SENSING SLOW MOBILITY AND INTERESTING LOCATIONS FOR LOMBARDY REGION (ITALY): A CASE STUDY USING POINTWISE GEOLOCATED OPEN DATA

M. A. Brovelli [a], D. Oxoli [a, *], M. A. Zurbarán [b]

[a] Dept. of Civil and Environmental Engineering, Politecnico di Milano, Como Campus, Via Valleggio 11, 22100 Como, Italy - (maria.brovelli, daniele.oxoli)@polimi.it
[b] Dept. of Systems Engineering, Universidad del Norte, Km 5 Via Puerto Colombia, Barranquilla, Colombia - mzurbaran@uninorte.edu.co

Commission II, ThS14 - Recent Developments in Open Data

KEY WORDS: Community, Mobility, Open Data, Open Source Software, Web Platform, User-Generated Content

ABSTRACT:

During the past years Web 2.0 technologies have caused the emergence of platforms where users can share data related to their activities which in some cases are then publicly released with open licenses. Popular categories for this include community platforms where users can upload GPS tracks collected during slow travel activities (e.g. hiking, biking and horse riding) and platforms where users share their geolocated photos. However, due to the high heterogeneity of the information available on the Web, the sole use of these user-generated contents makes it an ambitious challenge to understand slow mobility flows as well as to detect the most visited locations in a region. Exploiting the available data on community sharing websites allows to collect near real-time open data streams and enables rigorous spatial-temporal analysis. This work presents an approach for collecting, unifying and analysing pointwise geolocated open data available from different sources with the aim of identifying the main locations and destinations of slow mobility activities. For this purpose, we collected pointwise open data from the Wikiloc platform, Twitter, Flickr and Foursquare. The analysis was confined to the data uploaded in Lombardy Region (Northern Italy) – corresponding to millions of pointwise data. Collected data was processed through the use of Free and Open Source Software (FOSS) in order to organize them into a suitable database. This allowed to run statistical analyses on data distribution in both time and space by enabling the detection of users' slow mobility preferences as well as places of interest at a regional scale.

1. INTRODUCTION

The diffusion of GPS-enabled mobile devices has boosted an exponential growth of social networks as well as a new dimension to community-based web platforms through which an increasing number of users interact and share geolocated information about their activities through photos, location check-ins, relevant opinions etc. In the past, the availability of citizens location records was restricted due to the lack of these recent means. With the instant access provided by mobile technologies, location-based services (LBS) have been widely adopted, making available a rich source of contextual real-time information enabling detailed analysis on user's activities and mobility.

In the scientific community, location-based services has been a popular subject and numerous studies have been made taking advantage of geolocated user content to infer users' mobility patterns as well as identify the most visited location within a region.

As relevant example, Hassan et al. (2013) used Twitter (https://twitter.com) and Foursquare (https://it.foursquare.com) data for the cities of New York, Chicago and Los Angeles characterizing different human activities in the city life and the spatial concentrations of places intended for specific activities (i.e. work, eating, entertainment, etc.). As a result, they found a correlation between the historic popularity of a place and the possibility of users visiting these locations in the present. A similar study was conducted by Aubercht et al. (2014) using Foursquare check-ins in the city of Lisbon to compare urban activity during daytime and night time. Other related research was made for the cities of Norwich (UK), Rouen (France) and Koblenz (Germany). For this study, Van de Speck et al. (2009) recorded GPS tracks with the aim of evaluating the urban quality of the city centres by measuring people's mobility. The authors conclude that the analysis of GPS tracks can give insight in understanding hidden characteristics of urban systems. In another context, Brovelli et al. (2014) used Twitter data to explore the occurrence of floods and damage assessment for Italy using hashtags to filter information.

In this work we target slow mobility activities (e.g. trekking, hiking, running, cycling etc.) due to their growing consideration for the sustainable development of the territories (La Rocca, 2010). Slow mobility has a low environmental impact while generating a positive effect in the health and general well-being of the citizens. In fact, numerous Italian as well as European projects (see e.g. http://www.progetto.vento.polimi.it, http://www.pimms-transfer-eu.org) have been focusing on the promotion of alternative transportation and leisure travelling by leading to the implementation of ad hoc services and facilities for sustainable mobility.

As case study, we selected the Lombardy Region (Northern Italy), which includes different landscapes as well as

* Corresponding author

environments ranging from highly populated cities to the alpine glaciers, passing through its famous sub alpine lakes (e.g. Como Lake) and wide plains (e.g. the Po River valley). Thanks to this territorial variety, the Lombardy Region is a good candidate for studying slow mobility through its wide range of different environments to practice these kind of activities. Thus, an effective knowledge of the ongoing activities within the territory is a relevant issue to be tackled in order to better exploit this peculiarity in terms of territorial accessibility, environmental protection as well as tourism promotion.

According to this purpose, we present here an experimental procedure to highlight locations of interest as well as mobility patterns for the Lombardy Region through the analysis of geolocated content generated by the users -and available as open data- from different web platforms.

2. DATA COLLECTION AND PRE-PROCESSING

2.1 Selection of Data Sources

In order to infer and analyse users' behaviour with respect to slow mobility, different platforms and social networks were considered. These are: Wikiloc (http://www.wikiloc.com), Flickr (https://www.flickr.com), Twitter and Foursquare. The selection was based on their popularity and wide acceptance by users globally, also because they allow sharing geolocated contents that are made available for consulting according to the user's' preferences. It is important to note that for the purpose of identifying slow mobility patterns, Wikiloc promised to be the most relevant data source because of the nature of this network, which is intended to be used to report tracks -in GPS eXchange Format (GPX)- of outdoor activities such as hiking, running, cycling, etc. Conversely, the other platforms include heterogeneous user communities reporting activities that are not necessarily related to slow mobility. This is due to the broad type of contents these networks encourage to share.

The accessed data for each of the platforms were the following: metadata of geolocated photos for Flickr, geolocated tweets from Twitter, GPX tracks for Wikiloc and check-ins (i.e. user recorded position into specific venues such as restaurants, parks, hotels etc.) for Foursquare.

Flickr, Twitter and Foursquare provided a single pointwise information for any registered user content while Wikiloc provided GPX tracks formed by hundreds of points. For this reason, the pre-processing of the raw data was different based on these characteristics to enable the data be stored into a single database table as well as to obtain comparable results from the analysis. Further details about raw data collection and pre-processing are provided in the following sections.

2.2 Data Collection and Stored Attributes

Different strategies were adopted in order to collect raw data from the selected platforms. Flickr, Twitter and Foursquare offered the possibility to access their data through Application Program Interfaces (APIs). By using APIs it was possible to schedule the collection as well as to specify filters for the requested data by programmatic means. The common format through which data is served is JavaScript Object Notation (JSON). The Flickr API (https://www.flickr.com/services/ap) enabled to collect all the available user-generated content by specifying a bounding box and a time interval. The bounding box used was the minimum rectangle containing the whole Lombardy Region. The time interval selected for Flickr corresponded to the period from June 2015 to March 2016. Twitter Streaming API (https://dev.twitter.com/streaming/overview) also allowed to obtain real-time data in a rectangular user-defined bounding box, but did not offer the possibility to collect older data. For this reason, the collection was done from January 2016 to March 2016. The Foursquare APIs include different endpoints for various purposes. The selected one was the Venues explorer (https://developer.foursquare.com), which included the possibility to ask data within a circular buffer of a selected location. This produced duplicate results for the overlapping areas which were later removed from the dataset. The request was performed three times per day during morning, noon and in the afternoon peak hours from January 2016 to March 2016. Specific Python scripts were developed under the Django Framework (https://www.djangoproject.com) in order to connect to the services for the data collection on the three platforms. This code can be reused for any geographical area and is available to the public through GitHub (https://github.com/mazucci/geocollect)

As for the Wikiloc platform, it did not include APIs, therefore the way to retrieve data was by manually downloading GPX tracks related to the Lombardy Region. The points forming the tracks were extracted using a Python script and then recorded in the database. For storing all the collected information, it was used PostgreSQL (http://www.postgresql.org) database together with its spatial extension PostGIS (http://postgis.net).

For all the platforms was established a minimum set of attributes to be recorded for any geolocated pointwise information. Attributes were: latitude, longitude, timestamp and the corresponding source platform. This attributes unification allowed to harmonise the whole dataset in order to facilitate further spatial and temporal analyses on it.

2.3 Data Filtering

Additional to the aforementioned attributes, by using timestamps and coordinates of the points extracted from the Wikiloc tracks it was possible to compute the average speed at which the tracks were recorded by the user. This parameter was used to identify slow mobility activities by filtering the GPX tracks with an average speed less than 22 km/h. According to Gilani (2005), this threshold would cover mostly non-motorised transportation, which was present inside the Wikiloc database in some proportion. Moreover, a share of the collected geolocated information lay outside the Lombardy Region. This content was generated within the area covered by the searching bounding box but not contained within the regional boundaries. These points were filtered out by means of intersection with a shapefile representing Lombardy Region by using PostGIS functionalities.

The collected data within the specified periods resulted in a total amount of 2,298,395 pointwise geolocated information of which 2,163,101 corresponded to Wikiloc, 21,849 to Flickr, 101,032 to Twitter and 12,413 to Foursquare.

3. DATA ANALYSIS

The analysis of the dataset obtained in the previous steps aimed to highlight the variation of the concentration of points in both space and time. The main purpose was to understand if the pointwise geolocated information from the different platforms

showed specific spatial patterns inside the study area, in order to identify the most visited locations. By discriminating the information according to the time when it was generated, it was possible to introduce a temporal dimension in the analysis. This enabled to underline changes in spatial patterns for selected time periods such as weekdays or weekends.

In considering slow mobility, this temporal distinction is necessary to better focus on users' behaviour on non-routine activities which are expected to occur mostly during weekends. The dataset was then divided by both source platform and day type. A specific point shapefile was created for any of these subsets in order to visualize point patterns on maps.

One of the most common tools to visualize where a higher density of pointwise data occurs in space is the heat map. Heat maps can help to explore massive data sets in order to identify single instances or clusters of important data entities (Trame and Kßler, 2011). To create a heat map, point data is analysed to produce a raster map representing an interpolated surface showing the density of occurrence. To each raster cell is assigned a density value and the entire layer is visualized using a selected colour gradient. Heat maps were thus created using the QGIS (http://www.qgis.org) Heatmap plugin (http://tinyurl.com/zlp5chr) from the point shapefiles derived by the aforementioned data subsets. The main drawback of using heat maps lies in the fact that both the type of density function and the visualization parameters adopted strongly affect the result. Moreover, density maps such as heat maps can identify where data clusters exist but not if these are statistically significant.

In order to make less subjective the interpretation of the results a hot spot analysis was performed (Dempsey, 2014). Hot spot analysis uses the Getis-Ord local statistic - Gi*- (Getis and Ord, 1992) in order to define areas of high point density occurrence (i.e. hot spots) versus areas of low occurrence (i.e. cold spots). This technique required to aggregate pointwise data -called "events"- into weighted points (i.e. points collecting all the events within their area of influence). To perform this task we adopted as weighted points the centroids of the polygons representing the Lombardy Region municipalities by assigning to each of them the count of the events detected within the corresponding municipality area. This approach was also implemented in (Bhaduri et al., 2007) to represent blocks for geospatial modelling of population distribution and dynamics. Then, specific shapefiles of weighted points were created from the same data subsets used for heat maps calculations.

In order to compute the Gi* statistic, a conceptualization of the spatial relation between weighted points was identified by considering a fixed distance band. The most appropriated distance of reciprocal influence between weighted points was selected by calculating z-scores of the global Moran's Index (e.g. Ord and Getis, 1995; Dong and Liang, 2014). The selected distance for any of the input set of points was the one maximising the Moran's Index z-score value. At this point, z-scores resulting from Gi* statistic as well as p-values of the null-hypothesis (i.e. complete spatial randomness) were assigned to each weighted point. Very high or very low (negative) z-scores -associated with very small p-values- identified probable hot spots or cold spots in the data. Reference values for z-score and p-values are associated with the standard normal distribution and the thresholds adopted depend on the specific level of confidence at which the analyst is interested.

A Python script based on PySAL library (http://pysal.github.io) was created to compute the appropriated fixed distance band using the Moran's Index as well as Gi* statistics for each of the weighted point shapefiles. The results were then displayed using QGIS.

4. RESULTS

As previously explained, we turned our attention to identify spatial clusters of high density -distinguishing between weekdays and weekend- by means of heat maps and hot spot analysis.

In order to disclose the results and relate them to the different territorial features characterizing the Lombardy Region a reference map was included in Figure 1. These features helped to provide a framework for interpreting the results at a region level, focusing mostly on the kind of territory and the activities that can be practiced.

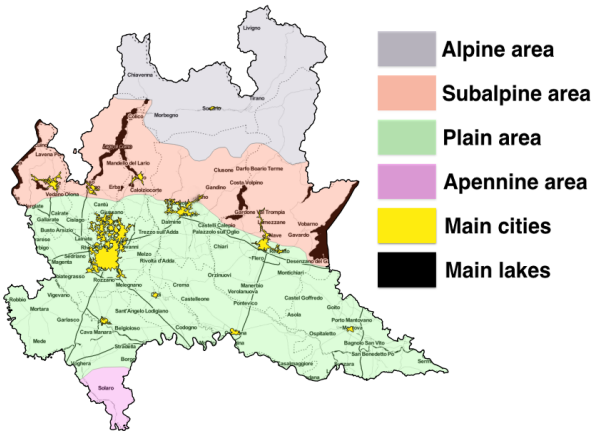

Figure 1. Schematic reference map including the main territorial features of Lombardy Region. (Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under CC BY SA.)

Regarding heat maps, we derived raster maps containing only the pixels of the original heat maps having density values higher than the mean added to the standard deviation from each single heat map. In this way it was possible to better highlight the locations at which high density of points occurred. As it is shown in Figure 2, high density clusters obtained from the different platforms showed to be concentrated in specific locations. This was particularly true for Flickr (Figure 2a), Foursquare (Figure 2b) and Twitter (Figure 2c) considering both weekdays (in yellow) and weekend (in red), orange areas result from the overlapping of the two. The highlighted locations were mainly cities with some exceptions for Flickr and Foursquare which produced small clusters scattered around lakes as well as the alpine area. A completely different situation was obtained by considering Wikiloc (Figure 2). The data showed different patterns for weekdays and weekend. During weekdays, it was possible to observe high activity around cities, lakes as well as in the alpine valleys. Data registered in weekend depict wide clusters concentrated especially along the subalpine area and alpine resorts.

The different behaviours during weekends and weekdays is evidenced also by looking at the share of data registered. In fact, 63.6 % of Wikiloc data was recorded during weekends; while

for the other platforms, lower percentages were obtained: 28.0 % for Twitter, 43.2 % for Flickr and 30.3 % for Foursquare.

For what it concerns the hot spot analysis, a specific styling of the point shapefiles containing the value from Gi* statistics was required. Different colours were used to distinguish between hot spots and cold spots, while different shades of these colours were adopted to highlight the level of confidence associated to any point. In Figure 3, the layers obtained from the hot spot analysis are presented. It can be noted that most of the cluster highlighted through heat maps for Flickr (Figure 3a-3b), Foursquare (Figure 3c-3d) and Twitter (Figure 3e-3f) were not visible in this case. Only the city of Milan resulted a high density cluster as well as a hot spot for these three platforms by considering weekdays and weekend. Besides being more numerous, hot spots from Wikiloc better reflect the patterns of the high density cluster from heat maps while bringing additional information. During weekdays, hot spots concentrated mainly around some of the main cities (i.e. Milan and Brescia) as well as in the alpine area (Figure 3g). During weekend, an important hot spot concentration appears all along the subalpine area (Figure 3h), while cold spots concentrate mainly in the plain area. With the Wikiloc data, the hot spot analysis clearly highlights the different concentration of activities depending on the territorial features of Lombardy Region.

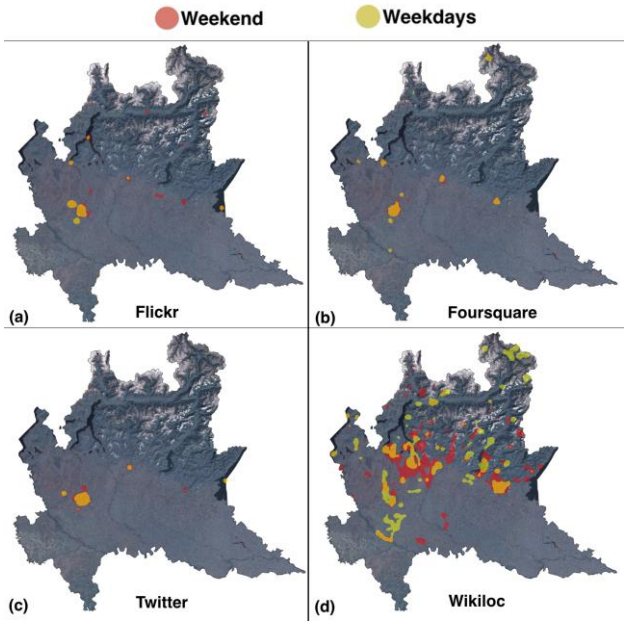

Figure 2. Density maps of the Lombardy Region for the different datasets of the selected social platforms during weekdays and weekends. (© 2016 MapQuest).

To summarize, it can be seen on maps produced from the Wikiloc data, that most hot spots as well as larger clusters are formed and more spread than in the other platforms, this is true during weekdays as well as in the weekend. This can be explain by the fact that Wikiloc is a specialized platform to record slow mobility activities, but also because of the difference that implies working with data from tracks as to pointwise locations. The map in Figure 3 showing Wikiloc activities during the weekends depicts a great concentration of hot spots within the subalpine area, indicating that the lakes and mountains are an important attraction for slow mobility. This may be explained by the variety of landscapes and the large amount of trails that this area offers.

On the maps from Flickr, Foursquare and Twitter, it can be seen that most of the clusters match each other. This redundancy shows places of interest across the region focused on the main cities, which are reasonably popular locations. In accordance with Wikiloc, Flickr and Foursquare also report a concentration of information along the alpine area and around the lakes, strengthening the validity of these locations as important attractions.

It is common for all the considered platforms that few or no clusters nor hot spots are found within the plain area, suggesting this as an unpopular area.

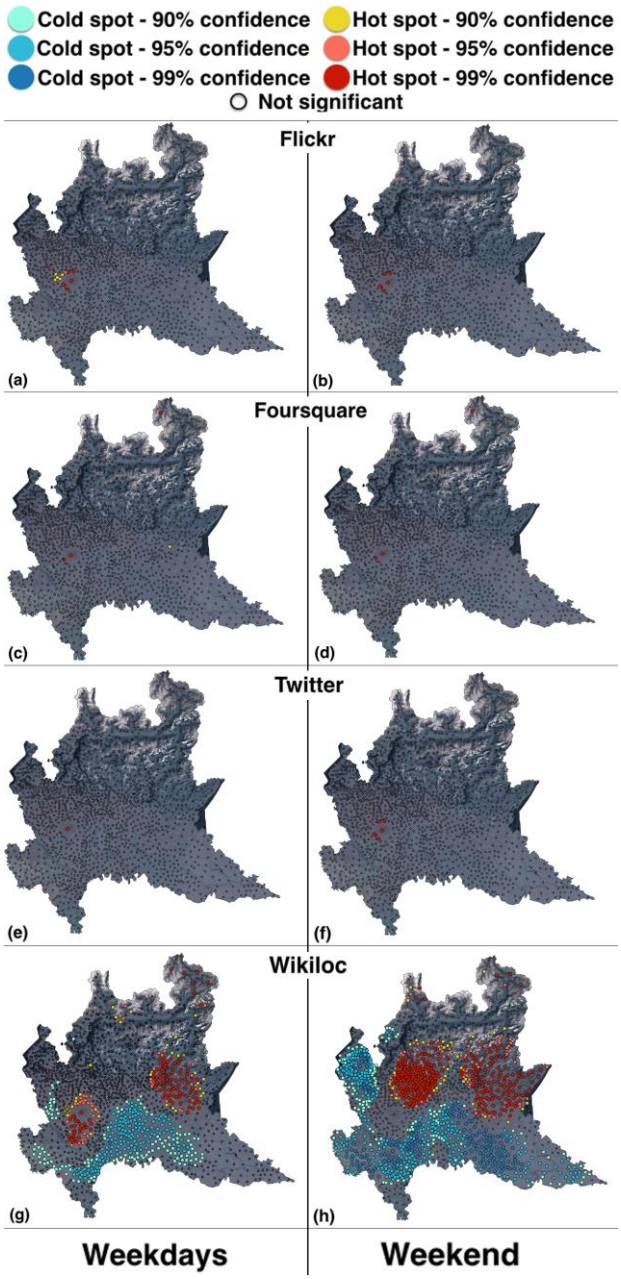

Figure 3. Hot spot analysis with different confidence levels for the different datasets of the selected platforms during weekdays and weekends. (© 2016 MapQuest).

5. CONCLUSIONS

We presented an experimental approach to exploit user-generated content in order to highlight users' mobility preferences and locations of interest for the Lombardy Region. During this experiment, we assessed that the Wikiloc dataset better described locations for slow mobility activities with respect to the other platforms. For these platforms, we expect that the results could be improved by using data that better matches the focus of the research, performing specific filtering (e.g. through keywords, hashtags, venues category etc.). Moreover, it could be important to include data recorded during longer periods in order to perform analysis over different seasons, which may affect the behaviour of users.

Besides the results achieved for this particular study, we can state general considerations about the adopted procedure. The main advantage of the proposed analysis framework relies on its capability to highlight important features of big geospatial datasets from different data sources. This eases the interpretation of the results for non-specialized users, who at a glance can get valuable insight through thematic maps. In addition, the use of open data and FOSS brings considerable added value to the analysis by enabling cooperation within the GIS community, which allows to customize the procedure for further improvement and its application on other possible contexts while being economically advantageous at the same time.

ACKNOWLEDGEMENTS

Acknowledgements to the Sustain-T Project (Technologies for Sustainable Development) by Erasmus Mundus for supporting the author and encouraging international cooperation in research.

REFERENCES

Aubrecht, C., Ungar J., Freire S., 2011. Exploring the potential of volunteered geographic information for modeling spatio-temporal characteristics of urban population: A case study for Lisbon Metro using foursquare check-in data. In: *Proceedings of the 7th International Conference on Virtual Cities and Territories*, 57-60.

Bhaduri, B., Bright E., Coleman P., Urban M. L., 2007. LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal*, 69 (1-2), 103-117.

Brovelli, M.A., Zamboni, G., Muñoz, C.A., Bonetti, A., 2014. Exploring twitter georeferenced data related to flood events: an initial approach. In: *AGILE 2014 International Conference on Geographic Information Science. Connecting a Digital Europe through Location and Place*. Castellon, Spain.

Dempsey, C. 2014. What is the difference between a heat map and a hot spot map?. *GIS Lounge* https://www.gislounge.com/difference-heat-map-hot-spot-map.

Dong, L., Liang, H., 2014. Spatial analysis on China's regional air pollutants and CO2 emissions: emission pattern and regional disparity. *Atmospheric Environment*, 92, 280-291.

Getis, A., Ord, J.K., 1992. The analysis of spatial association by use of distance statistics. *Geographical analysis*, 24(3), 189-206.

Gilani, H., 2005. Automatically Determining Route and Mode of Transport Using a GPS Enabled Phone. PhD thesis, University of South Florida.

Hasan, S., Zhan X., Ukkusuri S. V., 2013. Understanding Urban Human Activity and Mobility Patterns Using Large-scale Location-based Data from Online Social Media. In: *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, UrbComp 2013*, 6:1-6:8, ACM, New York, NY, USA.

La Rocca, R.A., 2010. Soft Mobility and Urban Transformation. *Tema. Journal of Land Use, Mobility and Environment*, 2.

Ord, J.K., Getis, A., 1995. Local spatial autocorrelation statistics: distributional issues and an application. *Geographical analysis*, 27(4), 286-306.

Trame, J., Kßler C., 2011. Exploring the lineage of volunteered geographic information with heat maps. In: *Proceedings of GeoViz: Linking Geovisualization with Spatial Analysis and Modeling*, 10-12, Hamburg, Germany.

Van der Spek, S., Van Schaick J., De Bois P., De Haan R., 2009. Sensing Human Activity: GPS Tracking. *Sensors*, 9(4), 3033-3055.