Overview of taxi database from viewpoint of usability for traffic model development: a case study for Budapest

Árpád Varga*, György Eigner†, Levente Kovács†, Imre Felde‡ and Miklós Mezei*

*John von Neumann Faculty of Informatics, Óbuda University, Budapest, Hungary
Email: felde.imre@nik.uni-obuda.hu
†Research, Innovation and Service Center of Óbuda University,
Physiological Controls Group, Óbuda University, Budapest, Hungary
Email: {eigner.gyorgy,kovacs.levente}@nik.uni-obuda.hu

***Abstract*—Forecasting and analyzing urban car traffic is an actual but still very complex problem. The modern car fleet handling IT systems designed for taxi and delivery service companies allows GPS coordinate data acquisition from large amount of vehicles for optimizing the ride and freight allocation. Since the database of these companies contains movement patterns belonging to multitude of vehicles, arise the question if these data, belonging to vehicles with special purpose, is suitable for representing the whole car traffic. To make the first step to answer this question, our case study utilizes the time-resolved GPS coordinate database of one of the largest taxi company in Budapest from year 2014.**

***Index terms*—Data mining, Taxi cab database, Data visualization, Urban traffic modeling**

I. Introduction

During the XX. century the proportion of the urban population show unprecedented increase. Nowadays more the 54% of total human population live in cities or in urban regions [1]. In case of Hungary, this proportion is 68%, and more than 16% of the country's population is concentrated surroundings the capital, Budapest [2]. The urban transportation networks are exposed due to the increased load. Thus, the optimization, design, and modeling of these systems became important topics. By the help of suitable arranged and organized urban transportation network considerable cost reduction can be achieved. On the other side, since the majority of the vehicles still use fossil fuel, the emission of exhaust gases can be moderated, which has positive feedback to the state of health of the urban population also.

To improve the design and optimization methods also require the development and elaboration of sophisticated mathematical models, which can describe and predict the main characteristics of the urban car traffic. In general, the traffic is characterized by two quantities, the car density ρ [vehicles/km], and the traffic flow q [vehicles/h], which can vary in time and space. For highway traffic, the connection between these two quantities in a given road section is approximately given by the "fundamental relationship" [3]. For low ρ values, the q is linear function of ρ. For higher q values, the individual vehicles tend to prohibit the movement of each other, resulting folding back and a decreasing segment in the fundamental relationship curve.

This nonlinear characteristics – in case of existence of several conditions – can cause the formation and propagation of congestion waves. For highway traffic, the evolution and propagation of these waves can be described and predict mathematically well by using nonlinear partial differential equations. Since the modern highways are densely equipped with monitoring stations, capturing the type (automobile/truck) and speed of the passing vehicles, these models can be validated relatively easily [4].

For urban traffic, both the model creation and its validation is far more complex task [5]. Unlike the highways, which are dominated by long, straight sections with relatively few junctions, the urban roads are connected with numerous intersections, producing a complex network with ten thousands of road sections and junctions [6]. The traffic flow capacity of the road sections and junctions varies between wide limits (small streets of the city center region and the suburbia vs. the slip road of the highways which connect the city with the national road system). While the traffic flow and the main travel paths show high level of predictable seasonality and periodicity in various time scales (school starting period vs. the holiday period in summer, working day vs. the end of the week, morning rush hours vs. the night traffic), unforeseen or proposed events, such as accidents, resurrections or large scale touristic (sport or cultural) programs can totally change this temporal pattern. Moreover, since the urban roads are

Gy. Eigner was supported by the ÚNKP-17-4/IV. New National Excellence Program of the Ministry of Human Capacities.

978-1-5386-3855-2/17/$31.00 ©2017 IEEE

characterized by high level of connectivity, the effect of events can propagate across the elements of the urban traffic network, causing strong deviations from the overall seasonal-temporal pattern even far away from the initial place of the event.

Among several issues of this kind of complex network modeling, also arises the question of validation: is it possible to find a database, which contains real traffic data from all regions of the diverse urban road network? There are also exist fix official installed traffic monitoring stations in the urban regions, but they are concentrated in the main junctions and may not represent the diversity of the urban roads. On the other hand, the modern car fleet handling IT systems designed for taxi and delivery service companies allow GPS coordinate data acquisition from large amount of vehicles for optimizing the ride and freight allocation [7]. Since the vehicles belonging to these companies travel considerable amount of distance daily, and they are both rove the inner and outer regions of the city, it is reasonable to use this kind of data to validate various traffic models. Moreover, these companies are often handling several hundred vehicles simultaneously, and each of them can be considered as a passive ”probe”, which is moving together with the surrounding traffic.

In our paper, we make an attempt to visualize and summarize the main temporal and spatial properties of the taxi GPS coordinate database belonging to one of the largest taxi company in Budapest, regarding to the year 2014.

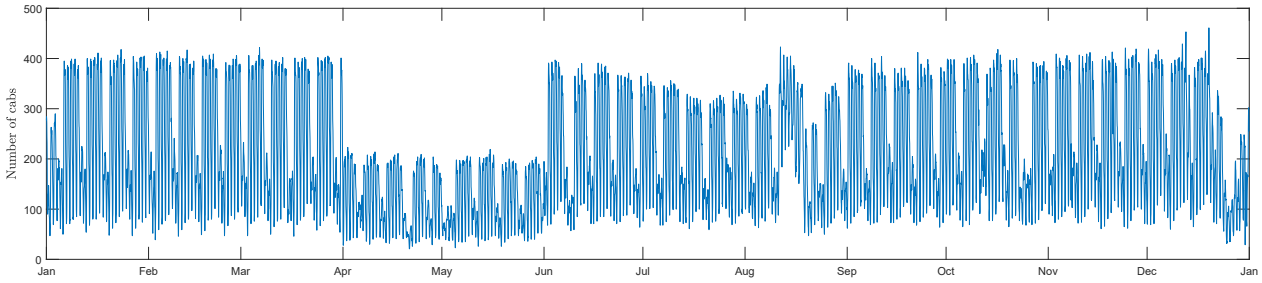

Figure 1: Variation of the number of active taxi cabs during the whole year of 2014.

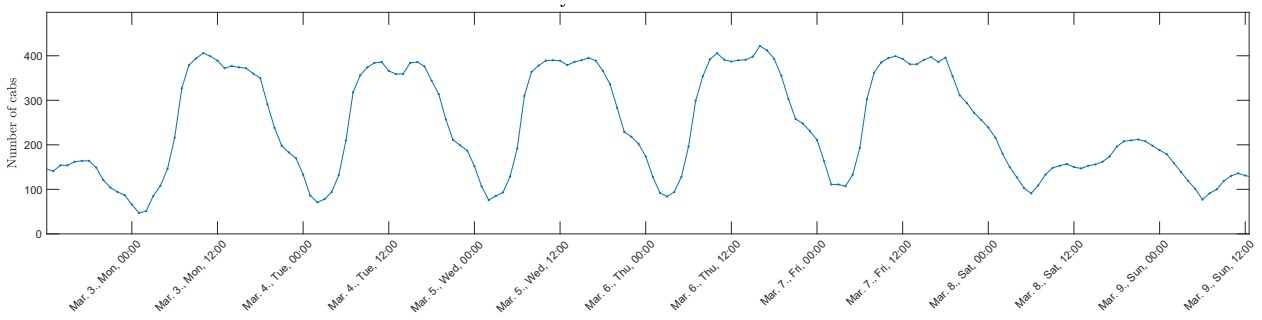

Figure 2: Variation of the number of active taxi cabs for a typical week of 2014.

II. Temporal Overview of Taxi Data

Each taxi cab from the database is equipped with commercial tablet PC, which ensures the connection with the dispatcher center via mobile internet connection. The tablets are also possess a built-in GPS receiver and the actual coordinates of the vehicle (geographical longitude and latitude) are determined from time to time, and sent to the dispatcher center together with the present state of the taxi (engaged, vacant, not in service, en route for an incoming carriage request). The instantaneous cab positions and states are used to allocate the incoming carriage requests. Several possible routes are calculated between the actual client calling places and the nearest vehicles, and the most preferable is chosen based on the shortest access time. The planned route is also sent to the tablet of the chosen vehicle, helping the work of the taxi driver. The central distribution system is also capable to consider the traffic situation and also try to uniform the load of the individual vehicles. Eg., if two taxi cab have similar access time to the same customer, then that cab is sent to the location of order, which has fewer carriage in the actual shift. The latitude, longitude coordinate logs, equipped with timestamp and a unique vehicle identification number, are stored in a central log file for later analysis and revision.

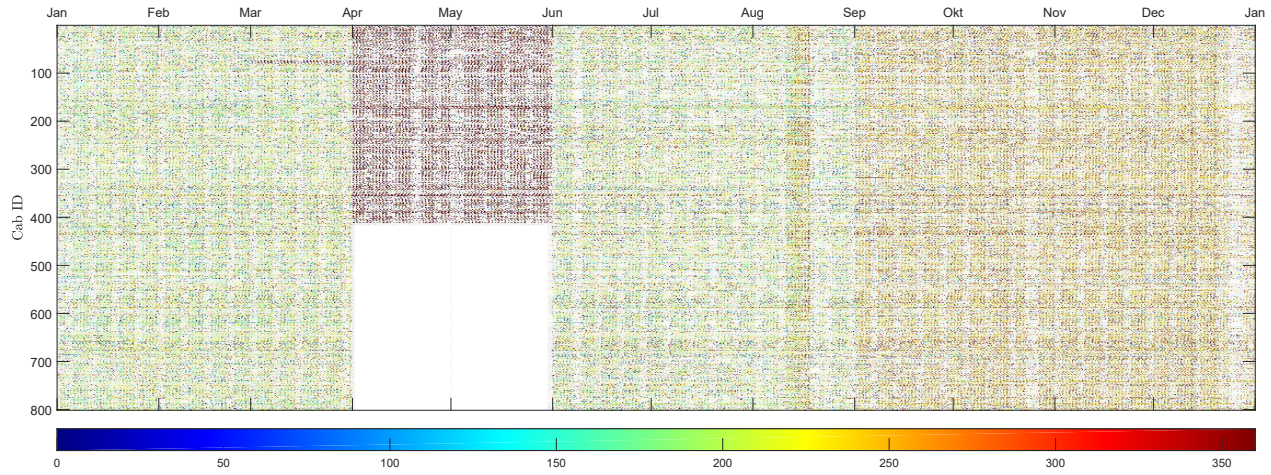

Figure 3: Variation of the number of hourly incoming coordinate logs for all taxi cabs during the whole year of 2014.

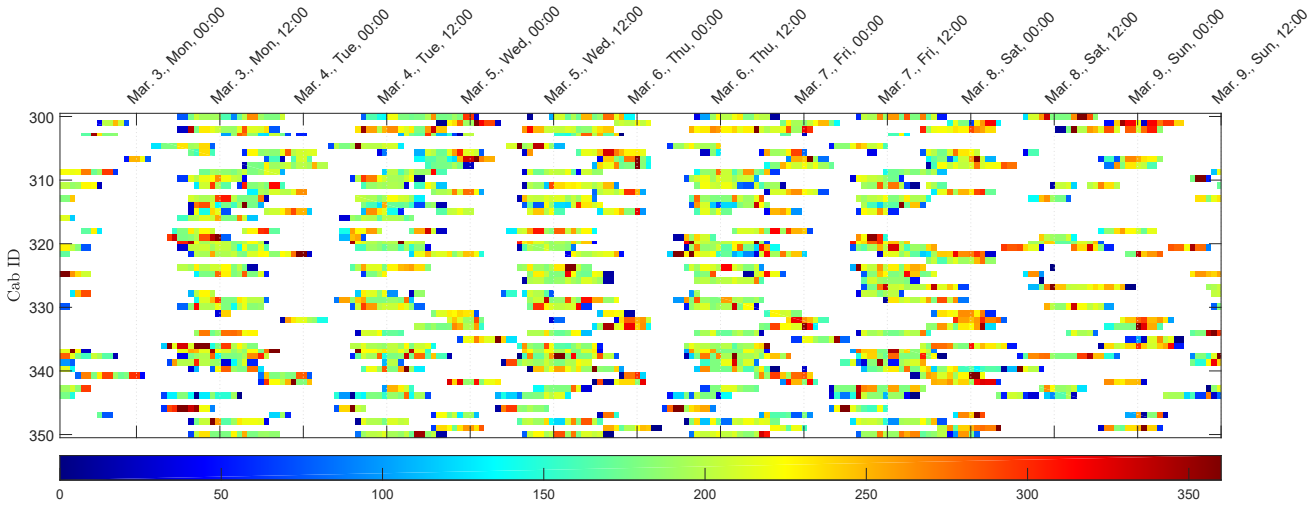

Figure 4: Similar to Fig. 3. shown for a typical week in case of taxi cabs with ID 300-350. Hours are colored with white if has no incoming coordinate logs.

In our case, the size of the central log file, concerning to the year 2014, is 36 GB and stored in plain text format. It contains more than $4.5 \cdot 10^8$ position logs belonging to 801 individual cabs. The incoming latitude and longitude coordinate pairs are denoted with a time stamp with 0.01 sec precision. To analyze the temporal distribution and continuity of the captured coordinates, the log file is parceled to 8760 section based on the time stamp, distinguishing the hours in year 2014. Then, the number of coordinate logs, belonging to different taxi cabs is counted in each hour. A taxi cab in a given hour is considered as an 'active' taxi cab, if there exists at least 1 coordinate log belonging to its ID in the given interval, regardless its state (since the vacant, free, and officially not in service vehicles can provide equally valuable information from viewpoint of the traffic monitoring).

The variation of the number of active taxi cabs per hour during the whole year and for a typical week is shown in Fig. 1-2. The minimum, maximum, and mean hourly active vehicle numbers are 21, 461 and 207.83. The fact that there is no hour passed without less than 21 active taxis, assume that the continuous traffic monitoring can be realized based on the taxi log data, since the presence of these vehicles in the traffic is constant. However, the number of active cars shows strong fluctuations, in accordance with the general travelling habits. The peak of the active hourly vehicle number are usually reached between 8:00-11:00 or 14:00-16:00 during and ordinary weekday, the minimum values are usually occurs in a period 2:00-3:00 at nighttime. Peaks belonging to the weekends or holidays reach only about the half of the weekdays. During April and May a strong decrease can be

observed in the active car numbers, since about the half of the cabs were temporarily withdrawn from circulation due to technical reasons.

Figure 5: Scatter plot of the captured position logs (”street scatter”) from all vehicles during Jan. 7., Tuesday, 2014. The position logs belonging to cab with ID No. 118. are connected and colored by the time of recording, calculated from the time stamp of the first position sample.

Another important question is the temporal density of coordinate logs incoming from the cabs, since it determines the temporal resolution of traffic report, based on taxi data. To get an overall view about the temporal log density, the hourly coordinate log numbers are counted for each individual taxi cab. Then, the hourly captured log numbers are depicted as a waterfall diagram (Fig. 3-4.), where the horizontal axis represents the date, and the cab ID-s are indicated on the vertical axis. Therefore, each square or pixel in the diagram represents an hour in a year, belonging to a specific taxi, colored by the log number. In Fig. 3 there is an empty territory without coordinate logs between April and May for cabs with ID 410-801, because the withdrawn of these vehicles from circulation were planned, since they had to prepared according to the requirements of the new taxi policy, which was introduced in 2013. The length of a typical shift for a taxi cab was 4-16 hours (of course, the drivers can be changed in that period). The number of hourly incoming coordinate samples shows high level of variability, but for the majority of the investigated hourly periods is between 180-360 log/hour, which means that the mean elapsed time between two coordinate sample is about 20-10 sec. It is also interesting, that the subtraction of about the half of the fleet results remarkable growth in the hourly log numbers, which indicates that the coordinate logging performance is sensitive to the number of the actual running cabs. From September there is again a noticeable rise in the hourly log numbers, since the IT systems of the company is developed.

III. Spacial Overview of Taxi Data

The precision of the positioning is approximately 5-10 m, and it is strongly depend on the surface conditions. Eg. in tunnels and in the vicinity of hills and larger buildings the signal of the GPS satellites are often overshadowed, resulting inaccurate position logs. To investigate the spatial distribution of the incoming coordinates in general, a scatter plot is drawn based on the coordinate logs from all vehicles on a typical weekday (Jan. 7., Tuesday, Fig. 5.), where each point represents an incoming coordinate sample. The structure of the downtown streets of Budapest is clearly outlined by the scatter, even though the considered data are concerning only a one day period. In the 2014 annual log file, several coordinates are labeled as a potential street crossing (registered manually by the taxi drivers), in Fig. 5. they are also depicted. The majority of the potential junctions are true crossings, but some of them depict false crossing points. To characterize the resolution of motion of an individual cab, the positions of a single vehicle are highlighted and consecutive positions of the chosen vehicle are connected. The movement of the vehicle can be followed even through the street network of the downtown, mainly characterized by short street sections.

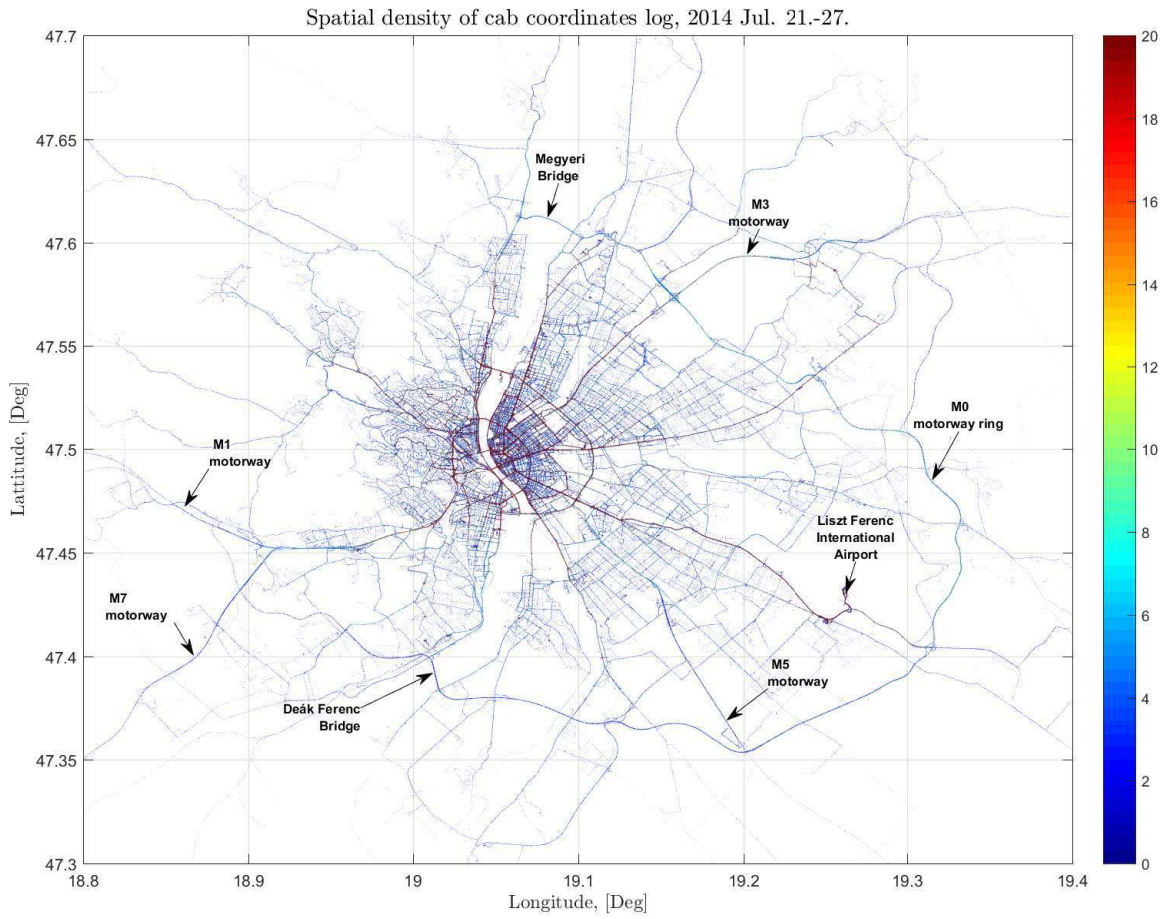

Figure 6: Taxi attendance histogram of Budapest corresponding to period Jul. 21. Mon.-Jul. 27. Sun. Empty cells are colored with white.

On the scatter plot it is clearly visible, that some streets and street sections are covered by more position samples than the others, indicating, that these routes are more often used by taxi cabs. However, for longer time periods, the scatter diagram is not suitable to visualize the "attendance" frequency of different streets by taxis, since the density of scatter points become indistinguishable. To make the depiction more expressive, first, the territory of Budapest and the connecting suburbia (a quadratic region with dimensions of 45×45 km), is parceled by a 8000×8000 quadratic mesh. The size of an individual mesh element is 5.625×5.625 m. Then, the number of position samples, falling in each quadratic cell, is counted, and colored by the count number, resulting an attendance histogram. For 1 week period (between Jul 21. Mon. – Jul. 27. Sun) for all taxi it is shown in Fig 6-7. As it is expectable, the downtown territories are far more attended in general, then the suburbia. At the same time, some suburban roads (typically: highway introductory sections, or road to the international airport) are also characterized by higher attendance frequency level. The most densely used road sections in the downtown area are the boulevards and the bridges. It can be also noticed, that some colored parcel is located in the Danube between the bridges (naturally, they represent invalid coordinate samples), which can be explained by the inaccuracy and limitations of the GPS system, as it was mentioned before.

IV. Conclusion

Several main issues of the urban traffic modeling are summarized and the operation of a coordinate data acquisition system, belonging to a modern typical taxi company is presented. Several possible graphical methods are displayed to investigate, understand and estimate the temporal and spatial qualities of such a complex and large-scale dataset. Since the taxi cab tracking data provide temporarily and spatially extensive picture about the movement of great amount of vehicles, it is possible to use it as a base for validation of traffic models. This validation will be done in our future research.

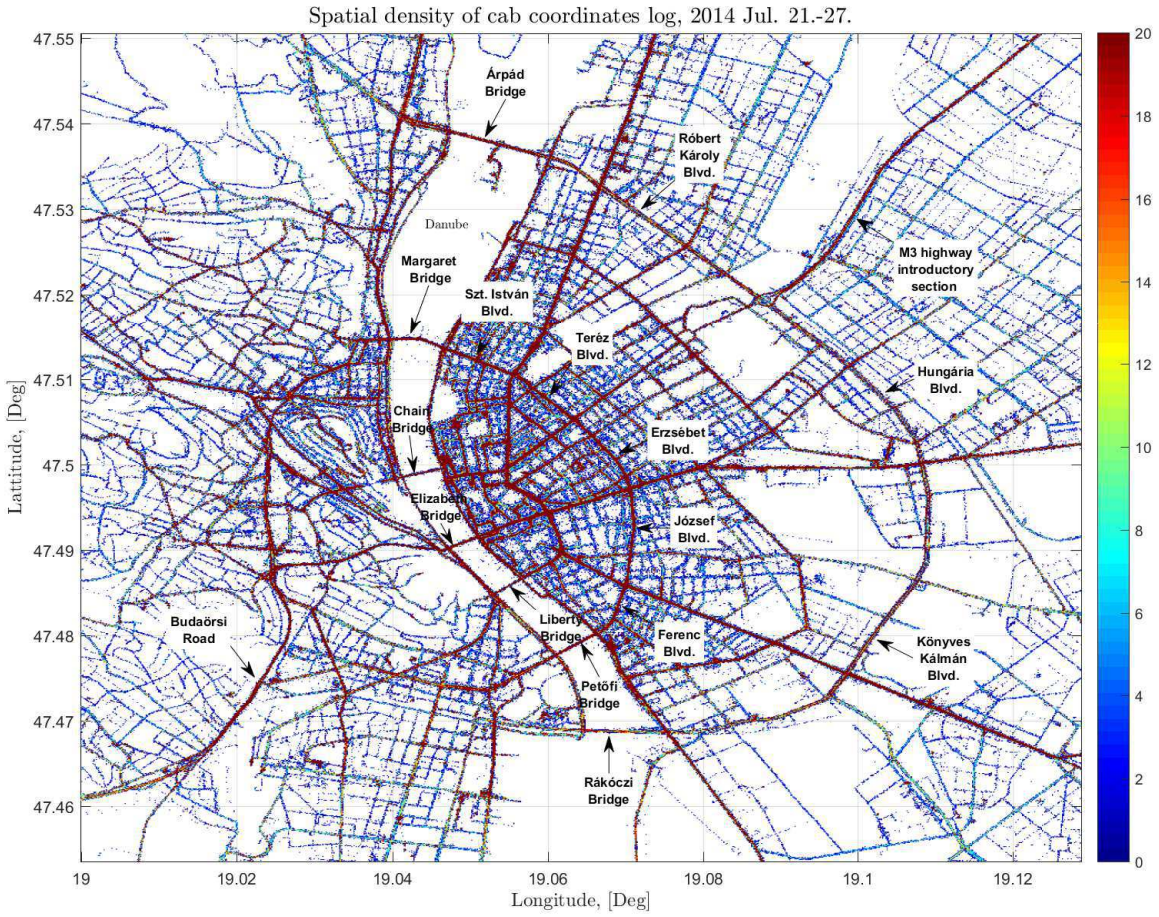

Figure 7: Same attendance histogram as in Fig. 6, but for the downtown of Budapest

ACKNOWLEDGMENT

Á. Varga, M. Mezei and Gy. Eigner acknowledge the support of the Robotics Special College of Obuda University and the Doctoral School of Applied Informatics and Applied Mathematics of Óbuda University. The research was also supported by the Research and Innovation Center of Óbuda University.

REFERENCES

[1] United Nations, "World urbanization prospects: The 2014 revision, highlights. department of economic and social affairs," *Population Division, United Nations*, 2014.

[2] *Population Census 2011*, 1st ed., Hungarian Central Statistical Office, 2012, 1. Preliminary data.

[3] M. Treiber and A. Kesting, "Validation of traffic flow models with respect to the spatiotemporal evolution of congested traffic patterns," *Transportation research part C: emerging technologies*, vol. 21, no. 1, pp. 31–41, 2012.

[4] ——, "Evidence of convective instability in congested traffic flow: A systematic empirical and theoretical investigation," *Transportation Research Part B: Methodological*, vol. 45, no. 9, pp. 1362–1377, 2011.

[5] A. Jamshidnejad, I. Papamichail, M. Papageorgiou, and B. De Schutter, "Sustainable Model-Predictive Control in Urban Traffic Networks: Efficient Solution Based on General Smoothening Methods," *IEEE Transactions on Control Systems Technology*, 2017.

[6] S. Zhao, P. Zhao, and Y. Cui, "A network centrality measure framework for analyzing urban traffic flow: A case study of Wuhan, China," *Physica A: Statistical Mechanics and its Applications*, vol. 478, pp. 143–157, 2017.

[7] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva, "Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2149–2158, 2013.