Exploring the Relationship between Mobile Phone Call Intensity and Taxi Volume in Urban Area

Marco Veloso, Santi Phithakkitnukoon, and Carlos Bento

***Abstract*— As urbanization increases rapidly, there is a need for better understanding of the city and how it functions. Increasing digital data produced by the city's inhabitants holds great potential for doing so. In this work, an analysis of mobile phone call intensity and taxi volume in Lisbon, Portugal was carried out. With one source of data describes how city operates socially over mobile phone network and the other characterizes urban dynamic in traffic network, we discovered the inter-predictability between them. Based on one month of observation, we found that the variation in the amount of mobile phone calls was strongly correlated with the taxi volume of the previous two hours. Hence taxi volume can be used to predict mobile phone call intensity of the next two hours. In addition, we found that the level of inter-predictability varied across different time of the day; taxi was a predictor during PM hours while mobile phone call intensity became a predictor for taxi volume in AM hours. Strong correlations between these two urban signals were observed during active hours of the day and active days of the week.**

I. Introduction

With the rapid growth of urbanization, the need for better services (e.g., public transportation, energy, communications) and urban planning (e.g., infrastructures, environments, policies) demands for better understanding of city dynamics. The development of pervasive technologies such as global system for mobile communications (GSM) and global positioning system (GPS) provides useful tools for sensing social and traffic activities in the city. Analyzing GPS-enabled vehicle traces and mobile phone activity thus provides to some extent an overview of how the city functions.

Today's taxis are equipped with GPS devices for better monitoring and dispatching. Their traces have been used to study different aspects of the traffic network as they provide fine-grained data that reflects the state of traffic flow in the city. Taxi traces typically carry occupancy information from which pick-up and drop-off location information can be easily inferred.

Mobile phone call data, on the other hand, has been used to study social aspect of the city. With its high penetration rate, mobile phone activity data can truly reveal the city's social characteristics.

By examining these two useful sources of data that describe city from different perspectives, in this study, we aim to explore hidden relationship between them – particularly the inter-predictability; *can one data source be used to predict the other?* Although they both explain the city in different ways, we believe that they are related in some way and we aim to explore the underlining relationship in this present work.

The remainder of the paper is organized as follows. Section II briefly describes related work in studying urban dynamics using taxi and mobile phone data. Section III describes the dataset used in this study. Our analysis and results are presented in Section IV. Finally, Section V concludes and summarizes our findings.

II. Related Work

With the advent of the pervasive technologies (e.g. GPS, GSM, Wi-Fi), several work have been presented to explore and improve urban mobility. Among them mining taxi trajectories has recently attracted much attention. Taxi-GSP traces have been used in a number of studies to develop better solutions and services in urban areas such as estimating optimal driving paths [1-3], predicting next taxi pick-up locations [4-8], modeling driving strategies to improve taxi's profit [8-9], identifying flaws and possible improvements in urban planning [10], and developing models for urban mobility, social functions, and dynamics between the different city's areas [11-12].

Yuan et al. [1] present the T-Drive system that identifies optimal route for a given destination and departure time. Zheng et al. [2] describe a three-layer architecture using the landmark graph to model knowledge of taxi drivers. Ziebart et al. [3] present a decision-modeling framework for probabilistic reasoning from observed context-sensitive actions. The model is able to make decisions regarding intersections, route, and destination prediction given partially traveled routes.

Yuan et al. [4] develop a recommender system for both taxi drivers and passengers that takes into account the passengers' mobility patterns and taxi drivers' pick-up traces. Chang et al. [5] propose a four-step approach for mining historical data in order to predict taxi demand distributions

M. Veloso is with Centro de Informática e Sistemas da Universidade de Coimbra, Portugal and Escola Superior de Tecnologia e Gestão de Oliveira do Hospital, Portugal (e-mail: mveloso@dei.uc.pt).

S. Phithakkitnukoon is with Culture Lab, School of Computing Science, Newcastle University, United Kingdom (e-mail: santi@newcastle.ac.uk).

C. Bento is with Centro de Informática e Sistemas da Universidade de Coimbra, Portugal (e-mail: bento@dei.uc.pt)

based on time, weather, and taxi location. They show that different clustering methods have different performances on distinct data distributions. Phithakkitnukoon et al. [6] present a model for predicting the number of vacant taxis for a given area of the city based on the naïve Bayesian classier with their developed error-based learning algorithm and a mechanism for detecting adequacy of historical data. Liu et al. [7] classify taxi drivers according to their income. They observe that top drivers operate in a number of different zones while maintaining exceptional balance between taxi demand and traffic conditions. Ordinary drivers on the other hand operate in fixed zones with few variations.

Ge et al. [8] present an approach for extracting energy-efficient transportation patterns from taxi traces and use it to develop a recommender system for pick-up locations and a sequence of waiting locations for a taxi driver. Zheng et al. [10] identify flawed urban planning in region pairs with traffic problems and the linking structure among these regions through their analysis of taxi traces. Qi et al. [11] investigates the relationship between regional pick-up and drop-off characteristics of taxis and social function of city regions. They develop a simple classification method to recognize regions' social areas that can be divided into scenic spots, entertainment districts, and train/coach stations. Veloso et al. [12] present an exploratory analysis of the spatiotemporal distribution of taxi pick-ups and drop-offs. They investigate downtime (time spent looking for next passengers) behavior, identify taxi-driving strategies, and explore relationship between area type (based on points of interest (POIs)) and taxi flow, as well as the predictability of a taxi trip.

In addition to the dynamic in vehicular network, the mobility of people at the individual level is equally important. With its ubiquity, mobile phones have become human probes for sensing human behavior and social dynamics. Therefore mobile phone data has been used increasingly in various studies aiming to develop universal laws that govern human behavior.

Song et al. [13] study the randomness in human behavior and to what degree individual human actions are predictable by analyzing mobility patterns of mobile phone users. Their results show 93% average predictability in people's mobility. Using cellular network data, Isaacman et al. [14] propose and evaluate three algorithms derived from a logistic regression-based analysis, and describe clustering techniques to identify important locations. They are able to detect home and work locations accurately, which is then used to perform an analysis of commute distance and estimate commuting carbon footprints.

Calabrese et al. [15] present an analysis of crowd mobility during special social events (e.g., sport game, concert) by analyzing mobile phone-location traces. Using data collected from nearly one million mobile phones, the authors are able to correlate social events that people go with their home locations. Using similar a dataset and POI information, Phithakkitnukoon et al. [16] develop the activity-aware map that describes the most probable activities associated with specific areas of a city. Their results show a strong correlation in daily activity patterns between groups of people who share common work area types. Traag et al. [17] describe an approach to correlate human mobility patterns with social events using trajectories of mobile phone users. A probabilistic framework is developed and used to determine the users who participate in a given social event.

The aforementioned studies focus solely on using either taxi or mobile phone call data to study urban functionality and develop intelligent systems. In contrast, this work investigates the relationship between these two data sources that describe city in different ways. To our knowledge, this is the first study to do so and we hope that this study will pave the way for more in-depth investigations in this direction.

III. DATASETS

This work analyses data of mobile phone call intensity and taxi volume in Lisbon, Portugal. The data was collected in December 2009 (a period of 31 days). The area of study corresponds to the municipality of Lisbon, of around 110 km^2, and a population of 800,000 habitants. The city downtown is characterized by a higher population density including touristic, historic and commercial areas. Encircling the city center, there are residential areas surrounding business areas with lower population density. Major infrastructures (e.g., airport and industrial facilities) are located in the city's periphery. The public transportation system consists on bus, metro, train, and ferry. All transportation systems are connected with stations in the city center.

Our taxi dataset was provided by GeoTaxi [18], a company that focuses on software development for fleet management, and holds about 20% of the taxi market share in Portugal. The dataset was composed of around 500,000 taxi-GPS location points and collected from 230 taxis. Along with the GPS location (latitude, longitude) information, it reported speed, bearing, engine status, and occupancy status of the taxi. The amount of pick-ups and drop-offs were inferred, which accounted for 44,731 distinct trips and was termed *taxi volume* in this study.

The overall taxi volume's spatial distribution in Lisbon is shown in Fig. 1 (on 500x500m^2-grid cells), where the number of pick-ups on each cell during the period under study is represented by a color scale (red corresponds to cells with a higher number of pick-ups). Some major locations are identified, such as city downtown (A), airport (B), train stations (C, D) and ferry dock (E). Different public transportation modalities (airport, train, ferry, bus) are well connected through taxi services.

The other dataset was mobile phone call intensity, which was provided by TMN [19], which is one of the main telecommunications operators in Portugal holding about 40% of the market share. The dataset contained information from the traffic channel, which carried speech and data traffic. The data was aggregated hourly, for each cell site, with cleaning and transformation procedures done by the data provider. For our study, only the number of connections successfully started for voice calls was considered for each cell site, which was defined as *call intensity*. Fig. 2 shows a spatial distribution of mobile phone call intensity where each dot represents the location of a cell site, and its size corresponds

to the average amount of calls per hour. Areas with higher call intensity usually present higher taxi volume.

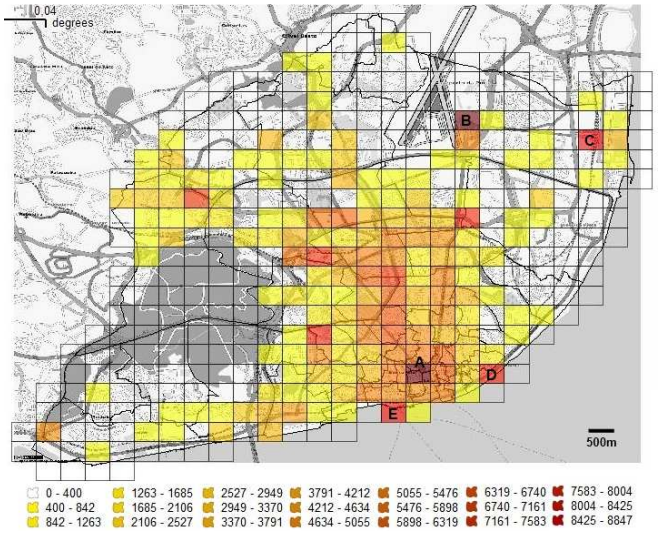

Figure 1. Spatial distribution of taxi volume (number of pick-ups).

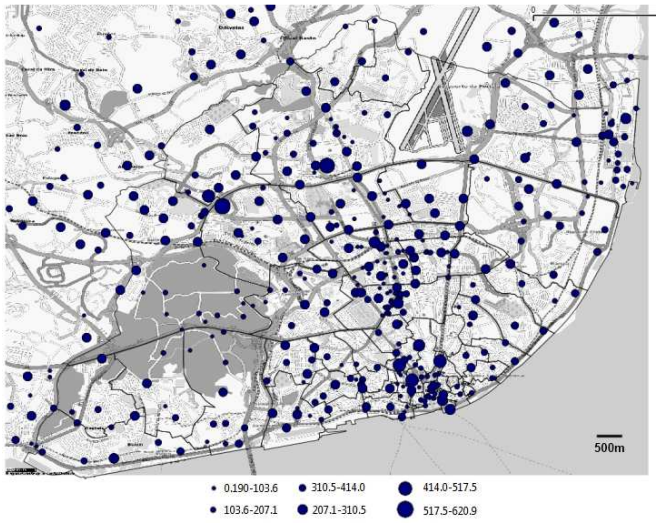

Figure 2. Spatial distribution of cell sites and corresponding mobile phone call intensity (average amount of calls per hour on each site).

IV. Analysis and Results

By examining the temporal distributions of taxi volume and mobile phone call intensity as shown in Fig. 3, we noticed their similar patterns; both gradually increase in the morning around 7am, stay highly active, and then drop down slowly in the evening around 7pm. In addition, we observed that mobile phone call intensity appeared to follow the taxi volume with an approximate gap of about 1-2 hours.

To further explore this relationship, we extracted data as hourly aggregated time series. Since they both have different units, the time series were thus normalized (by the sum) to [0, 1]. We overlaid these time series on the same plot as shown in Fig. 4 and observed similar temporal patterns. Both exhibited daily cycles. Mobile phone call intensity reached almost zero (minimal activities) between midnight and 6AM while high values appearing around noon. Taxi volume time series appeared to follow this similar pattern with low values emerged during off-peak hours (little after midnight until early morning). It is also observable a reduction of the activity from both services on weekends and holidays.

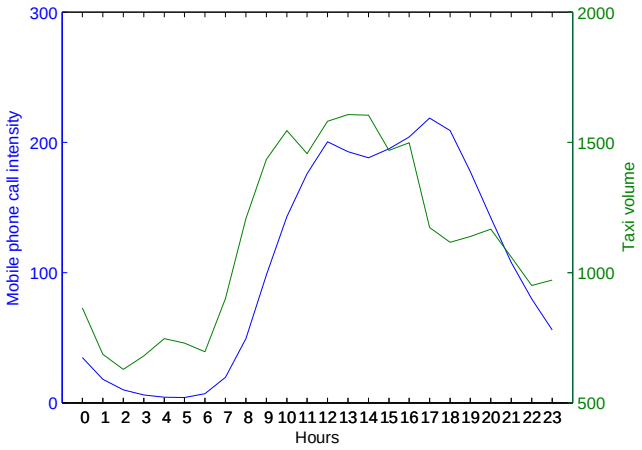

Figure 3. Temporal distribution of mobile phone call intensity (blue) and taxi volume (green) across different time of the day where 0 implies midnight to 1AM, 1 implies 1AM-2AM, and so on.

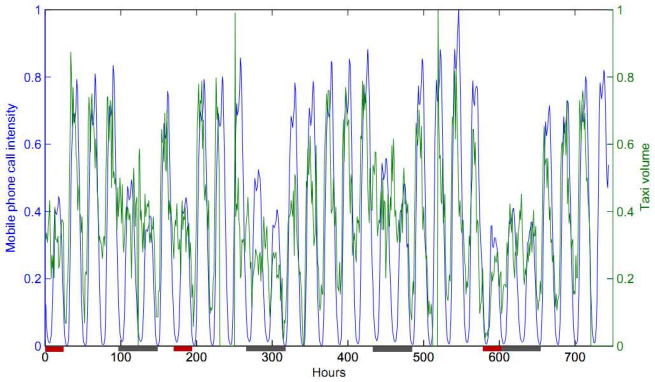

Figure 4. Normalized time series of mobile phone call intensity (blue) and taxi volume (green) over 31 days of observation. The grey line on x-axis represents the weekend periods while the red line corresponds to the holidays (December 1st, 8th, and 25th).

To quantify the difference between these two time series, we computed the Euclidean distance (ED) as follows:

$$ED_i = \sqrt{(g_i - t_i)^2} = |g_i - t_i| \qquad (1)$$

where g_i represents the mobile phone call intensity at hour i and t_i denotes taxi volume at hour i. Hence $G = \{g_1, g_2, \ldots, g_n\}$ and $T = \{t_1, t_2, \ldots, t_n\}$ represent the normalized time series of GSM call intensity and taxi volume of length n, respectively.

Euclidean distance of these time series turned out to be 0.2100 and its hourly distances are shown in Fig. 5.

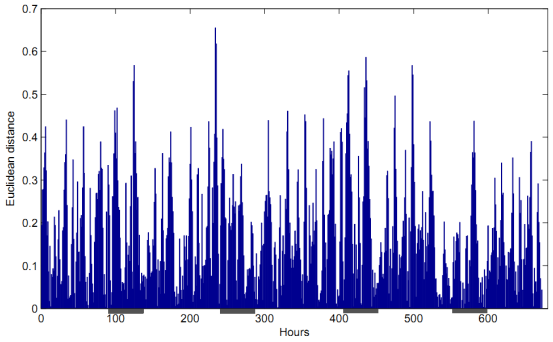

Figure 5. Hourly Euclidean distance of the normalized time series of mobile phone call intensity and taxi volume. The grey line on x-axis represents the weekend periods.

Furthermore, we observed daily and weekly cycles. Through our examination of the data, we found that the highest similarity between these time series was during 8AM to 10PM (active hours) with the Euclidean distance of 0.1917. The hourly distance is shown in Fig. 6.

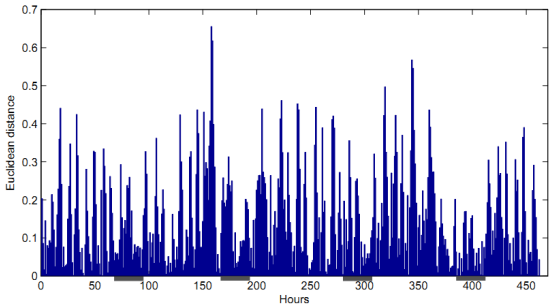

Figure 6. Hourly Euclidean distance of the normalized time series of mobile phone call intensity and taxi volume during 8AM to 11PM for which the overall distance was found to be the lowest at 0.1917. The grey line on x-axis represents the weekend periods.

From weekly cycle perspective, weekdays that are associated with more activities (mostly repeated activities in temporal orders such as commuting to work, having lunch at the same time and same restaurant, making a phone call before arriving at home, and so on) than weekends unsurprisingly yielded more correlated behaviors between mobile phone calls and amount of taxis. The Euclidean distances were 0.2094 and 0.2251 for weekdays and weekends, respectively. The Standard Deviations were 0.13278 and 0.1401 for the same periods. Fig. 7 shows hourly distance of weekdays.

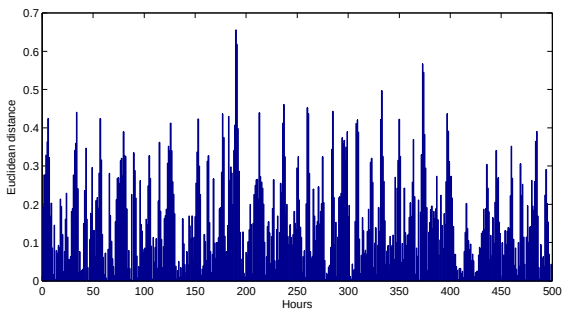

Figure 7. Hourly Euclidean distance of the normalized time series of mobile phone call intensity and taxi volume during weekdays.

We have so far observed that there is a correlation between the two i.e., their values vary in a similar way, especially during active hours of the days (8AM-10PM) and active days of the week (weekdays). We then wanted to investigate further in terms of predictability between them. More specifically, *can one data source be used to predict the other and to what extent?*

To do so, we employed the *coefficient of determination* or R^2 (that is widely used for regression analysis) to measure the interdependency between these two urban signals for different time shifts. The time shifting was used here to examine the predictability that one had on the other. For example, one-hour lag of X yields a high R^2 value with Y implies that X is likely a one-hour predictor of Y i.e., the variation in values of X suggest a similar variation in values of Y of the next hour.

The coefficient of determination or R^2 can be calculated as:

$$R^2 = \frac{\sum_i (y_i - \bar{y})^2 - \sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \quad (2)$$

where $\bar{y}$ is the mean and $\hat{y}$ denotes the predicted value of y (i.e., $\hat{y}_i = a + bx_i + \varepsilon_i$).

By fixing mobile phone time series and shifting taxi time series between -5 hours to +5 hours (e.g., -5 hours of time shift means considering mobile phone data at time t against taxi data at time t-5 hours), we discovered that at time shift of -2 hours the two data sources had the highest correlation. As shown in Fig. 8, at time shift of -2 hours the Euclidean distance and R^2 values were 0.1563 and 0.8512, respectively. This suggests that generally the taxi volume is a 2-hour predictor of mobile phone intensity. In other words, *the variation in the amount of taxis is an indicative variable for the mobile phone call intensive of the next two hours.*

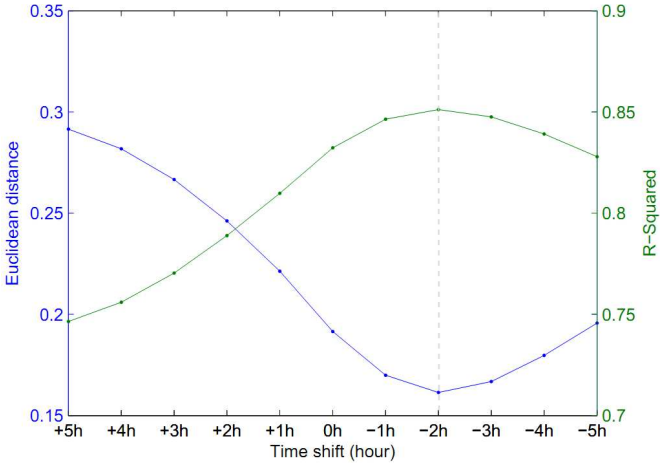

Figure 8. Fitting results for the sliding window between GSM and taxi data.

The hourly Euclidean distance of this 2-hour difference comparison is shown in Fig. 9. The plot of the normalized taxi volume against the normalized mobile phone call intensity is shown in Fig. 10 along with the fitted linear function $y = p_1 x + p_2$, where p_1 = 0.88417, p_2 = 0.092295, and R^2 = 0.8512.

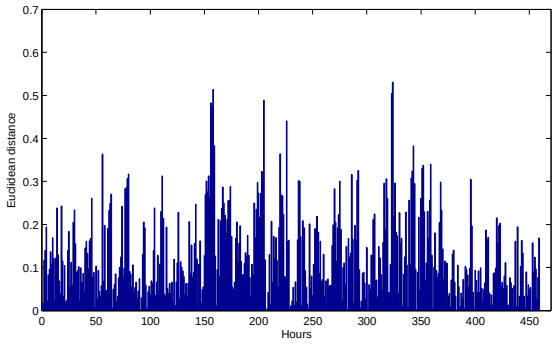

Figure 9. Hourly Euclidean distance of the normalized time series of mobile phone call intensity and taxi volume of the time shift of -2 hours (i.e., comparing mobile phone data at time *t* with taxi data at time *t*-2.

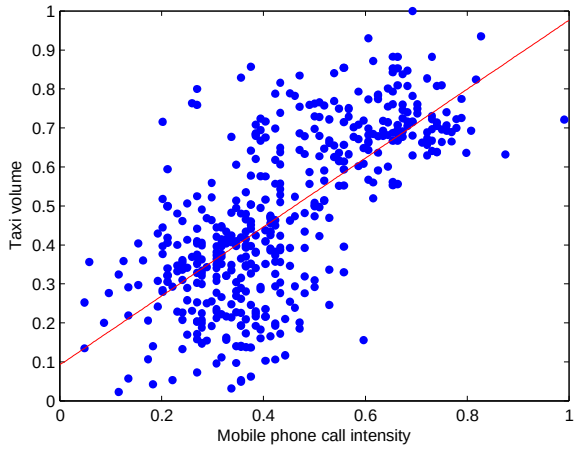

Figure 10. The fitted linear function of the normalized taxi volume (at time *t*-2) against the normalized mobile phone call intensity (at time *t*) with R^2 = 0.8512.

Having observed strong correlations at active hours of the day and active days of the week, as well as 2-hour time difference led us to a further investigation of how this inter-predictability varies across different time of the day.

Similar to the previous approach, by keeping the normalized mobile phone time series fixed while shifting taxi time series between -5 and +5 hours, we computed R^2 values across varying time shifts for each different hour of the day. The result is shown in Fig. 11 where this inter-predictability was observed to change over time. It turned out that there were strong inter-predictabilities (correlations) during active hours of the day, which was line with our previous observation. *Interestingly, we found that during the active hours, mobile phone call intensity was a predictor for taxi volume in AM hours and the relationship was reversed as the taxi volume became a predictor for mobile phone call intensity in the PM hours.* Hence at noon hour there was a strong correlation at 0 time shift. In other words, variations in both urban signals were well synchronized at around midday.

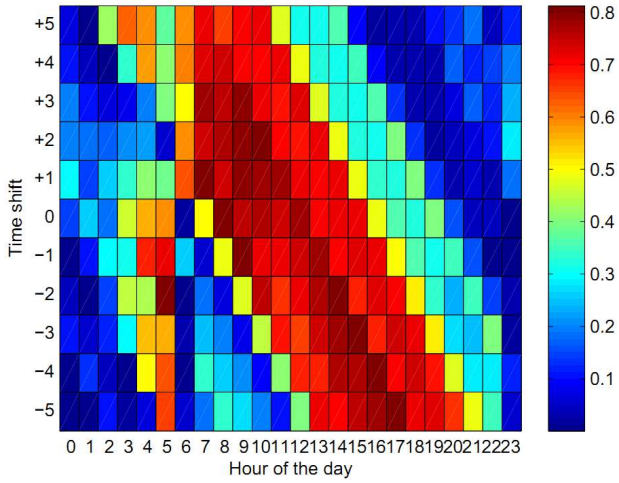

Figure 11. Pseudocolor plot of R^2 values across varying time shifts of different hours of the day.

We believe that our findings to some extent unveil relationship between two different urban signals; as one describes sociality of the city while the other characterizes state of traffic flow. The findings are useful for developing efficient intelligent transportation systems as they provide the link between social and transportation networks.

V. Conclusion

In this work, we explored a relationship between the taxi volume and mobile phone call intensive in Lisbon, Portugal. Particularly we were interested in the inter-predictability between these two urban signals. Based on one-month of data, we found a strong correlation between them during active hours of the day (8AM-10PM) and active days of the week (weekdays). Moreover, we also discovered that mobile phone call intensity had a strong correlation with taxi volume of the previous two hours, which means that the amount of taxis can be used to predict the intensity of mobile phone calls of the next two hours. Furthermore, we found that this inter-predictability varied across different time of the day. Intensity of mobile phone calls was a predictor of taxi volume in morning hours while the amount of taxi flow became a predictor of mobile phone calls in the afternoon and evening.

Nonetheless, there were a number of significant limitations to our study. The first of these is the limited amount of data used. Only one month of data were available to us at time of this study, which limited our observation from which our results were obtained. Another potential limitation is the linear relation that was assumed between our two data sources in this study. Further investigation thus needs to be done in finding the most suitable function for their relationship. A final limitation related to the extent to which our findings are applicable beyond the city of Lisbon. As urban area of a First World (developed) country and a member of the Schengen area, Lisbon has significant similarities with many European and other developed cities in the world. We thus believe that the findings are likely to

be applicable to cities with broadly similar social, cultural, and economic profiles.

This study sheds light on the multi-source urban data fusion for better understanding of urban functionality and developing efficient transportation systems. We hope that our findings suggest new ways to use multi-source data to investigate the interplay between different urban entities.

VI. ACKNOWLEDGMENTS

This research was carried out under the framework of the project CityMotion (Data Fusion for Mobility Consumers, Providers, and Planners) in the MIT-Portugal program. The authors gratefully acknowledge TMN and GeoTaxi for providing the data for this study.

REFERENCES

[1] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, Y. Huang,"T-Drive: Driving Directions Based on Taxi Trajectories," in Proc. ACM SIGSPATIAL GIS 2010, Association for Computing Machinery, Inc. 1 (2010), 99-108.

[2] Y. Zheng, J. Yuan, W. Xie, X. Xie, G. Sun, "Drive Smartly as a Taxi Driver.," in 7th Int. Conference on Ubiquitous Intelligence & Computing and 7th Int. Conference on Autonomic & Trusted Computing (UIC/ATC) (2010), 484-486.

[3] B.D. Ziebart, A.L. Maas, A.K. Dey, J.A. Bagnell, "Navigate like a cabbie: probabilistic reasoning from observed context-aware behavior," in: UbiComp '08: Proc. of the 10th int. conf. on Ubiquitous computing, New York, NY, USA, ACM (2008), 322-331.

[4] J. Yuan, Y. Zheng, L. Zhang, X. Xie, G. Sun, "Where to Find My Next Passenger?," in 13th ACM Int. Conf. on Ubiquitous Computing (UbiComp 2011), China (2011).

[5] H. Chang, Y. Tai, J.Y. Hsu, "Context-aware taxi demand hotspots prediction," in Int. J. Bus. Intell. Data Min. 5(1) (2010), 3-18.

[6] S. Phithakkitnukoon, M. Veloso, C. Bento, A. Biderman, C. Ratti, "Taxi-Aware Map: Identifying and predicting vacant taxis in the city," in Proc. AmI 2010, First International Joint Conference on Ambient Intelligence (2010), 86-95.

[7] L. Liu, C. Andris, A. Biderman, C. Ratti, "Uncovering cabdrivers' behavior patterns from their digital traces," in Computers, Environment and Urban Systems, 2010.

[8] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, M. J. Pazzani, "An Energy-Efficient Mobile Recommender System," in Proc. KDD 2010, ACM Press (2010): 899-908.

[9] L. Liu, C. Andris, A. Bidderman, C.Ratti, "Revealing taxi drivers mobility intelligence through his trace," Movement-Aware Applications for Sustainable Mobility: Technologies and Approaches, (2010), 105-120.

[10] Y. Zheng, Y. Liu, J. Yuan, X. Xie, "Urban Computing with Taxicabs," in 13th ACM Int. Conference on Ubiquitous Computing (UbiComp 2011), China (2011).

[11] G. Qi, X. Li, S. Li, G. Pan, Z. Wang, D., Zhang, "Measuring Social Functions of City Regions from Large-scale Taxi Behaviors," in PerCom- Workshops 2011, pp. 21-25, Seattle, USA, (2011).

[12] M. Veloso, S. Phithakkitnukoon, C. Bento, P. Olivier, N. Fonseca, "Exploratory Study of Urban Flow using Taxi Traces," in First Workshop on Pervasive Urban Applications (PURBA) in conjunction with Pervasive Computing, San Francisco, California, USA (2011) .

[13] C. Song, Z. Qu. N. Blumm, A. Barabási, "Limits of Predictability in Human Mobility," in Science Vol. 327 no. 5968 pp. 1018-1021, (2010).

[14] S. Isaacman, R. Becker, R. Caceres, S. Kobourov, M. Martonosi, J. Rowland, A. Varshavsky, "Identifying Important Places in People's Lives from Cellular Network Data," in Proc. of the Int. Conference on Pervasive Computing, San Francisco, California, USA, (2011).

[15] F. Calabrese, F. Pereira, G. Di Lorenzo, L. Liu, and C. Ratti, "The geography of taste: analyzing cell-phone mobility and social events," in Proceedings of the 8th International Conference on Pervasive Computing. Springer (2010).

[16] S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti, "Activity-aware map: Identifying human daily activity pattern using mobile phone data," in Proceedings of the 10th International Conference on Pattern Recognition. IEEE (2010).

[17] V.A. Traag, A. Browet, F. Calabrese, F. Morlot, "Social Event Detection in Massive Mobile Phone Data Using Probabilistic Location Inference," in IEEE SocialCom (2011).

[18] Geotaxi, http://www.geotaxi.com/.

[19] TMN - Telecomunicações Móveis Nacionais, http://www.tmn.pt/.