Clustering of Mobile Subscriber's Location Statistics for Travel Demand Zones Diversity

Marcin Luckner, Aneta Roslan, Izabela Krzemińska, Jaroslaw Legierski, Robert Kunicki

Clustering of mobile subscriber's location statistics for travel demand zones diversity

Marcin Luckner[1], Aneta Rosłan[1], Izabela Krzemińska[2], Jarosław Legierski[1 2], Robert Kunicki[3]

[1] Warsaw University of Technology
Faculty of Mathematics and Information Science
00–662 Warsaw, ul. Koszykowa 75, Poland
mluckner@mini.pw.edu.pl,aneta.roslan@gmail.com
[2] Orange Polska S.A.
Orange Labs Poland
02-691 Warsaw, ul. Obrzeżna 7, Poland
izabela.krzeminska@orange.com, jaroslaw.legierski@orange.com
[3] The City of Warsaw
Department of Computer Science and Information Processing
00-095 Warsaw, pl. Bankowy 2, Poland
rkunicki@um.warszawa.pl

Abstract. Current knowledge on travel demand is necessary to keep a travel demand model up to date. However, the data gathering is a laborious and costly task. One of the approaches to this issues can be the utilisation of mobile data. In this work, we used mobile subscriber's location statistics to define a daily characteristic of mobile events occurrences registered by Base Transceiver Stations (BTS). For types of preprocessed data were tested to create stable clusters of BTS according to registered routines. The obtained results were used to find similar travel demand zones from the Warsaw public transport demand model according to a daily activity of the citizens. The obtained results can be used to update the model or to plan a cohesive strategy of public transport development.

1 Introduction

Knowledge on travel demand is a key aspect of a transport development. It can be gathered using a detailed survey and passenger counting and used to build up detailed travel demand models.

However, the model may be very fast out of date in dynamically growing metropolises. It is not possible to repeat wide passengers research very often. We need a method to update the travel demand model continuously.

Mobile subscriber's location statistics is a data source that can be used to update the model. A daily distribution of events registered in a Base Transceiver Station (BTS) creates a characteristic of the area covered by BTS's signal. Next, similar characteristics can be grouped to create an area of common daily characteristic. The areas represent a different daily distribution of the citizens and can be used to characterise travel demand zones.

In this work, we used the Self-Organizing Map (SOM) to cluster similar mobile subscriber's location daily statistics. Four types of preprocessed input data were discussed. The data were used to create clusters of the same characteristic. The clusters with the most appropriate properties were merged with the existing travel demand zones in the City of Warsaw. As the result, we obtained the travel demand zones divided into clusters of the same daily citizens' activity.

The rest of this work is structured as follows. Section 2 briefly describes related works. Section 3 presents used data: Mobile subscriber's location statistics and the travel demand zones. Section 4 describes the used clustering model and a data preparation. Section 5 discuss the obtained results. Finally, Section 6 presents conclusions and future works.

2 Related work

In our model, we try to update a travel demand models using mobile data. Work [2] presented an alternative solution of counting and surveying to clarify the demand data. The paper dealt with the issue of needed sample size to produce reliable results. The same authors proposed in Work [1] a theory of iterative estimation method of reliable passenger data. Work [4] presented the system based on transportation software-VISUM. The system integrated real–time data with static planning model. A similar solution is applied in our work, where the static travel demand zones are merged with the dynamic mobile subscriber's location statistics. Work [8] presented an alternative calibration of a Dynamic Traffic Assignment (DTA) model. Demand matrices were estimated by a bi–level optimisation problem and manually adjusted supply parameters.

In literature can be found mobile phone data analysis based on Call Detail Record (CDR) exploration [5],[7]. Using CDR data we can recognise and analyse trajectories for each user separately. The trajectories are built on the base of the user activity in the mobile network when each event – originating and answering the call or sending/receiving xMS message – can be correlated with its location. Therefore, CDR data are easier to analyse and bring more information like additional information about the direction and speed of movement. Also, the level of analysing is different because a single mobile terminal is an observable marker of the users. However, these analyses are not legally permitted – according to EU law directives – without the consent of the user in most European Union countries. That moves the scope of the analysis, rather in the area of theoretical possibilities with no chance for practical use. The existing number of mobile devices tracked with the users' consents is insufficient for big studies. Another problem is the unknown bias of the consents on the collected data.

3 Data

3.1 Mobile subscriber's location statistics

Mobile subscriber's location statistics contains information on the number of terminals communicated with given cells of the Public Land Mobile Network

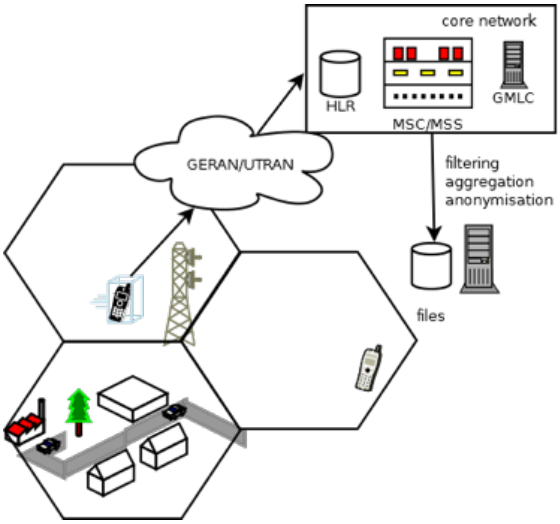


Fig. 1: Mobile subscriber's location statistics collection system

(PLMN) of Orange telecommunication provider in Warsaw municipal area. Based on that, the assumption is made, that at least given number of terminals, in given period of time was in the area covered by given cell. Cellular systems are separate for each telecommunication technology (2G/3G/LTE). The number of cells in given area is derivative of technology and capacity of BTS.

The coverage of single Cell can be different depending on area topology (up to 1 km in urban area, 20 km in rural) so the accuracy of this location is not as well as e.g. for GPS based solutions. When adding to it information that every cell has different area coverage in each technology (2G/3G/LTE) it brings complexity to the analysis. .

Another problem which requires a separate solution is the rule of network paging. A network paging contains a current location status and it is always done together with users actions like calls or SMS exchange and it brings the localisation note (cellid). But inactive terminals logged in the network but without any action of the user are periodically updated accordingly to network and terminal settings. Usually, it is done every 1-2 hours after the last event with location status.

The terminal location is detected on the basis of network events (13 different events are taken into account) that are triggered together with voice and xMS communication.

For statistics collection dedicated system on telecommunication provider presented in Figure 1 was used.

Statistics were collected in a quarter long, a hour long, and daily intervals.Statistics were also collected in a division into events generated by active subscribers and a periodic location update for terminals. Additionally, the overall statistics were created.

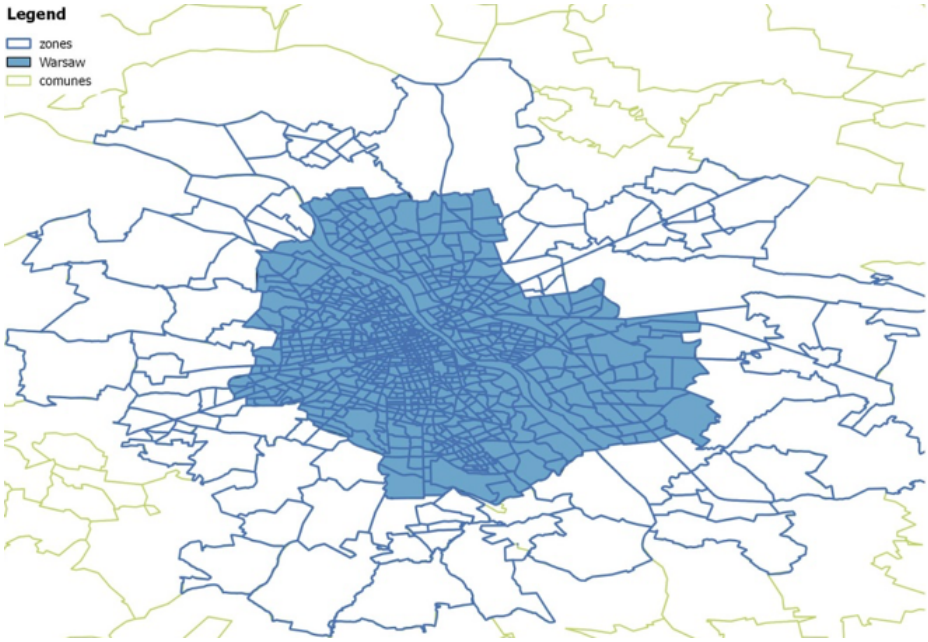


Fig. 2: Travel demand zones for Warsaw and neighbourhood communes [3]

In our work, we possessed samples of data from the urban area from selected cells located in Warsaw and also from selected cells from the Warsaw suburban area. The raw stream of data from the mobile cells in Warsaw was between 300 and 400 events per second. The average file size with the raw data from 24 hours for Warsaw area is about 100 MB. However, based on legal regulation in Polish Telecommunication Law, analysing this data in the raw state is not allowed without consent of the users. Therefore, the internal mechanism was implemented, and the data were collected from the early beginning as total sums in flat files containing the following columns: a latitude and longitude (in the WGS84 coordinate system) of the centre of the cell, the radius of the cell, the number of terminals in period for the defined cell, and the time stamp.

What is important, this method protects the privacy of the network users in the best possible manner. The data are collected only in the form of statistics which means: total sum of detections for a cell and there is no possibility of reversing this process. The processed data never originate from individual terminals. Statistics describe rather the load of the BTS than any characteristic of the individual users.

3.2 Travel demand zones

The travel demand zones for Warsaw are presented in Figure 2. The zones were created during the Warsaw Research Movement project in 2015 [3]. The project

developed the traffic model by a consortium composed of PBS Sp. o.o. (leader), Cracow University of Technology, and Warsaw University of Technology, on a request of the Capital City Warsaw.

Warsaw was divided into 896 interior areas, where 801 are municipal and 25 are from suburban ones. The centre of every region is considered as the start and end points of every travel area. During the determination of the regions, there were many factors taken into account, like the type and its function. Occurring communication barriers, like railway lines, were marked as the boundaries between the areas.

For each interior area, for the existing condition and forecasting periods, were assigned 186 variables. The variables include, among others, the number of inhabitants, workplaces, places in schools, the surfaces of objects selected category (residential, services, offices). These data may be used to estimate how many people start or finish trips in each region. Details are given in [3].

In our project, we used spatial characteristics of the zones – delivered as SHP files – to localise BTSs on their area. The geometry of the zones was transferred from the EPSG: 5300 coordinate system to the WGS84 (EPSG:4326) coordinate system used to describe BTSs localisations. Next, the geometry was validated to remove self–intersections from the zones.

4 Clusterisation

4.1 Self-Organizing Map

Collected data were clustered using the Self-Organizing Map (SOM) [6]. SOM is a fully connected single-layer linear network, where the output is organized in a two–dimensional matrix of nodes. The fundamental of the SOM is the soft competition between the nodes in the output layer. The winner shares the success with the neighbour nodes.

The SOM maps from the input data space $\mathbb{R}^n$ onto a two–dimensional matrix of modes called a map grid. A parametric reference vector $y_i \in \mathbb{R}^n$ is associated with i–th node on the map.

An input vector $x \in \mathbb{R}^n$ is compared with the y_i. The SOM can use any metric. In our model the winning node is calculated by

$$y = \arg\min_{y_i} \left(\sum_{i=1}^{n} (|x_i – y_i|)^2 \right) \tag{1}$$

and x is mapped onto y according to the parameter values y_i.

The distance is a commonly used modified Euclidean distance because we are not concerned with the actual numerical distance from the input. The computed value is some sort of uniform scale in order to compare each node to the input vector. This equation provides that, eliminating the need for a computationally expensive square root operation for every node in the network.

The map is being updated by an iterative presentation of learning data.

$$y_i(t+1) = y_i(y) + n_i(t, x(t) - y_i(t)). \tag{2}$$

The function n_i defines a neighbourhood and may vary in time according to the learning rate.

The SOM was implemented in R using the kohonen package [9]. During our tests the implementation was started with default parameters. Data were presented to the network 100 times. At the same time, the learning rate declined linearly from 0.05 to 0.01. The radius of the neighbourhood started with a value that covers 2/3 of all unit-to-unit distances. The initial weight given to the $X \subset \mathbb{R}^n$ map in the calculation of distances for updating $Y = \biguplus_i^n y_i$, and to the Y map for updating X was going linearly from 0.75 to 0.5 during the training. The initial values of y_i were selected randomly.

4.2 Preparation of input data

Each BTS_i is characterised by a radius BTS_i^r, a latitude BTS_i^{lat}, and a longitude BTS_i^{long}. The number of events depends on a discrete time and is defined as $BTS_i^e(t)$. The vector of $BTS_i^e(t)$ for $t \in [1, \ldots, 24]$ defines a daily distribution of events registered on BTS_i.

A BTS radius BTS_i^r can vary from 1 to 3933 meters. Raw data $BTS_i^e(t)$ – calculated per hour – are subscribers' logins amount and does not consider BTS range and area. During one day, 24 values of subscribers amount per hour gives subscriber's distribution. Distribution range is then arbitrarily large.

For analysis we used both the raw and normalised data. The normalisation considered two differences in area and time distribution. Four data sets were created: *abs_inf*, *per_inf*, *abs_nrm*, and *per_nrm*.

The sets with the *abs* prefix contained absolute values of events $BTS_i^e(t)$ not normalised according to BTS_i range. The sets with the *per* prefix contained values calculated per area

$$|BTS_i^e(t)| = \frac{BTS_i^e(t)}{\pi (BTS_i^r)^2} \tag{3}$$

The next step was a normalisation according to daily distribution. The sets with the *inf* suffix contained values of the events not normalised according to the daily distribution.

The sets with the *nrm* suffix contained values normalised to the range $[0, 1]$. According to the previous normalisation the sets consist of values

$$\begin{aligned} \overline{BTS_i^e(t)} &= \frac{BTS_i^e(t)}{\max_{t=1}^{24} \left(BTS_i^e(t)\right)} \\ \overline{|BTS_i^e(t)|} &= \frac{|BTS_i^e(t)|}{\max_{t=1}^{24} \left(|BTS_i^e(t)|\right)} \end{aligned} \tag{4}$$

For clusterisation based on BTSs the SOM uses vectors from R^{24} of values $BTS_i^e(t)$, $|BTS_i^e(t)|$,$\overline{BTS_i^e(t)}$, or $\overline{|BTS_i^e(t)|}$. Let us define $\boldsymbol{BTS_i}$ as a vector of daily events for BTS_i without deciding what kind of normalisation was used to calculate the vector. The vectors were used directly to train a SOM using BTSs.

Table 1: Clusters for raw data

	Count		Percentage	
Id	BTS	Nodes	BTS	Nodes
1	13	1	0.32	4.00
2	118	4	2.94	16.00
3	322	7	8.02	28.00
4	3528	12	87.83	48.00
5	36	1	0.9	4.00

To train a SOM using the zones a conversion of a set of vectors $\mathbb{BTS}$ connected with BTSs into vectors connected with the zones is required. The simplest method is to assign whole BTS and all its measurements to the zone which contains BTS's coordinates (geometry centre). Zones are distinct and so each BTS is contained by no more than one zone. It is the simplest and the least expensive assignment. the method has some disadvandages. The actual range of BTS and zone are ignored, it behaves as all subscribers were in the BTS centre.

Let us define a membership function of zone Z_i connected with a polygon Z_i^P. For a BTS with the coordinates $(BTS_j^{lat}, BTS_j^{long})$ the membership function is

$$\mu(i,j) = \begin{cases} 1 & \text{if } (BTS_j^{lat}, BTS_j^{long}) \in Z_i^P \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

Using (5), we can define the vector of daily events for the zone Z_i as

$$\boldsymbol{Z_i} = \sum_{j=1}^{|\mathbb{BTS}|} \mu(i,j)\boldsymbol{BTS_j}. \tag{6}$$

5 Tests

The aim of the test was to find BTSs with common events distribution during the day. A 5×5 SOM network model was created and trained with data from working days. The given data were from 16.05.2016 (16th May) to 26.06.2016 (26th June). From this range, all weekend were removed as well as 26.05.2016, which was a Holiday (Corpus Christi). As a testing day 17.05.2016 (Tuesday) was used. The data were clustered into 5 groups. The number of groups was a balance between a disparity of classes and a diversity of classes. The increase of the number of classes created small clusters. The reduction of the number of classes gave an extensive cluster that contained various types of the characteristics.

Tests on raw data For the first test the raw data were used. Data were collected in the *abs_inf* dataset. Table 1) contains information on the created clusters. A single cluster covered a majority of BTS and almost half of all SOM's nodes.

Table 2: Clusters for area normalised data

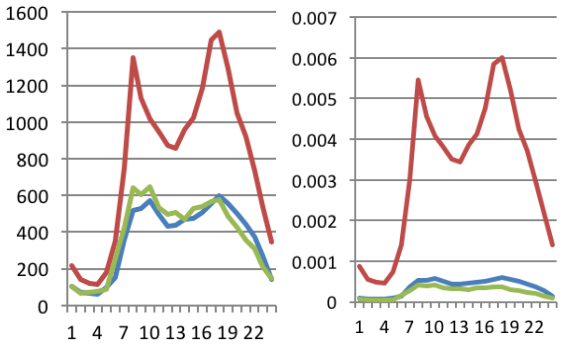
	Count		Percentage	
Id	BTS	Nodes	BTS	Nodes
1	1	1	0.02	4.00
2	4012	21	99.88	84.00
3	1	1	0.02	4.00
4	2	1	0.05	4.00
5	1	1	0.02	4.00

The results were not helpful to analyse the events distribution. Most of the BTS were in nodes with a flat similar daily characteristic.

Tests on data normalised by area In the second test, data from the *abs_inf* set were normalised per area using normalisation formula (4). As the result the *per_inf* dataset was created.

Data normalised per area did not improved the obtained results. Table 2 shows that nearly all BTSs were grouped in a single cluster. Increasing of the number of the iterations in the learning process did not change the results. Every but 5 BTSs are in the single cluster.

The area normalisation flats characteristics of the BTSs so much they cannot be distinguished any more. Figure 3 compares characteristics of three BTSs. Figure 3a shows the daily characteristics calculated on raw data when Figure 3b shows the characteristics for the same BTSs but on data normalised by area.



(a) Original subscribers amount

(b) Subscribers amount per area

Fig. 3: Three sample BTS characteristics before and after recalculate per area

Table 3: Clusters for daily normalised data

	Count		Percentage	
Id	BTS	Nodes	BTS	Nodes
1	938	5	23.35	20.00
2	480	4	11.95	16.00
3	429	4	10.68	16.00
4	1926	9	47.95	36.00
5	244	3	6.07	12.00

The comparison shows that two characteristics were flatten by the normalisation. The example conveys a reason of the reduction of data to a single cluster.

Tests on daily normalised data In the next tests, data from the *abs_inf* and *per_inf* were daily normalised using formulas (3). As the result the *abs_nrm* and *per_nrm*. datasets were created.

Both datasets gave the same result. The previous normalisation did not influence on the normalisation process where input data are scaled to $[0; 1]$ range.

Table 3 shows the obtained results. The results are more balanced than before. There still exists a domination of one cluster over the rest in sense of grouped BTSs but each of groups contain at least hundreds of BTSs. Additionally, in all cluster except one the number of used nodes is very balanced.

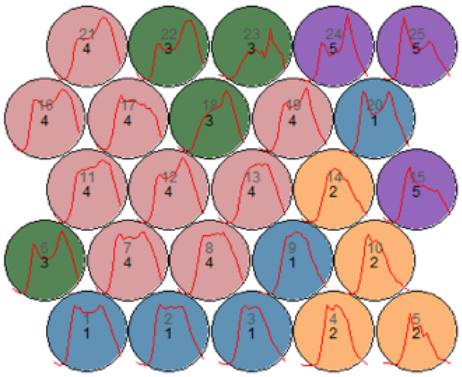


Fig. 4: Distribution of clusters among SOM nodes

Figure 4 shows codes that are values associated with SOM nodes. The codes show a variety of characteristics in the clusters. Thanks to the normalization and vertical alignment of characteristics the groups can be interpreted. The SOM identified forms with two clear peaks. Some of them have an equal height and some have one higher peak. There are also forms with the mild wide top.

X

5.1 Clusters distributions

For the future test have been done on the *per_nrm* dataset that contained daily normalised data. The SOM algorithm was started with default parameters but with three different seeds. The seed value defined random initialisation of the clusters. The aim of the test was to check a stability of the created clusters.

Figure 5 shows the obtained clusters. The obtained clusters differ but there are areas that stay in the same cluster for various initialisation parameters. The examples of the areas were identified manually and marked in red.

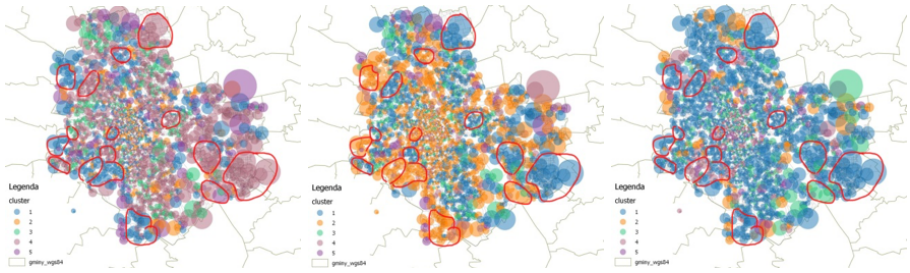


Fig. 5: BTS clusters

The last test was an aggregation of BTS data with the zones. The greedy method was used to group BTS data in the zones according to formula (6). A BTS was assigned to the zone that contained its centre. The BTS data were a daily normalised data.

Figure 6 shows three example of clustered zones. All SOM parameters were default only the seed was changing. We can show static areas as in the case of the clusters created directly on BTSs. The blue cluster contains zones placed on the east bank of the river in a south part and goes diagonally from south-west to north-east in the north part of the city.

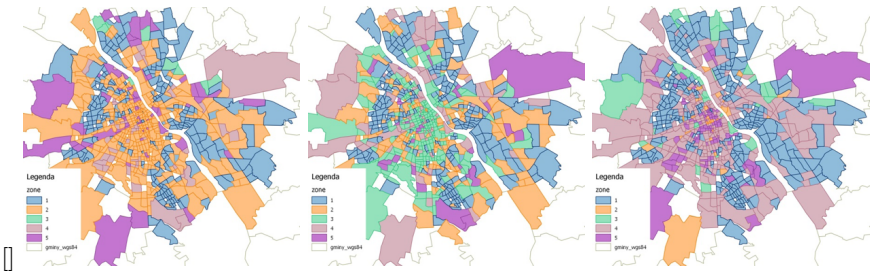


Fig. 6: Demanding zones clusters

The obtained results from both approaches cover each other at least partially. And in both cases, stable clusters can be defined. These results are most promising. Even the initial segmentation on all distributions already shows differentiation on the map where you can classify urban areas as a typical night or day. One can see large "day" clusters in the form of shopping centres and offices and "night" clusters of sleeping quarters.

For example office zones Służewiec Przemyslowy (Industrial Sluzewiec) in the south of Warsaw and Al Jerozolimskie in the west are visible. The major shopping centres in Warsaw such as Galeria Mokotów, Złote Tarasy and Arkadia can be also identified. The sleeping quarters Tarchomin in the north of Warsaw, Ursus on the west side and Wawer in the east are marked.

At this moment all zones of the same type – residential, services, offices – are detected manually using visualisations of several iterations of the clustering algorithm. An automation of the process is planned in the future.

6 Conclusions

In this work, we proposed how to convert the mobile subscriber's location statistics into a diversity of the transport demand zones according to a daily activity of the citizens.

Presented analysis shows that even aggregates in units of time can provide useful information after the use of appropriate methods of statistical processing. Topic requires further research, but even those so far are very promising and brings a lot of valuable information.

The obtained results are not as valuable as tracking the mobile device can be. But it is a good trade off to have less amount of information from big and reliable data rather than to have a big amount of information from a small not representative probe. Presented analysis shows that even aggregates in units of time can provide useful information after the use of appropriate methods of statistical processing. Topic requires further research, but even those so far are very promising and brings a lot of valuable information. The obtained diversity may be used to detect starting areas of travels for the given time of a day. Similarly, possible aims of travels can be detected.

The definition of the similar transport demand zones can be also useful in anomaly detection when an observed schema in the given zone does not fulfil the expected pattern for a type of the zone. However, in this case, the resolution of the measures should be higher than one hour.

Finally, the results may be concluded with strategy decisions of a public transport development. One of the applications of our work can be the demarcation of the paying parking zone. The longer observation of daily routines should tell us which zones should be included. Similarly, in work [10] toll on a regional expressway were estimated.

Acknowledgements

This research has been supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 688380 *VaVeL: Variety, Veracity, VaLue: Handling the Multiplicity of Urban Sensors.*

References

1. Horváth, B., Horváth, R.: Real network test of an iterative origin-destination matrix estimator in urban public transport. In: 2014 18th International Conference on System Theory, Control and Computing (ICSTCC). pp. 715–719 (Oct 2014)
2. Horváth, B., Horváth, R.: Estimation of sample size to forecast travel demand in urban public transport. In: 2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS). pp. 300–303 (June 2015)
3. Kostelecka, A., Szarata, A., Jacyna, M.: Warsaw' traffic measurement 2015. Tech. rep., PBS Sp. z o.o, Cracow University of Technology, Warsaw University of Technology (2015), http://transport.um.warszawa.pl/warszawskie-badanie-ruchu-2015/model-ruchu
4. Liu, H., Sun, J., Zhu, Z.: Study on processes reengineering of transportation planning. In: 2008 International Conference on Intelligent Computation Technology and Automation (ICICTA). vol. 2, pp. 494–497 (Oct 2008)
5. Lorenzo, G.D., Sbodio, M., Calabrese, F., Berlingerio, M., Pinelli, F., Nair, R.: Allaboard: Visual exploration of cellphone mobility data to optimise public transport. IEEE Transactions on Visualization and Computer Graphics 22(2), 1036–1050 (Feb 2016)
6. Murtagh, F., Hernández-Pajares, M.: The kohonen self-organizing map method: An assessment. Journal of Classification 12(2), 165–190 (1995), http://dx.doi.org/10.1007/BF03040854
7. Pinelli, F., Nair, R., Calabrese, F., Berlingerio, M., Lorenzo, G.D., Sbodio, M.L.: Data-driven transit network design from mobile phone trajectories. IEEE Transactions on Intelligent Transportation Systems 17(6), 1724–1733 (June 2016)
8. Seyedabrishami, S., Nazemi, M., Shafiei, M.: Off-line calibration of a macroscopic dynamic traffic assignment model: Iterative demand-supply parameters estimation. In: 17th International IEEE Conference on Intelligent Transportation Systems (ITSC). pp. 2035–2040 (Oct 2014)
9. Wehrens, R., Buydens, L.: Self- and super-organising maps in r: the kohonen package. J. Stat. Softw. 21(5) (2007), http://www.jstatsoft.org/v21/i05
10. Zhang, M.M., Wei, J.: Optimization for urban traffic assignment by congestion toll levied on regional expressway. In: 2008 Workshop on Power Electronics and Intelligent Transportation System. pp. 472–475 (Aug 2008)