Big data analytics for smart mobility: a case study

Roberto Trasarti[1] Barbara Furletti[1]

Lorenzo Gabrielli[1] Mirco Nanni[1] Dino Pedreschi[1,2]

[1] KDD Lab - ISTI - CNR
Pisa, Italy
name.surname@isti.cnr.it

[2] University of Pisa
Pisa, Italy
pedre@di.unipi.it

1. APPLICATION SCENARIO

This paper presents a real case study were several mobility data sources are collected in a urban context, integrated and analyzed in order to answer a set of key questions about mobility. The study of the human mobility is a very sensitive topic for both public transport (PT) companies and local administrations. This work is a contribution in the understanding of some aspects of the mobility in Cosenza, a town in the South of Italy, and the realization of corresponding services in order to aswer to the following questions identified in collaboration with the PT experts.

Question 1: How is PT able to substitute private mobility? The objective is to compare public and private mobility to verify the capability of PT to satisfy the user mobility needs.

Question 2: How different zones of the city are reachable using PT? This question focuses on understanding how much different zones of the city are served by PT considering different times of the day.

Question 3: Are there usual time deviations between real travel times and official time tables? We want to verify if usual time deviations between real travel times and official time tables exist highlighting chronic delays in the service.

Question 4: Can we spot visitors and commuters by their behavior? We aim at identifying important categories of people estimating their segmentation in order to evaluate the corresponding demand of services.

For this case study we use data from Cosenza area: a GSM dataset [1], a GPS dataset [2], and data from the PT system [3]. GSM data contains 25 mln of phone calls made by about 350K distinct users from 15 October to 9 November 2012. GPS dataset contains about 1.5 mln of private vehicle tracks gathered in February-March and July-August 2012, while PT data consist of a set of GPS logs obtained by the on-board tracking system of the Cosenza's PT and the PT official time table containing the scheduled times of the arrival of the buses at their stops.

2. METHODOLOGY AND RESULTS

To answer the questions posed by the PT manager we developed and implemented a set of methodologies and processes, and we integrated the corresponding services in M-Atlas [3], a larger mobility data analysis framework developed in our laboratory.

For Question 1 we study the PT capabilities to replace the private mobility in a city. We use the GPS logs of the buses, a *real* time table computed starting from the real buses movements, and the GPS tracks of the private vehicles. We map the PT system to a spatio-temporal network, where nodes are bus stops labeled with name and position, while edges are the connections labeled with origin-destination stops and timestamp. Then, we map the GPS tracks on the PT network and we compute the shortest way to satisfy the users' mobility using an agent-based algorithm that simulates the human mobility in a network [1]. To evaluate the efficiency of the PT we compute the percentage of travels satisfied by the public transport considering a temporal and spatial tolerance (*Coverage*), and the distribution of delays accumulated by the user using the PT instead of the car (*Distribution of time deviations*). Using a maximum *walking distance* of 2 km and applying a temporal constraint of 1 hour as maximum delay, we obtain that the percentage of the user's car travels fully made by using PTs without taking more than 1 hour of extra time is 24%. If we further investigate the delay of the PTs travels w.r.t. the car ones, we find that the delay distribution is affected by the seasonality: in summer the average delay is 29 minutes (with a variance of 26), while in winter is 16 minutes (with a variance of 15). Going back to the trajectory data and extracting the starting points of the users which are not served by the public transport, we can discover which areas are disconnected from the network. By using a clustering algorithm on the starting points of GPS tracks that are not fully covered by the PT we identify two peripheral areas, one industrial and one residential, that are not reached by the bus service (Fig. 1). This result suggests the introduction of new lines or the addition of new bus stops to an existing line passing near those areas. This service is very effective in discovering the real needs of the population and how the network

[1] Wind Telecom S.p.a http://www.wind.it/

[2] Octotelematics S.p.a. http://www.octotelematics.com/

[3] Amaco S.p.a. http://www.amaco.it/

(c) 2014, Copyright is with the authors. Published in the Workshop Proceedings of the EDBT/ICDT 2014 Joint Conference (March 28, 2014, Athens, Greece) 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

can handle them, and the analysis may highlight potential customers which can be served by the public transport and therefore good candidates for specific marketing campaign.

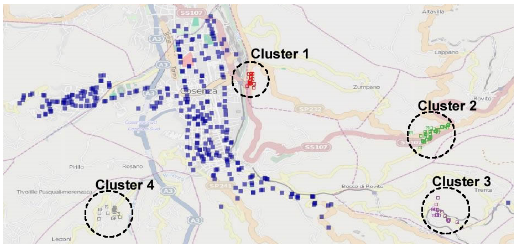

Figure 1: Areas that are not served by the PT service (blue dots are bus stops).

For Question 2 we try to understand which areas of the city can be reached starting from a specific bus stop at a specific time of the day, having a fixed amount of time on the PT network. As a result we find that particular areas of the city can be reached by the PTs in a fixed amount of time only in certain time slots, as shown in Fig. 2. This service allows the PT manager to add lines or modify the bus schedule for analyzing the impact of his choice in the PT system.

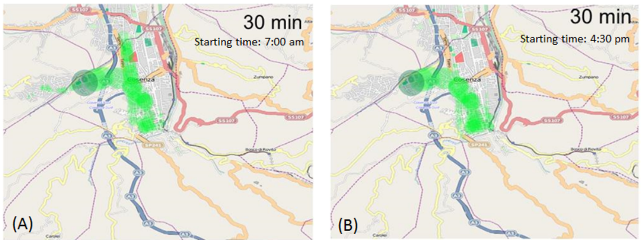

Figure 2: The reachability of the city starting from the darker point at 7:00 am (A), and 4:30 (B) and having 30' of time available.

We answer to Question 3 by computing the differences between the expected duration of the bus as stated in the official time table and the one inferred by the bus log. Fig 3 shows that almost all the buses are late in a range $[10, -10]$ min. except for bus CVR A which has an average delay of 17 min. with a very small variance. This kind of information is very useful to spot problems in the buses management, i.e. to improve the service or to highlight too strict schedules which can't be respected by the buses in reality. As a result we draw a complete map of the typical behaviors of the buses and we identify the most critical lines. The last consideration is about the buses which makes the travel faster than expected, these are clearly buses which try to reduce the delay accumulated in previous travels. This behavior is harmful and makes the time table unreliable.

To answer to Question 4 we apply the analytic process described in [2] which analyzes the calling behavior of the users in order to classify them into three categories: *Resident*, *Commuters* and *Visitors*. People that appear only once (i.e. that make only one call in all the period of observation) before disappearing are separately classified in the *In Transit* category. We obtain the following segmentation: 23.12% of Residents, 14.56% of Commuters, 26.45% of Visitors, and 28.74% of In transit. The 7.13% are *unclassified* due to their unclear profile. GSM data are a good proxy to compute people presence in a territory with a certain regularity and with an economic convenience because survey campaigns are expensive and time consuming. Our indicator based on GSM data helps to manage and re-arrange the resources and services w.r.t. the user demand.

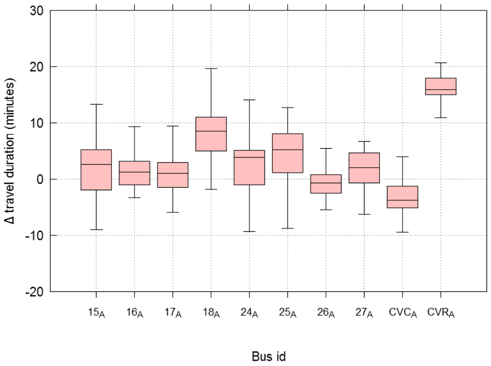

Figure 3: Official schedule vs. inferred one.

The collaboration with the public administration helped us to identify several key questions concerning the mobility needs and the transportation offers. By exploiting the peculiarities of the different data sources that were available in the application context we answered producing a set of analyses and implementing a set of services for extracting useful and new knowledge. The results have been tested on the field, allowing a continuously monitoring of the general status and health conditions of the urban traffic, in terms of impact of PT, actual mobility demand, and mobility profiles of citizens living in the area. We consider this as a preliminary work towards the definition of a sort of dashboard for a mobility manager composed of a set of end-user services and indexes to evaluate the transport system of a city.

3. ACKNOWLEDGMENTS

This work has been partially funded by the European Union under the National Operational Program Research and Competitiveness 2007-2013: Project Tetris n. PON1_00451; and under the FP7-ICT Program: Project DataSim n. FP7-ICT-270833.

4. REFERENCES

[1] F. Pinelli et all. *Space and time-dependant bus accessibility: a case study in Rome.* Proc. of the 12th IEEE Conf. on ITS, 2009.

[2] B. Furletti et all., *Analysis of GSM Calls Data for Understanding User Mobility Behavior.* IEEE Big Data, 2013

[3] F. Giannotti et all., *Unveiling the complexity of human mobility by querying and mining massive trajectory data.* VLDB Journal Special issue on Data Management for Mobile Services (2011).