Advances in Science and Technology Research Journal

Advances in Science and Technology Research Journal 2021, 15(3), 224–232
https://doi.org/10.12913/22998624/140540
ISSN 2299-8624, License CC-BY 4.0

Received: 2021.06.14
Accepted: 2021.07.20
Published: 2021.08.16

Travel Time Forecasting Based on Fuzzy Patterns

Adam Kiersztyn[1*], Agnieszka Gandzel [2], Leopold Koczan[2], Maciej Celiński[2]

[1] Department of Computer Science, Lublin University of Technology, ul. Nadbystrzycka 36B, 20-618 Lublin, Poland
[2] Faculty od Technology Fundamentals, Lublin University of Technology, ul. Nadbystrzycka 38, 20-618 Lublin, Poland
* Corresponding author's e-mail: a.kiersztyn@pollub.pl

ABSTRACT
Estimating travel time is one of the most important processes in logistics as well as in everyday life. In particular, when it comes to transportation services, efficient time management can be a competitive advantage, not to mention customer satisfaction, which can be easily translated into business success. Therefore, in this study we analyze various travel time estimation methods in combination with a well-known Fuzzy *C*-Means clustering algorithm. The proposed FCM-based solution has significant advantages, allowing for the determination of the optimal travel time. In an extensive numerical experiment, we present the application of the proposed method to estimate the time of a taxi trip around New York. Due to division of the city area into detailed areas and taking into account information about the travel time in the analysis, a model was obtained, that perfectly forecasts speed of taxi travel. In this study we consider various, competitive approaches to build such a model.

Keywords: travel time forecasting, fuzzy clustering, fuzzy patterns, transportation data.

INTRODUCTION

The problem of estimating travel time is a widely analyzed issue [1÷4]. There are a number of methods for estimating travel time depending on the available data. Recently, methods based on artificial intelligence techniques have gained popularity [5÷7] . The detection of patterns in the analyzed data is also considered [8]. Another approach consists in multi-dimensional clustering of available data [9]. Many studies consider the estimation of taxi travel time [10÷13], which depends to a large extent on the traffic volume in the [14, 15].

The main goal of this study is to combine the most commonly used approaches and propose to estimate the time of a taxi trip based on multi-level fuzzy clustering combined with the construction of a model using modern data analysis techniques. At the stage of building the model, apart from classical approaches, also ones based on artificial intelligence techniques were used, in particular Gradient Boosted Trees [16], Random Forest, or Tree ensembles [17]. The novelty of the proposed solution consists in the use of fuzzy clustering with the use of Fuzzy C-Means [18] of geographic data in combination with skillful determination of patterns. The essence of the method consists in skillful determination of patterns in data at various angles and using the average calculated for the elements of particular patterns to build the model.

The work is organized as follows. The second section contains a description of the proposed method. The third one presents an expanded, deepened example of the application of the innovative approaches discussed. The last part provides conclusions and future work directions.

THEORETICAL DESCRIPTION OF TRAVEL TIME ESTIMATION

The operation diagram of the proposed solution allowing to estimate the travel time based on the information available at the beginning of the journey is presented in Figure 1.

The point of departure in the proposed innovative solution is to build a model whose explanatory variables will be the values of the average time

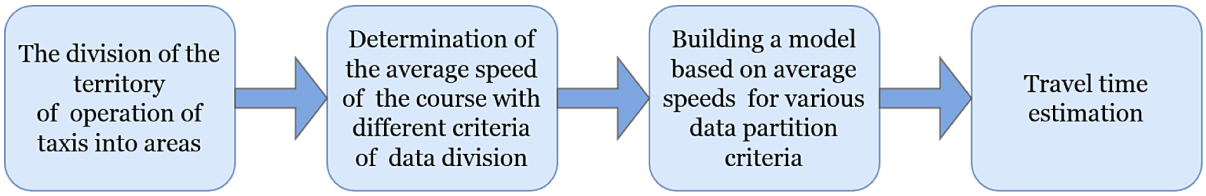

Fig. 1. Schema of the proposed method

of a taxi journey, assuming the occurrence of a certain condition. The natural conditions proposed for consideration should include the coordinates of the beginning and ending points of the journey. It is obvious that we are not able to consider all the possible combinations of start and end points. Therefore it is necessary to group and divide the city into certain areas. The natural solution seems to be to use fuzzy clustering with Fuzzy C-Means (FCM) [18]. As a result of the FCM algorithm, the degrees of membership in individual clusters are obtained. The value of the average travel speed is taken as a representative result for all elements belonging to a given cluster.

The division of a town area into clusters should be a compromise between the accuracy of the division (the larger number of clusters bringing a greater award) vs. speed (fewer clusters). The approach proposed in this study consists in designating 10 areas, which are then further detailed and each of the areas is divided into 5 subareas. As a result, a division into 50 clusters is obtained, but the use of cascade clustering allows for a significant reduction in computational complexity, and thus shortening the operation of the algorithm. Obviously, another number different than 10 is also possible to choose. However, the value of 10 was obtained from preliminary testing for the database under consideration.

It is proposed to use several independent divisions and determine the average travel speed for each of them. The use of the coordinates of the travel start points as the first division of the taxi operation area is suggested as a natural division. In the next step, it is advisable to make a division based on the coordinates of the end point of the journey and finally to simultaneously consider the coordinates of the start and end point of the journey as variables used in the FCM algorithm.

Basing the average travel speed prediction model on the travel start and end coordinates, despite its obvious advantages, may be an incomplete solution. It also seems to be justified to include information about the time of the travel start in the model. It is obvious that in large metropolises the speed of moving around the city largely depends on the traffic volume. The average speed in the morning or afternoon rush hours is much lower than the speed of driving in the night hours. In addition, the day of the week also undoubtedly affects the speed. For the set of explanatory data in the model, it is also advisable to use a conjunction of the previously considered conditions.

EMPIRICAL EXAMPLE

The operation of the proposed novel forecasting method based on fuzzy patterns is presented on the example of estimating the travel time in New York by taxi. The analysis used sample data [19] limited to 100,000 random elements in the training set and 2200 random elements in the test set. The adopted numbers of the training and test sets were determined as follows to increase the transparency of the results visualization. All calculations and visualizations of the results were made using the KNIME Analytics Platform (https://www.knime.com/). In accordance with the adopted methodology, all available points reflecting the location of the travel start point were divided into 10 clusters. The results of the division can be seen in the Figure 2.

Then, each of the areas was additionally divided into 5 sub-areas. Average velocity was determined for elements belonging to individual clusters. A similar division was made for the points determining the coordinates of the end of the journey. The third division into groups was made on the basis of the coordinates of both the start and the end of the journey together.

As a result of these operations, a total of 180 clusters were obtained, consisting of 30 main divisions and 150 detailed divisions. Based on the data from the training set, the average speed was calculated for the elements belonging to each cluster. These results differ significantly from each other, which can be seen in Figure 2. The points representing the highest average speed journeys

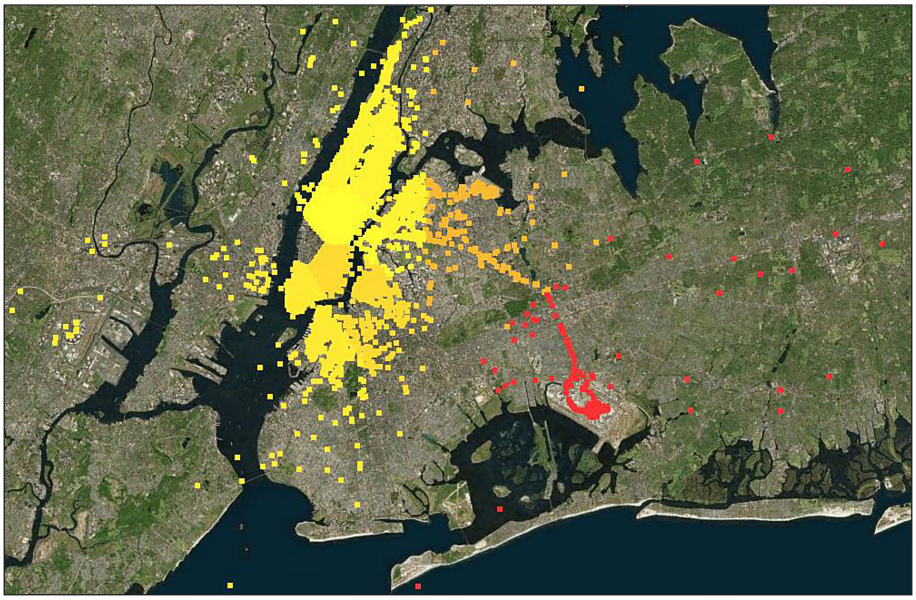

Fig. 2. Average speed for different sub-areas

are marked in red. It can be seen that the highest average speed is achieved for journeys departing from the outskirts of the city. Courses held in the city center are characterized by a lower average speed, which is marked in yellow in Figure 2.

Then, the data contained in the training set was divided according to the time of travel start. Instead of using the standard division by hour, FCM was also used, thanks to which the obtained division is more dependent on the analyzed data and more suited to their specificity. The visualization of the obtained division can be observed in Figure 3, where individual colors are responsible for belonging to the winning cluster. At first glance, the results of such a division may seem insignificant, but they play a huge role in building the model. The visualization presented in Figure 3 shows the lack of dependence between the time and place of the beginning of the journey.

For the points belonging to the clusters obtained in this way, the average travel speed was also determined. The values of this mean are presented in Figure 4. As predicted, journeys with the highest average speed take place during the night. The starting point is irrelevant in this context of data analysis.

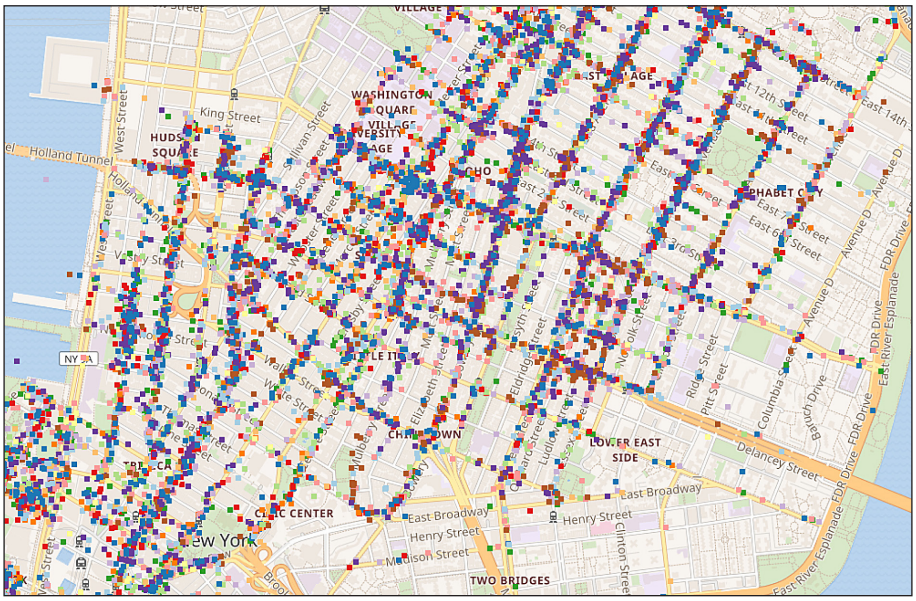

Fig. 3. Time of day clusters

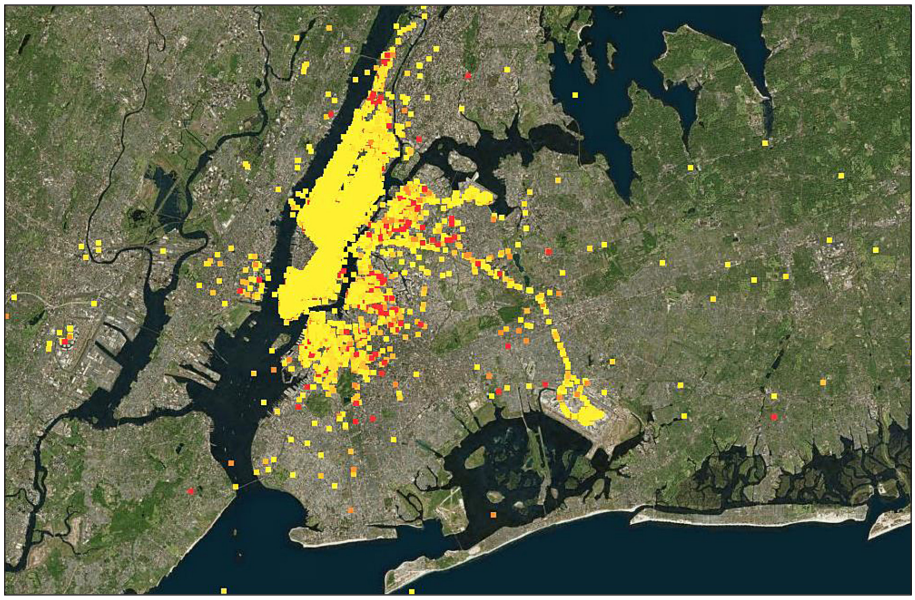

Fig. 4. Average speed for the split, taking into account the time of day

The next division of the elements of the training set was made according to the day of the week on which the trip took place. It turns out that the number of elements belonging to each group varies, which may suggest that there are different numbers of taxi journeys on different days. The distribution of the number of taxi journeys by day of the week can be seen in Figure 5.

Evidently, the percentage of taxi trips taken on Sunday is the smallest, which is probably due to much less traffic on that day. Less city traffic on Sunday allows for a smoother ride, which has at least two consequences. First, the higher average speed of journeys on that day. Second, and more important from a modeling point of view, greater traffic flow reduces the chances of traffic disruptions and at the same time increases the precision of the estimates.

Another axis of data division is to distinguish both the day and time of the beginning of the taxi course. Such a division of data allows to capture complex relationships reflecting the variability of weekly and daily cycles in the intensity of urban traffic. The penultimate division category takes into account the time of day as well as the location of the beginning of the taxi course. The last category of division takes into account both the coordinates of the start and end of the journey and the time of the day when the journey takes place. The last two categories make it possible to include in the estimates both the daily volume of city traffic and the traffic volume in specific city locations. The Figure 6 shows the average speeds

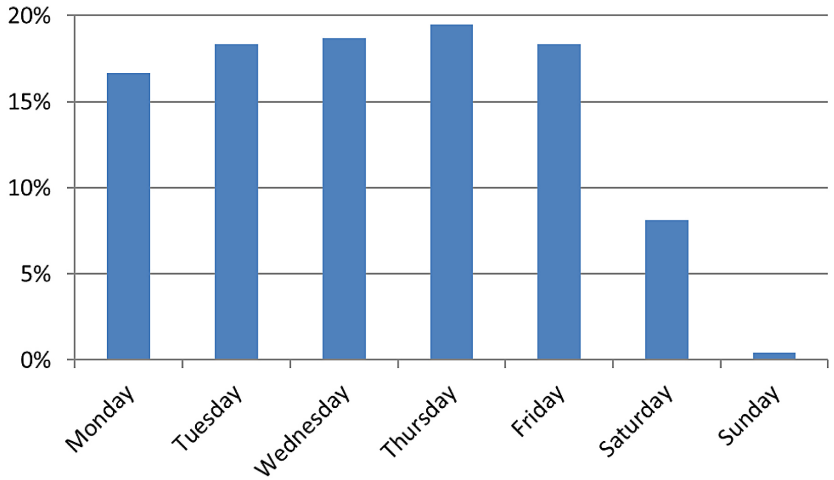

Fig. 5. Percentage of the number of taxi journeys by day

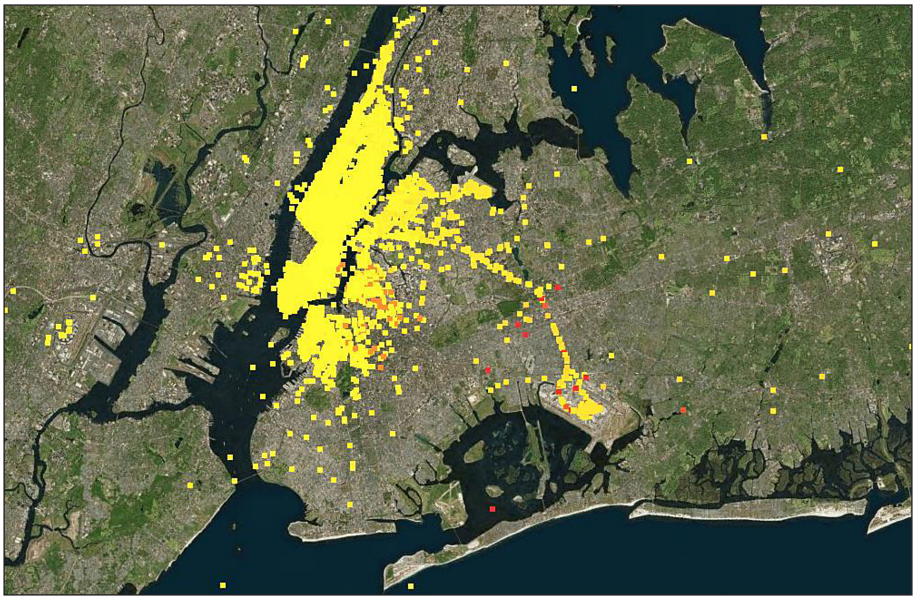

Fig. 6. Average speed depending on the time of day and location of the beginning of the course

for individual groups taking into account the time of day and the location of the beginning of the taxi course. The color of the point corresponds to the average speed value in a given class.

After in-depth analysis of the data presented in Figure 6, it can be seen that the highest average speed of over 105 mph is recorded for journeys departing from JFK airport during night hours. On the other hand, the journeys with the lowest average speed starts downtown during the afternoon rush hour. In this case the average speed does not exceed 9 mph.

Based on the data divisions of the training set and the average speeds determined for them with which the taxis traveled, a model forecasting the average speed of a new journey was built. Several competing approaches have been considered for comparison. A linear model, a polynomial model, a Gradient Boosted Trees model [16], a Random Forest model and a tree ensemble [17] model were determined.

Information from the test set was used to verify the effectiveness of individual models. Based on the patterns describing individual clusters in FCM, the affiliation of each course from the training set to individual clusters was determined in each of the analyzed approaches. In this way, a series of values was obtained corresponding to the average speeds determined on the training set. These values were substituted into appropriate models and an estimate of the average speed was obtained. Basic statistics for the relative error of estimation are presented in Table 1.

When analyzing the results contained in Table 1, it should be noted that the best model is obtained after applying the Gradient Boosted Tree (GBT). The average value of the relative error for the GBT-based model is only 9\%, which is significantly lower than the other models. After a more in-depth analysis of the remaining statistics, it should be noted that first quartile of the estimates for the training set is less than 2\%. The value of the third quartile is 11\%, which confirms the high efficiency of this model. These values indicate a very good estimation of the average travel speed using the model based on average speeds determined for different data divisions for the training set. It should also be noted here that simple models, such as a linear model or a polynomial model, do not give satisfactory results.

Table 1. Basic statistics for the relative error

Statistics\Model	Linear	Polinomial	GBT	Random Forest	Tree Ensemble
Average	0.42	0.35	0.09	0.16	0.16
Quartile 1	0.12	0.13	0.02	0.05	0.05
Median	0.25	0.22	0.06	0.10	0.10
Quartile 3	0.48	0.41	0.11	0.17	0.16
Standard deviation	0.70	0.46	0.15	0.25	0.23

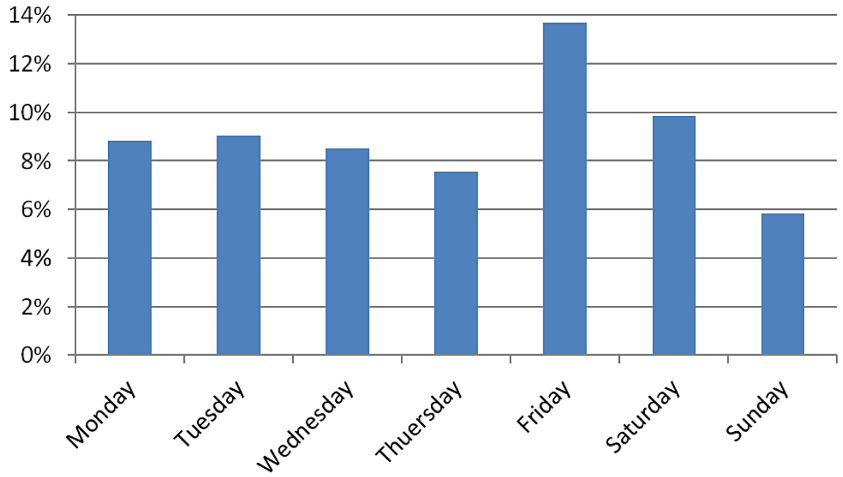

Fig. 7. Average value of the GBT model relative error for each day

It is also advisable to investigate how the relative error is distributed for different data partition criteria. For example, Figure 7 shows the average relative error of the GBT model broken down into consecutive days of the week.

It can be observed, that the day of the week may have a bearing on the estimation error. The slightest error is achieved for trips that take place on a Sunday. This is a relatively natural observation, because on Sundays there is usually less traffic in cities. Due to the reduced traffic, it is easier to estimate the unknown average travel speed, because fewer factors can distort correct forecasting.

On the other hand, the value of the average relative error after taking into account the division into the main areas of travel commencement is presented in Figure 8.

Figure 8 shows the average values of the error estimates for the items belonging to the super-cluster of cluster number 2 and number 9, as in the test set of such elements are not found. The geographical location of individual areas is shown in Figure 9. It can be seen that some areas do not appear in the graph either. This is due to the fact that the elements located off the coast of Africa and Europe were allocated to two clusters. The area of data presentation was limited to the New York area and the clusters were not mapped in Figure 9.

The elements belonging to a given cluster according to the proposed method are then reclassified using the FCM method. The results of an exemplary detailed division can be seen in Figure 10.

It can be seen that the additional division of the travel start area is of particular importance in the city center.

The comparison of the actual speed value with the best estimation methods is shown in Fig. 11, where the travel time is marked on the X-axis. It is easy to notice that one of the observed values differs significantly from the others. However,

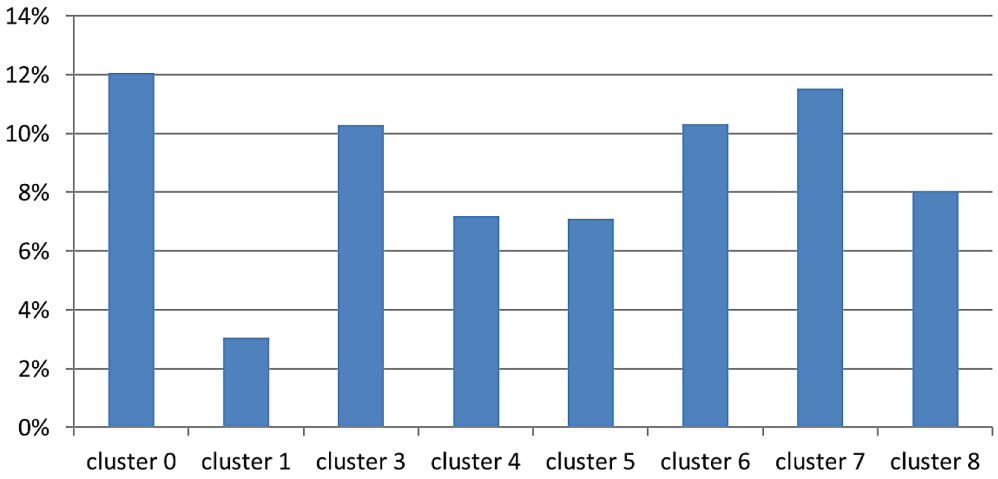

Fig. 8. Average estimation error due to membership of the cluster designated by the beginning of the journey

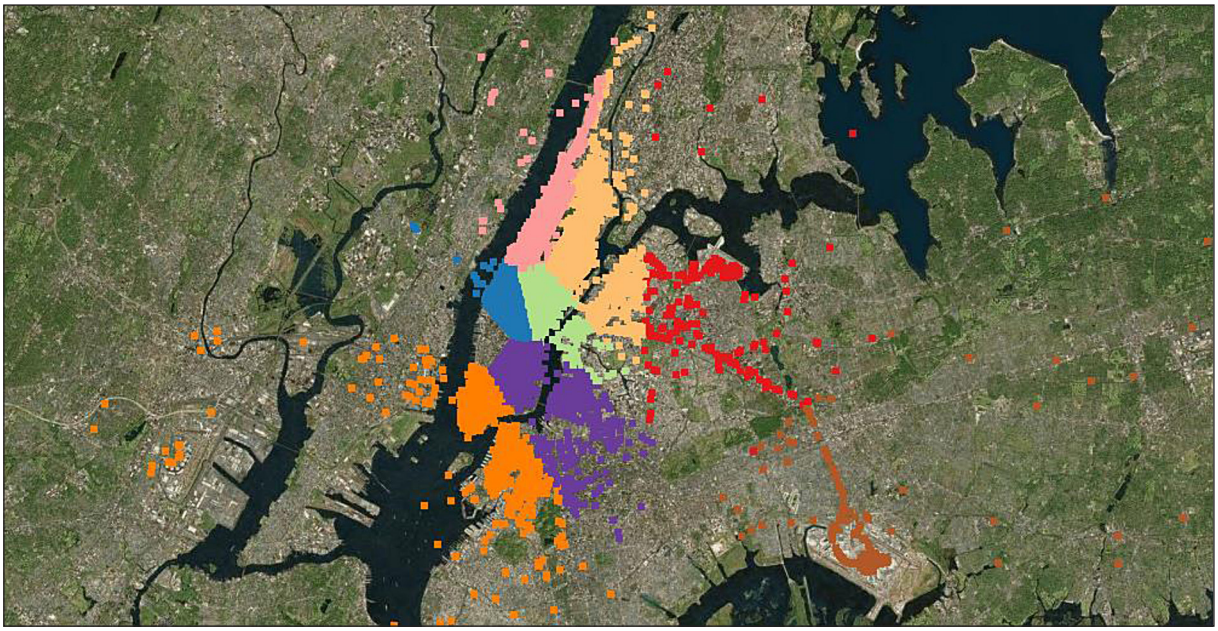

Fig. 9. The main clusters for division according to the starting point

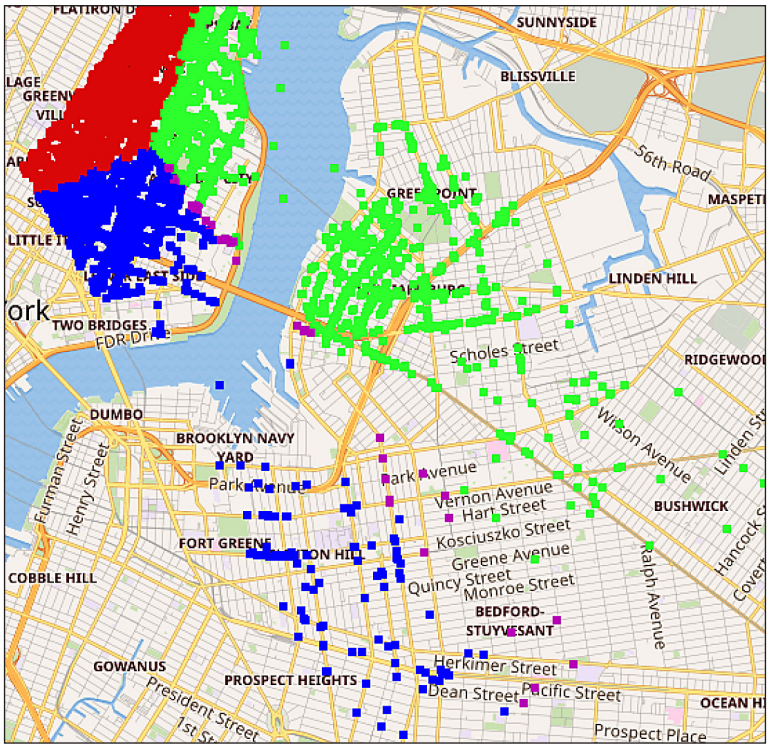

Fig. 10. Division into sub-areas of elements belonging to cluster number 3

most models were able to estimate this peak well in the data well.

On the basis of the estimation of the average speed with which the journey will take place, it is also possible to easily estimate the travel time. It is enough to use any tool determining the distance between two points and on this basis calculate the travel time.

CONCLUSIONS AND FUTURE WORK

The method of estimating the time of a taxi trip, based on historical data, proposed in the study gives the possibility of planning a trip and allows better time management. The results of numerical experiments confirm the potential of the proposed approach. Based on a relatively small

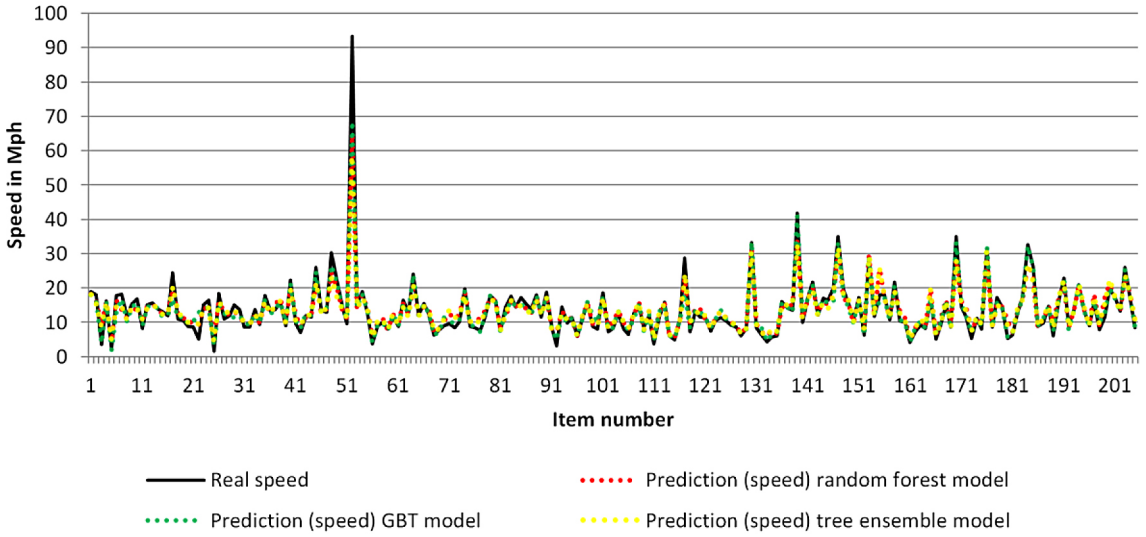

Fig. 11. Comparison of real speed with the best estimation methods

training set, it was possible to build a model in which the average relative error is at the level of 9%. High efficiency in the estimation of the travel speed clearly translates into the effectiveness of the estimation of travel time.

There are plans for further research on the development of the proposed concept involving the use of a larger number of elements in the training set. In addition, it is planned to introduce more conditions to be included in the analysis and to introduce more clusters in FCM.

Acknowledgments

Funded by the National Science Centre, Poland under CHIST-ERA programme (Grant no. 2018/28/Z/ST6/00563). The work was co-financed by the Lublin University of Technology Scientific Fund:*FD-ITIT-KIER*.

REFERENCES

1. Guo G., Zhang T. A residual spatio-temporal architecture for travel demand forecasting. Transp. Res. Part C Emerg. Technol. 2020;115:102-639.
2. Li Y., Lu J., Zhang L. Zhao Y. Taxi booking mobile app order demand prediction based on short-term traffic forecasting. Transp. Res. Rec. 2017;2634(1):57–68.
3. Park D., Rilett L. R., Han G. Spectral basis neural networks for real-time travel time forecasting. J. Transp. Eng. 1999;125(6):515–523.
4. Yang M., Liu Y., You Z. The reliability of travel time forecasting. IEEE Trans. Intell. Transp. Syst.11. 2009;1:162–171.
5. Liao S., Zhou L., Di X., Yuan B., Xiong J. Large-scale short-term urban taxi demand forecasting using deep learning, of23rd Asia and South Pacific Design Automation Conference (ASP-DAC). IEEE. 2018;428–433.
6. Luo H., Cai J., Zhang K., Xie R., Zheng L. A multi-task deep learning model for short-term taxi demand forecasting considering spatiotemporal dependences. J. Traffic Transp. Eng. (Engl. Ed.). 2020.
7. Rodrigues F., Markou I., Pereira F.C. Combining time-series and textual data for taxi demand prediction in event areas: A deep learning approach. Inf. Fusion. 2019;49:120–129.
8. Zhang Z., Wang Y., Chen P., He Z., Yu G. Probe data-driven travel time forecasting for urban expressways by matching similar spatiotemporal traffic patterns. Transp. Res. Part C Emerg. Technol. 2017;85:476–493.
9. Davis N., Raina G., Jagannathan K. A multi-level clustering approach for forecasting taxi travel demand, IEEE of 19th International Conference on Intelligent Transportation Systems (ITSC). IEEE. 2016;223–228.
10. Chintakayala P.K. & Maitra B. Modeling generalized cost of travel and its application for improvement of taxies in Kolkata, Journal of Urban Planning and Development. 2010;136(1):42–49.
11. Faghih-Imani A., Anowar S., Miller E.J., Eluru N. Hail a cab or ride a bike? A travel time comparison of taxi and bicycle-sharing systems in New York City, Transportation Research Part A: Policy and Practice. 2017;101(C):11–21.
12. Markou I., Rodrigues F., Pereira F.C. Real-time taxi demand prediction using data from the web, 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE. 2018;1664–1671.
13. Tang J., Zhang S., Chen X., Liu F., Zou Y. Taxi trips distribution modeling based on Entropy-Maximiz-

ing theory: A case study in Harbin city – China, Physica A. 2018;493(C):430–443.

14. Nie Q., Xia J., Qian Z., An C. Cui Q. Use of multisensor data in reliable short-term travel time forecasting for urban roads: Dempster–Shafer approach, Transp. Res. Rec. 2015;2526(1):61–69.
15. Castro P.S., Zhang D., Li S. Urban traffic modelling and prediction using large scale taxi GPS traces, in: J. Kay, P. Lukowicz, H. Tokuda, P. Olivier, and Kr¨uger A., (eds.) Springer, Pervasive Computing. 2012;7319:57–72.
16. Friedman J.H. Stochastic gradient boosting, Comput. Stat. Data Anal. 2002;38(4):367–378.
17. Kocev D., Vens C., Struyf J. D˘zeroski S. Tree ensembles for predicting structured outputs, Pattern Recognit. 2013;46(3):817–833.
18. Bezdek J.C., Ehrlich R., Full W. FCM: The fuzzy c-means clustering algorithm, Comput. Geosci. 1984;10(2-3):191–203.
19. Donovan B., Work D. New York City Taxi Trip Data (2010-2013); 2014.