Validating general human mobility patterns on massive GPS data

Luca Pappalardo, Salvatore Rinzivillo,
Dino Pedreschi, and Fosca Giannotti

KDDLab, Institute of Information Science and Technologies (ISTI),
National Research Council of Italy (CNR)
{luca.pappalardo,rinvizillo,fosca.giannotti}@isti.cnr.it

Discussion Paper
Original paper is in [1]

Abstract. Are the patterns of car travel different from those of general human mobility? Based on a unique dataset consisting of the GPS trajectories of 10 million travels accomplished by 150,000 cars in Italy, we investigate how known mobility models apply to car travels, and illustrate novel analytical findings. We also assess to what extent the sample in our dataset is representative of the overall car mobility, and discover how to build an extremely accurate model that, given our GPS data, estimates the real traffic values as measured by road sensors.

Keywords: Human mobility, Data Mining

1 Introduction

The analysis of human movement has received increasing attention in the last decade, due to the emergence of big mobility data, portraying mobility activity at an unprecedented scale and detail. Data collected by wireless technologies, such as GPS and mobile phone networks, constitute a brand new social microscope, which promises to help us discover the hidden patterns and models that characterize the trajectories humans follow during their daily activity. This direction of research has recently attracted scientists from diverse disciplines, notably data mining and network science, also given its importance in domains such as urban planning, sustainability, transportation engineering, public health, and economic forecasting.

In this paper, we focus on mobility by *car*, the most popular private means for transportation in the current society. We have access to a unique dataset consisting of the detailed spatio-temporal trajectories of approximately 10 million travels, accomplished by more than $150,000$ cars in central Italy during the month of May 2011. Starting by the preferential return model for human mobility introduced by Barabási and others in [3, 4], which explains the patterns and laws governing the key phisical quantities of human movements, we address the

following question: does this model apply to car travel? Models in [3, 4] are developed with reference to mobile phone data, which have two main differences with respect to our GPS data: mobile phone data pertain to general mobility (while GPS data pertain only to cars) and are much less detailed than GPS trajectories, the latter providing for the precise spatio-temporal records of each travel with high exact geo-location and high sampling rate. It is therefore legitimate to investigate to what extent the previous models apply to GPS data, which deviations are observed, and which new analytical opportunities are provided by the finer spatio-temporal granularity. On the other hand, it is compulsory to investigate to what extent our GPS data are representative of the overall vehicular mobility, in order to generalize the validity of our findings. To this purpose, we use independent ground-truth measurements of global traffic volumes obtained by sensors placed in a set of locations during the same observational period of our GPS data, and show that the GPS data are an extremely accurate estimation of the overall volumes in each location.

2 Related works

In the past few years, the exploding prevalence of mobile phones, GPS navigators, and other handheld devices allowed scientists to track human mobility and to test mobility models on a fertile ground. González *et al.* [3] analyzed a massive mobile phone dataset, and found a power law in the distribution $P(r_g)$ of the radius of gyration r_g, the characteristic distance traveled by a user. Authors of [4] discovered that the number of distinct location visited by humans is sublinear in time, while the probability of a user to visit a given location presents a Zipf-like distribution. Bazzani *et al.* [5] discussed an exponential law for the trajectories distribution in a urban road network, using a GPS dataset on private cars similar to ours. In [6], Lee *et al.* clustered visit points from a GPS dataset to form hot spots, finding that the pausetime distribution in hot spot follows a truncated power law consistently with [5]. From the data mining community, authors of [2] presented an extensive set of analyses on large sets of GPS data.

Predicting and estimating the number of vehicles is a crucial component of advanced traffic management and information system. Factor approaches are the most popular methods, and are generally implemented by developing a set of factors from historical data and applying them on new data to make predictions [7]. Locally weighted regression is a memory-based algorithm that learns continuous mappings from real-valued input vectors to real-valued output vectors. It assigns a weight to each training observation depending on the location of the training point in the input variable space relative to that of the point to be predicted [8]. Artificial neural networks and support vector machines also have proved to be valid alternatives for modeling and predicting traffic counts [9].

3 Understanding the patterns of car travel

Our GPS dataset stores information of approximately 9.8 Million different car travels from 159,000 cars tracked during one month (May 2011) in an area corresponding to central Italy (a 250km×250km square). The GPS traces are collected by a company that provides a data collection service for insurance companies, covering around 2% of the total registered cars[1]. The GPS device is automatically turned on when the car is started, and the global trajectory of a vehicle is formed by the sequence of GPS points that the device transmits each 30 seconds to the server via a GPRS connection. When the vehicle stops no points are logged nor sent. We exploit these stops to split the global trajectory into several sub-trajectories, that correspond to the travels performed by the vehicle. Clearly, the vehicle may have stops of different duration, corresponding to different activities. To ignore small stops like gas stations, traffic lights, bring and get activities and so on, we chose a duration threshold of at least 20 minutes: if the time interval between two consecutive observations of the car is larger than 20 minutes, the first observation is considered as the end of a travel and the second observation is considered as the start of another travel.

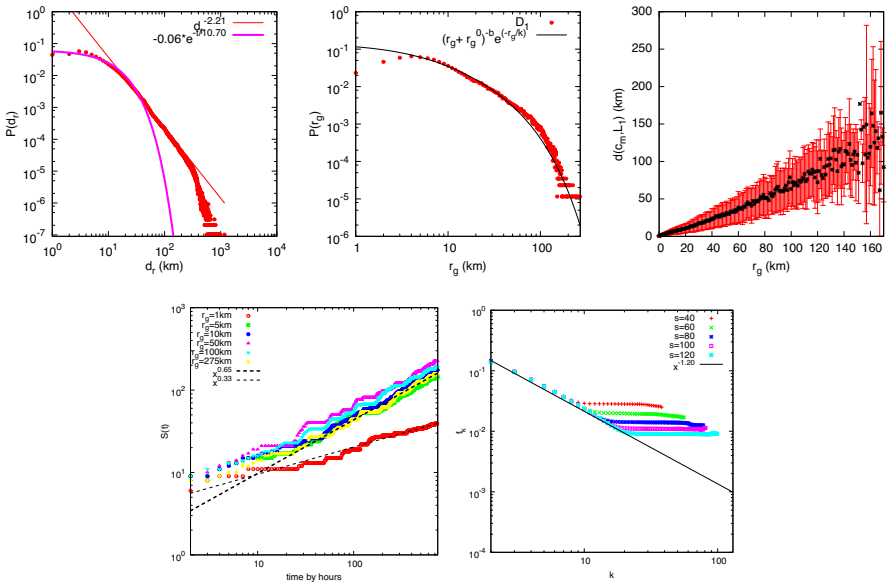

Fig. 1. *(Top Left)* Probability density function of travel distances in kilometers. *(Top Center)* Distribution of the radii of gyration. *(Top Right)* Correlation between r_g and the distance $d(c_m, L_1)$. *(Bottom Left)* Number of visited distinct location over time for different r_g groups. *(Bottom Right)* Visitation frequency of the kth most visited location, for users that have visited $s = 40$, 60, 80, 100 and 120 different locations.

[1] `http://www.octotelematics.it/`

Our first measurement is the distribution of travel length, from which two different regimes clearly emerge (Figure 1, left). The first one (from 1 to about 20 km) corresponds to low range travels, mainly located within the cities, and is characterized by an exponential distribution. The second regime corresponds to inter-city travels (i.e. travels connecting different urban areas), and is governed by a power law distribution with exponent $\beta = 2.53$. Here, the observed scaling exponent is different from the one observed for GSM data ($\beta = 1.75$) [3], since it is influenced by a reduced range of distances in the GPS dataset. Indeed, both geographical and physical boundaries (the region considered has a length of about 500 km and travels longer than 500 km are rare) provide a limitation to the range of possible distances. In order to characterize human mobility patterns emerging from available trajectories, we used the radius of gyration as the characteristic travel distance covered by each individual, a measure of how far a car is from its center of mass (mean location) [3]. Formally, the radius of gyration of a user u is defined as

$$r_g^u = \sqrt{\frac{1}{n_u}\sum_{i=1}^{n_u}(\boldsymbol{r}_i^u - \boldsymbol{r}_{cm}^u)^2}, \tag{1}$$

where $\boldsymbol{r}_i^u$ represents the $i = 1, \ldots, n_u$ positions recorded for the user u, and $\boldsymbol{r}_{cm}^u = \frac{1}{n_u}\sum_{i=1}^{n_u}\boldsymbol{r}_i^u$ is the center of mass of the trajectory. For each user in the dataset D_1, we calculated his radius of gyration by taking all points composing his sub-trajectories as the n_u recorded positions. Then, we plotted the distribution of r_g, observing a power law with an exponential cutoff, $P(r_g) \sim (r_g + r_0)^{-\beta}\exp(r_g/\tau)$ where $r_0 = 5.54$, $\beta = 1.13$ and $\tau = 39.76$ (Figure 1, center). Such curve agrees with the previous results found on GSM data (power law with $\beta = 1.65$) [3], confirming that the majority of users travel within a small distance, but some of them carry out long journeys. The difference between the predicted distribution and the observed behavior for people with low r_g (up to 5km) is presumably due to the tendency of covering small distances by foot, bike, or bus, resulting in a low probabilty to find such travels in our dataset. Figure 2 shows the spatial distribution of the centers of mass and most frequent locations, with the color representing the value of relative r_g. People with high radius of gyration concentrate their center of masses in the countryside, in the mountains (Appennines) and on the coast, whereas those with lower r_g are mainly located in urban areas. Another interesting characteristic of individual mobility we consider is the most frequent location L_1, i.e. the zone where a vehicle can be located with highest probability when it is stationary, most likely his home or work. To estimate L_1 for a user u, we calculated all the locations where he goes by extracting origin and destination points of his sub-trajectories, without taking into consideration the time spent in each location by the vehicles. Then, we applied on such points the Bisecting K-means clustering algorithm, an extension of K-means algorithm that splits the set of all points in two clusters, dividing recursively the obtained clusters until they have a radius smaller than or equal to a threshold, set in our experiments to 250 meters. The centroid of the cluster with the highest frequency is chosen as L_1 for the user u.

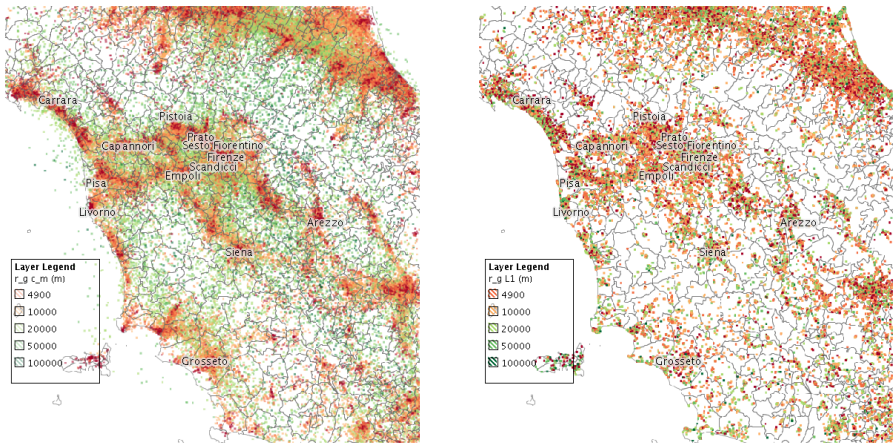

Fig. 2. Spatial distribution of centers of mass (left) and most frequent locations (right) on the map of central Italy.

The most frequent location does not necessarily coincide with the center of mass, and the distance $d(L_1, c_m)$ tends to grow with r_g. The strong correlation in Figure 1 (Top Right) shown is not obvious, and it is presumably due to the systematic nature of human motion. Indeed, if a person travels arbitrarily on any direction from and to the same preferential location, then the distance between c_m and L_1 would tend to zero, and the radius of gyration would have no relation with such distance. On the contrary, since each vehicle follows systematic travels among few preferred places, the center of mass is pulled by these trips towards the mean point of the frequent locations. Therefore, the more a vehicle travels away from its L_1, the more the center of mass tends to be distant from the most frequent location. Figure 1 (Top Right) also suggests that people with low r_g have a larger probability to be located near the place where they live or work. On the contrary, people traveling at large distances tend to be located in distant places, depending on the fact that they are moving or not. Such phenomenon is confirmed by plotting on the map L_1 instead of $\boldsymbol{r}_{cm}$, and noting that points corresponding to users with high radius of gyration move towards urban centers, showing a power of attraction of cities on mobility by car (Figure 2).

In order to estimate the trend of people to visit new distinct locations, we extracted the number of clusters $S(t)$ visited by a user, finding a power law $S(t) \sim t^{\mu}$. For users having a small r_g (within 1km), the exponent of power law is $\mu \approx 0.3$, while it grows for users traveling at large distances from the center of mass, $\mu = 0.65 \pm 0.03$ (Figure 1 bottom left). In both cases, the fact that $\mu < 1$ indicates a decreasing tendency of users to visit previously unvisited locations. Moreover, the visitation frequencies of individuals, that measures to what extent individuals return to the same place over and over, follow a Zipf's like distribution $f_k \sim k^{-1.2}$ (Figure 1, bottom right), confirming the pattern found in [4].

The results of our analysis substantially confirm and refine the mobility patterns found on GSM data [3], with a difference in the population of very short range travelers which is underrepresented with respect to the prediction. This suggests that movements by car represents a significant portion of human travels, serving as a good social microscope that enables us to observe habits, trends and patterns in human mobility behavior.

4 Inferring traffic count by GPS data

GPS data representing movements of cars traveling within an urban territory could be very useful to address urban traffic monitoring and prediction, provided that this data are a trustable proxy of ground-truth. This is also important to assess the generality of our analytical results: to what extent our 2% sample of tracked cars is representative of the overall mobility?

In this study, we use as ground truth a dataset D_2 composed by logs collected in May 2011 from twelve Variable Message Panels (VMP). VMPs are devices situated in the outer belt of the city of Pisa with the purpose of counting hour by hour all the vehicles entering the entry gates of the city. Exploiting the spatial precision of GPS data, we simulated the number of GPS vehicles crossing a VMP gate, by considering a buffer of 30 meters radius around the position of the road sensors, and by aggregating hour by hour the number of GPS vehicles crossing those areas. We found a good match between the curves, which essentially differ for a scaling factor (Figure 3).

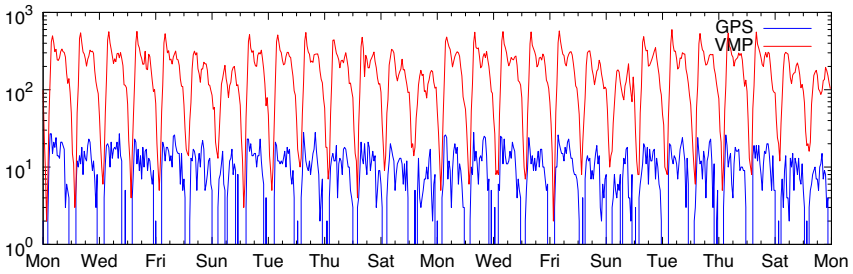

Fig. 3. Traffic sensed by a VMP device and GPS traffic volume in one entry gate.

In order to estimate accurately such scaling, we considered VMP and GPS traffic counts related to each gate as discrete signals, and decomposed them through a Discrete Wavelet Transform (DWT). DWT is a mathematical tool that projects a time series onto a collection of orthonormal basis functions and produces a set of coefficients, capturing information from the time series at different frequencies and distinct times. Given two decompositions, representing the VMP and GPS traffic count of a gate, we exploit the produced coefficients to build a model able to infer real traffic from a GPS sample.

To check the validity of such approach and evaluate the performance of the model, we measure the error with respect to the observed VMP traffic counts in all locations. In Figure 4 is showed the real VMP series, the scaled GPS signal, and the measured relative error at a selected VMP location. The error is low when the GPS traffic is high. During the night hours the relative error tends to grow since there are too few circulating GPS vehicles, but the absolute error is still negligible.

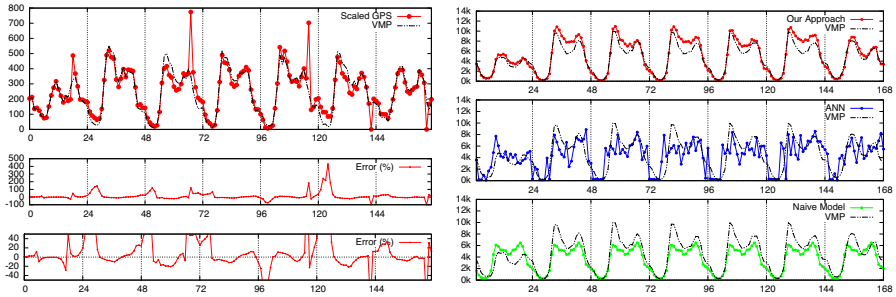

Fig. 4. *(Left)* Real traffic volume versus scaled GPS signal in a single location. Plot at the bottom is the relative error. *(Right)* Real traffic data versus our approach (upper), a neural network approach (center), and a naive approach (bottom).

It is possible to generalize the approach to scale the GPS traffic observed in a single VMP location to the total traffic entering the city, i.e. the total traffic measured by all the VMP sensors. This approach can enable a real time traffic estimation based on the observation of the GPS vehicles alone, reducing the need for ad hoc installation of new panels. To generalize the inference to the real time scenario, we need to create a model trained on historic data capable of giving precise estimates for the traffic situations. To this aim, we create the signal $\boldsymbol{v}$, that represents the total real traffic obtained by summing hour by hour the traffic volume of all VMP devices. Then, we divide $\boldsymbol{v}$ into a training set $\boldsymbol{v}_{TR}$, used to learn the model; and a test set $\boldsymbol{v}_{TS}$ used to evaluate the performance of the extracted model. As training set we used the sub-signal corresponding to the first week of our dataset. The remaining weeks are used as test set. Our method consists in learning a model by extracting the scaling factors, and then by estimating the signal of the unseen traffic. The resulting series is then compared with the real signal $\boldsymbol{v}_{TS}$.

To evaluate the accuracy of our model we performed the evaluation using two other approaches: a Backpropagation Multilayer Feedforward Neural Network [12] and a naive predictor, learned on the training set by averaging, for each hour, the values observed during the training week. Figure 4 (Right) compares the estimations made by the three approaches. Our DWT method maintains the general shape of the curve but tends to overestimate the real traffic, especially in daylight hours. The ANN approach provides volumes that are comparable

with the VMP observations but does not preserve the general shape of traffic. Finally, the naive approach shows how the phenomenon can not be captured by a static model. Although the shape of the naive curve is similar to the VMP curve, it tends to underestimate the real traffic and in particular the first peak in the morning, corresponding to people going to work. Furthermore, since it is a simple daily mean of the traffic in the observation period, it is not able to discriminate between working and nonworking days. Of the three approaches, ours gives a better approximation of evolution of traffic during the week capturing the crucial peaks during the rush hours.

5 Conclusions

In this paper we studied the patterns of human mobility by car using a large dataset of GPS traces collected in central Italy. Since the GPS data pertain only private cars movements, we used our data to assess the validity of the general laws of mobility derived from individual movements observed by means of GSM data. Moreover, we focused on the analysis of local behavior and validity of the dataset by comparing the observations with the ground-truth provided by real-traffic sensors. The experimental settings showed a close correlation between the real traffic volume and the scaled GPS flows obtained by means of a machine learning approach. The final part of the paper introduces a method, based on historic data analysis, to monitor real time traffic.

Acknowledgements. The research reported in this article has been partially supported by European FP7 project DATASIM (`http://www.datasim-fp7.eu/`).

References

1. L. Pappalardo, S. Rinzivillo, Z. Qu, D. Pedreschi, F. Giannotti, Understanding the patterns of car travel, European Physics Journal Special Topics 215, 61-73 (2013).
2. F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, R. Trasarti, VLDB Journal **20**, 2011 695-719.
3. M. C. González, C. A. Hidalgo, A. L. Barabási, Nature **454**, (2008) 779-782.
4. C. Song, T. Koren, P. Wang, A. L. Barabási, Nature Physics **6**, (2010) 818-823.
5. A. Bazzani, B. Giorgini, S. Rambaldi, R. Gallotti and L. Giovannini, Journal of Statistical Mechanics: Theory and Experiment, May 2010.
6. K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, Demystifying levy walk patterns in human walks, NCSU, Technical Report, 2008.
7. N.J. Garber, L.A. Hoel, *Traffic and Highway Engineering*, Brooks/Cole Publishing Company, 1999.
8. J. H. Friedman, *Intelligent local learning for prediction in high dimensions*, International Conference on Artificial Neural Networks (ICANN 95), Paris, 1995.
9. P. Lingras, S. Sharma, M. Zhong, Journal of the Transportation Research Board **1805**, (2002) 16-24.
10. G. Strang, SIAM Review **31**, (1989) 614-627.
11. N. M. Temme, Applied and Computational Harmonic Analysis **4**, (1997) 414428.
12. D. Svozil, V. Kvasnickab, J Pospichalb, Chemometrics and Intelligent Laboratory Systems **39**, (1997) 43-62.