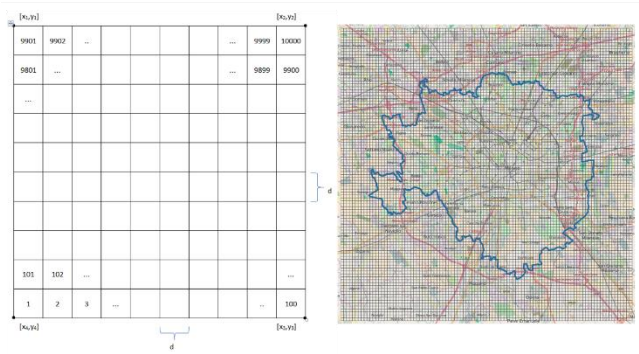

Figure 1. Representation of the Milano grid, with d = 235 meters

On the CSV files, each record is related to a specific cell id of the Milano grid, in such a way that each record can be referenced to each grid cell. Once the data is represented spatially, it can be seen as a series of raster maps, one for each time stamp (i.e. 144 raster map per day for each variable).

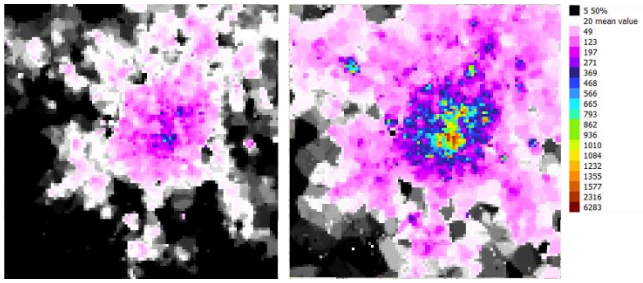

Figure 2. Mean values (left) and maximum values (right) of Outgoing calls in the city of Milan between November – December 2013

Figure 2 summarizes the spatial behavior of the data. Mean and maximum values cluster on areas characterized by a high density of buildings, population and activities, such as new residential centers, airports, industrial zones, etc.

2.1.2 Temporal distribution of the mobile network data

Figure 3 shows a clear view of data temporal behavior. These are radial box plots, where data is summarized by timestamp, meaning that the value of each hour is the sum of all grid values of a specific time stamp.

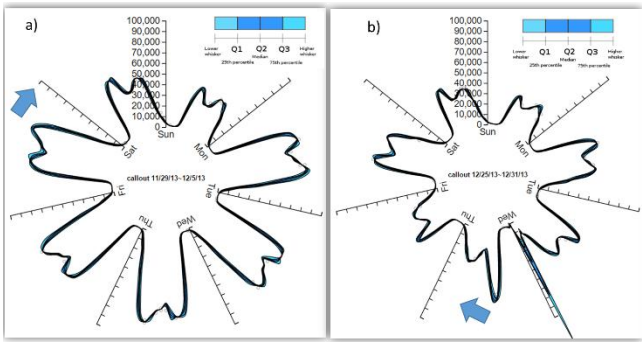

Figure 3. Radial box plot of incoming calls in the city of Milan. a) 11/29/13 and 12/05/13 (normal week). b) 12/25/13 and 12/31/13 (last week of the year, new year's eve).

The left plot shows data from December's first week, which is representative of all the rest of the weeks between November and December 2013. It evidences a strong daily seasonality of *incoming calls* corresponding with working and non-working hours; the same behavior is observed for the rest of the variables (outgoing calls, incoming and outgoing SMS, Internet connection) indicating a temporal human behavioral pattern. Likewise, there is a weekly seasonality between working days and weekends, with Sunday the day with less activity. The right plot shows data from December's last week, where peaks for December 25th and 31st can be seen clearly.

2.2 Precipitation data

Precipitation maps for Milan city between November and December 2013, comes from Lombardia's regional agency for environmental protection ARPA[2]. An Optimal Interpolation (OI) method, explained in [5], was used by [4] to interpolate on a regular gid of 1.5 km, the hourly-cumulated precipitation using data from ARPA Lombardia's mesoscale meteorological network. Figure 4 shows the maximum values during November and December 2013, where high values are concentrated outside the city of Milan, especially on the southeast part of the city.

[2] http://ita.arpalombardia.it/ita/index.asp

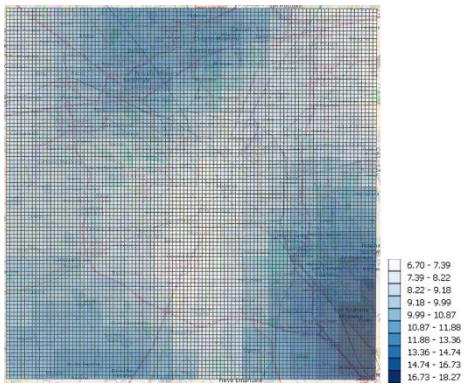

Figure 4. Milan's precipitation map of maximum values (mm) between November – December 2013

In Lombardy, the meteorological autumn and winter 2013 were very mild, rainy and with exceptionally heavy snowfall [6]. The cause is the succession of disturbed situations, characterized by currents coming from southern quadrants, which favored an almost continuous supply of warm, moist air in particularly from the end of December onwards.

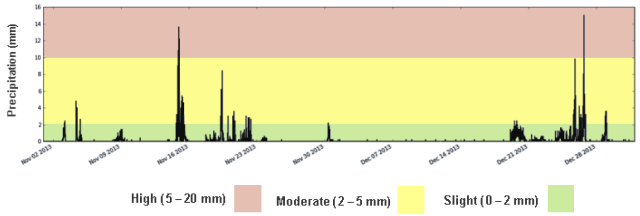

Figure 5. Milano grid precipitation map of maximum values (mm) between November – December 2013

Figure 5 shows rain distribution within the Milano grid area, as in the boxplots, summarized by time stamp. The different colors divide the threshold of precipitation intensity[3]. There are two main peaks on November 16th and December 27th.

2.3 Methods of analysis

The Kruskal-Wallis test is the non – parametric counterpart to the analysis of variance ANOVA test. It allows to compare samples of the same variable by their difference in their medians [7]. To detect any differences of telecommunications data activity between the levels of precipitation intensity (high, moderate, slight and no rain) a Kruskal-Wallis was used in each cell (square) of the Milano grid: the analysis was made for each location and all time stamps (see figure 6).

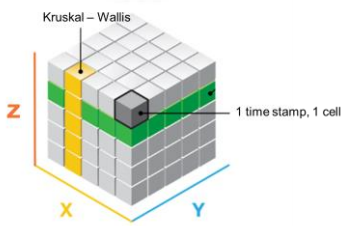

Figure 6. Schema of the map arrange to perform the Kruskal – Wallis test.

This test was used because assumptions of normality could not be confidently made for each of the cells of the Milano grid. On this analysis, the *response variable* is the telecommunications activity, spatially distributed within the city of Milan, in zones on which the precipitation can be high/moderate/slight (*factors*) or not be present at all, and its behavior can be different whether is a weekday or a weekend. The Kruskal-Wallis test was used to test the null hypothesis that the precipitation intensity levels have equal population medians. We want to know if people behavior in telecommunications activity can change: e.g. Are people's outgoing calls median significantly on days of heavy rain?

2.4 Data processing

Precipitation maps and the telecommunications activity records were analyzed on a common space-time basis using python with Spyder Scientific PYthon Development EnviRonment[4] and scipy.stats[5] and pandas[6] libraries. For data visualization QGIS[7] and d3.js[8], ricksaw.js[9], rbox.js[10] javascript libraries.

The original files were imported in MongoDB non sql database[11]. Different python scripts were used to data processing:

- Asciitotable.py: Conversion of precipitation maps into data frame data structures.
- *exploration.py:* Basic data statistics, distribution fitting and histogram analysis.
- *pre_kw.py:* Sum values ignoring country code, Aggregate data by hour, Join of telecom a precipitation data, Creation precipitation intensity and days of the week categories.
- *kruskal.py:* Kruskal-Wallis calculation by location.

All scripts used can be found on Github: https://github.com/carolinarias/Kruskal-Wallis-Spatial.git

3. RESULTS AND DISCUSSION

We calculated a series of Kruskal-Wallis maps for each of the telecommunications variables (incoming and outgoing SMS/calls).

High (heavy) rain was not considered because the number of measurements was less than five on the majority of the cells; for Kruskal – Wallis test to work, the samples must have more than

[3] http://www.arpa.piemonte.gov.it/rischinaturali/tematismi/meteo/osservazioni/radar/intensita-precipitazione.html?delta=0

[4] https://pythonhosted.org/spyder/

[5] https://docs.scipy.org/doc/scipy/reference/stats.html

[6] http://pandas.pydata.org/

[7] http://www.qgis.org/it/site/

[8] https://d3js.org/

[9] http://code.shutterstock.com/rickshaw/

[10] https://bl.ocks.org/davidwclin/ad5d13db260caeffe9b3

[11] www.mongodb.com

five observations. Figure 7 shows an example for outgoing calls (that represent the results also for the other variables), where the value 1 indicates that the test passed: Being the null hypothesis that the population medians are all equal, a P-value $\leq \alpha$ (0.05 in our case) means that the differences between the medians are statistically significant. The Kruskal-Wallis test reveals that the medians for the telecommunication activity are significantly different across the different precipitation intensity levels: moderate, slight and no rain.

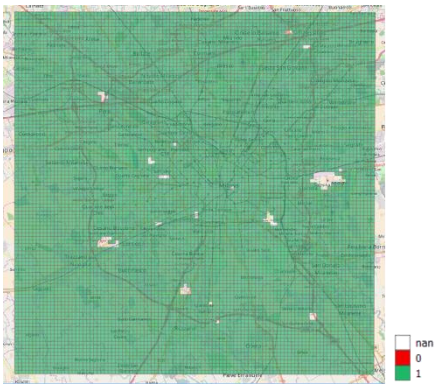

Figure 7. Kruskal / Wallis map for outgoing calls between November – December 2013

The test also identified areas where certain levels of precipitation are common (i.e. areas with the same value of moderate precipitation = 2.5 mm), identify on the map as *nan*.

4. CONCLUSIONS AND FUTURE WORK

The results discussed above are a promising step towards a holistic understanding of the complex relationship between environmental and social dynamics, and a starting point for further smart cities and human geography analysis.

According to the Kruskal-Wallis test results, we can conclude that for a confidence interval (95%) the null hypothesis of equality of medians can be rejected: there is a significant relationship between telecommunications data and precipitation intensity levels.

A following analysis would try to identify the causality of the relationship between precipitation and telecommunications activity: e.g., how strong/weak is the relationship? is there any primary process or feature which may have a spatial and/or temporal component?

We will continue this research taking into account:

- Citizen - generated geographic content vs. official sensor data.
- Test not only precipitation but other weather variables like temperature sun radiation, wind direction, etc.
- Using a larger data sample (i.e. one-year time series of data).
- Integrate additional data (i.e. traffic data, census data, historical social media data, etc.).

We hope to explore further the hypothesis of predicting weather conditions / environmental stress with the help of mobile data.

5. REFERENCES

[1] Sagl, G., Beinat, E., Resch, B., & Blaschke, T. (2011, June). Integrated geo-sensing: A case study on the relationships between weather and mobile phone usage in northern italy. In Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2011 IEEE International Conference on (pp. 208-213). IEEE.

[2] Craveiro, P., Ramos, F.M.V., Kanjo, E., Mawass, N.E., 2013. Towards an early warning system : the effect of weather on mobile phone usage A case study in Abidjan 1–11.

[3] Barlacchi, G., De Nadai, M., Larcher, R., Casella, A., Chitic, C., Torrisi, G., Antonelli, F., Vespignani, A., Pentland, A., Lepri, B., 2015. 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.

[4] Lussana, C., Salvati, M.R., Pellegrini, U., Uboldi, F. 2009. Efficient high-resolution 3-D interpolation of meteorological variables for operational use. *Adv. Sci*. Res. 3, 105–112.

[5] Uboldi, F., Lussana, C., Salvati, M., 2008. Three-dimensional spatial interpolation of surface meteorological observations from high-resolution local networks. *Meteorol. Appl*. 15, 331–345.

[6] ARPA Agenzia Regionale per la Protezione dell'Ambiente, 2015. Sintesi Meteoclimatica Inverno 2013/2014.

[7] Wheeler, D., Shaw, G., Barr, S., 2013. *Statistical Techniques in Geographical Analysis, Third Edition*. Taylor & Francis.