BALÁZS HORVÁTH
dr, Széchenyi István University,
Department of Transport Hungary,
email:balazs.horvath@sze.hu
VIKTOR NAGY
Széchenyi István University,
Department of Transport Hungary

Zone estimation with cluster analysis of public transport stops[1]

Streszczenie: Obecnie nieustannie powstają różne zbiory danych, generowane przez różne urządzenia i systemy. Transport publiczny nie jest w tym zakresie wyjątkiem. Nowoczesne systemy monitorowania oparte na GPS i bilety elektroniczne wytwarzają duże ilości danych, i moglibyśmy je wykorzystać dla poprawy poziomu usług. Z jednej strony dane te są przechowywane, a dostawcy usług nie mają do czynienia z zawartością informacji, ale z drugiej strony, być może są one po prostu usuwane, aby ograniczyć obciążanie cyfrowej przestrzeni. Dane te mogą być przetwarzane dzięki nowoczesnym urządzeniom i metodom, i możemy je wykorzystać do uzyskania informacji. Dzięki rozprzestrzenianiu eksploracji danych, narzędzia te pojawiają się nie tylko w badaniach marketingowych, ale w większości różnych działań badawczych, i reklamują one nową rewolucję naukową. Chociaż znaczenie tych źródeł danych jest zasadnicze, nie jest to rozpowszechnione w planowaniu transportu, a jedynie w pewnych określonych obszarach [1], tak jak pisze Csiszár i in. W artykule przedstawiono możliwości zastosowania materiałów nieprzetworzonych, biorąc jako podstawę informacje o pasażerach podróżujących transportem publicznym. Stworzono metodę trzech kroków, która może być przydatna do automatycznego kształtowania strefy lub do nadzorowania granic stref utworzonych konwencjonalnie. Może też przydać się przy kontroli gruntów użytkowych. Procedura ta jest skuteczna w tworzeniu łańcuchów podróży z danych z kart inteligentnych oraz w tworzeniu macierzy żródło-cel na podstawie danych zameldowania. W artykule pokazano, w jaki sposób możliwa jest dystrybucja stref, za pomocą różnych metod pomiaru odległości i procedur gromadzenia, i zaprezentowano tego efekty na przykładzie wybranego miasta.

Słowa kluczowe: transport publiczny, big data (duże zbiory danych), szeregi czasowe, podobieństwa macierzy, grupowanie

Introduction

The cognition of the traveling behaviour is essential of the maintenance and correction of the service level of the public transport. The demands are changing continuously. We can talk about daily, weekly and seasonal fluctuation. The traffic demands evolve because of the different functions of areas, so if we know the character of the area, we can deduce the demands too. This could be true vice versa. If we knew the position and the time of demands, we would know the land-use character of the given part of the city. This article is to examine and confirm this thesis. In other words, how we can create traffic zones with modern tools and in the knowledge of passengers' check-in check out data.

The smart card systems are storing the number of boarding passengers, and in some cases also the alighting values. In our case the passengers' data was known from a passengers counting, which was executing in Győr. The database contains all of the stop points in the city, the boarding and alighting information and we can also extract the time of these. The main goal of the research is to explore the area's behaviour with data mining techniques. In the first step, the method assigns boarding data per hour to every single stop points. From the passengers' boarding and alighting information in a stop point, we created time series, which are showing the behaviour type of the given stop points, presented on graphic curves. Based on these time series we are able to deduce to the characteristics of the stop point's environment, since the different land usage yields dissimilar stop usage, with well-defined peak hours.

In the second step, the method compares the stop points to each other and adds a dissimilarity value based on the boarding data. With the distance measurement of time series it is possible to define, that how similar are two selected stop points and their environment to each other. For such kind of distance measurement are more methods are known.

In the next step, with the help of different clustering processes and the usage of "R" data miner software, these distance data can be turned into groups, and we can observe these groups of stop points. These stop point clusters are defining separated zones, what is a basic step in transport modelling, and the production of them were coming true with manual methods usually, until now.

Data description

The given data counted in Győr in 2012. The population of the city is 129.372 and its territory is 174,56 km^2. The city is served by buses with 451 stop points. The database contains the lines, the directions, the vehicle types and the capacity, the schedule based and the real departure and arrival times. It contains furthermore the name and the code of the involved stop points, the distance of them, the time from start to reach the given point, and also the boarding and alighting numbers of passengers in the given stop points. With the changing of structure and filtering the data we create the following datatypes: stop point code, boarding number, time. In the research we work from these data. We take into consideration only the boarding numbers and we treat the alighting numbers as irrelevant.

We did not need a precision in minutes, and to treat the different lines in different rows neither. We assign the boarding information to a stop points, we summarize the boarding numbers per hours, so we get one row (a time series) for each stop points.

[1] ©Urban and Regional Transport, 2016.

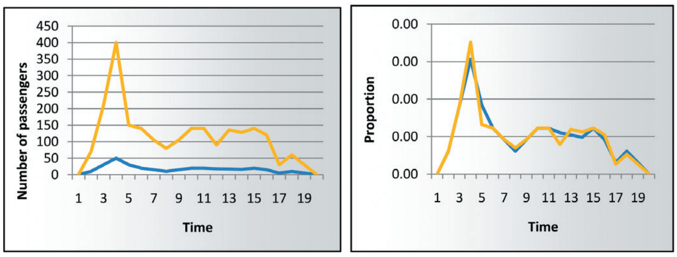

Fig. 1. The exact numbers and the percent values of boarding passengers

Thus the cells of the received table contain the sum of one hour boarding passengers of a given stop points. The total value of passengers using one stop point is obviously different. If we represented them in a diagram, we can see that similar kind of points (typically working places, housing zones …) maybe seem different, and we cannot compare them. For the proper comparison we did not use the exact number, but the proportion of the given hour in the stop point, so the values turn into comparable, as it can be seen in Figure 1.

Time series comparison

As we can see int he previous chapter, all the stop points belong to a time series, which allows us to compare the records. There are multiple methods for time series distance measurement. In this chapter we will introduce some of them and finally choose one and present its results. Henceforward the distance expression means deviation in nature and not physical distance.

Manhattan distance

The Manhattan (or in other words city block) distance is used in the case of normal street networks originally. The Manhattan method counts the walking distance instead of the straight line distance, but it is also used to time series comparison. The advantages of this method are the simplicity and the fast computation. The counting method (where „a" and „b" are the chosen stop points and „i" means the hours) is the following [2]:

$$d_{MAN} = \sum_{i=4}^{22} |a_i - b_i| \quad (1)$$

Euclidean distance

This method is known from geometry and it is an often used mode for time series distance measurement too. This is the most known method, which based an the Pythagorean theorem. It can be calculated as follows [3]:

$$d_{MAN} = \sum_{i=4}^{22} |a_i - b_i| \quad (2)$$

The advantage of the method is the low computation time, the disadvantage is that we compare the pairs only, so a little shift in the time series causes difference, but in real life these series mean same kind of stop points.

DISSIM

The name DISSIM came from the word dissimilarity. The method examined the difference between two time series as follows [4]

$$d_{DIS} = \int_{i=4}^{i=21} |a[i] - b[i]| di \quad (3)$$

It trains the absolute value of the difference of two coherent time values. After that, the equitation counts the definite integral of the curve. The advantage of the method is that it counts not only the difference between two point pair, but the narrow environment too. Also the counting process lasts not so much longer comparing with the previous mentioned methods.

Dynamic Time Warping

The Dynamic Time Warping (DTW) method compares not only the values of a given hour, but the environmental values of the hour too. The method „warps the time" [5]. This method is favourable in our case. It is common in public transport, that the demand appear in an hour late, in a long transport line, or maybe the demand outlast. However two stops point and territory are similar to each other.

The advantage of the DTW method is that we can put some parameters to the equitation, but the appropriate calibration is difficult. The computation time is also longer, than in the case of other mentioned methods. The DTW compares all the hours of all the stop points to each other. Here it means, that a 19x19 (the vehicles make journeys in 19 hour a day) matrix belongs to all cells of a 451x451 symmetric matrix.

The result of distance measuring

The methods specify the distance between two stop points, and because of the different measuring manner, the result values are in variant scales. For the better comparability and transparency, we define the results in a 0 to 100 scale. For this, we used the following equitation:

$$d = 100 - 100 * \left(\frac{d_{max} - d_{ij}}{d_{max}}\right) \quad (4)$$

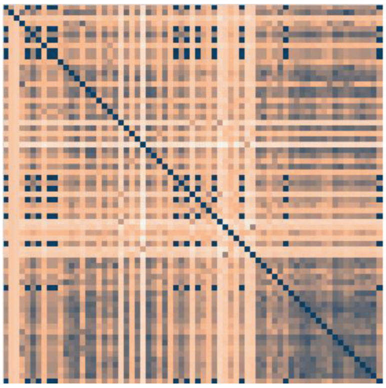

Fig. 2. Similarity matrix sample

We are figuring the result in a matrix form. If the crossings were bluer, the stop points are more similar to each other, while the orange colour means total difference. In this paper, we are submitting the results of Euclidean distance measurement, which found as the most appropriate method from the selected ones, because of its speed and

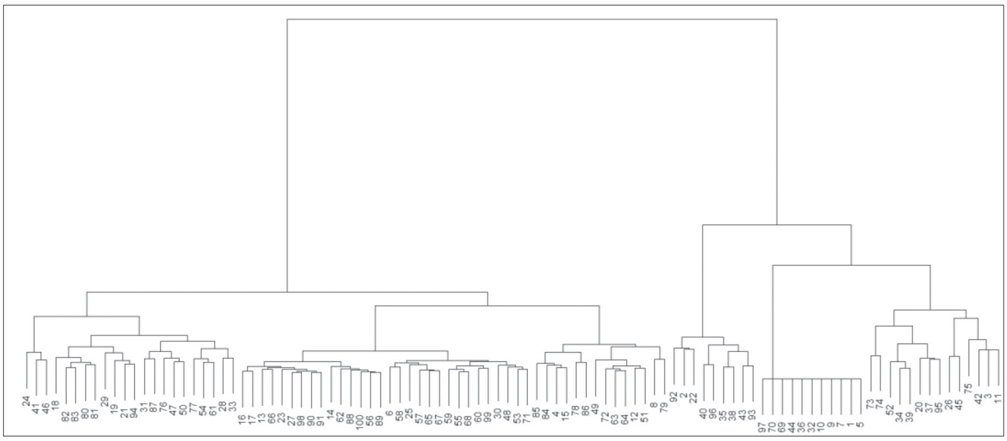

Fig. 3.
The sample dendrogram of Ward method

usability. It takes more time to use the Dynamic Time Warping and without the correct calibration it is not better, than the others. Besides, based on the similarity matrix, the Euclidean method was the only one which shows appropriate diversity and not only extremity, so we thought that this method is properly subtle and not so rough. The matrix is mirrorable and symmetric.

Clustering

The clustering methods are the first and most popular data mining techniques. It is used in different scientific areas, such as medicine, astronomy, social sciences. The method creates groups based on the different attributes of elements. Opposite to classification, this method does not need predefined classes so clustering techniques also can be called as unsupervised learning [6]. The method is able to arrange the elements into groups without samples and the knowledge of group numbers and attributes. We examined the data with three hierarchical clustering methods, which were the single linkage clustering, complete linkage clustering respectively the Ward method [7]. The final result can be represent in a dendrogram in the case of hierarchical methods, which shows that how link to each other the single elements and groups. It also shows the distances between each other. The advantage of this is that we do not need to find the proper group number previously, just cut the tree at the selected value.

We used the R data miner software for the examination [8]. The inputs were the dissimilarity values, which was counted in excel. The results of the simple linkage and the complete linkage methods did not show proper form, so we used the Ward method finally, which seemed promising for further investigation. Figure 3 shows the results of a sample of 100 stop points for the better perspicuity. It can be seen, that the dendrogram split into well separated groups.

We defined the ideal number of clusters on the Ward dendrogram. We circled the separated clusters and counted them. We found 14 separated groups. The program shows that which stop points or clusters are grouped together in which step and how far were these groups to each other. If we illustrated the distances in every steps, we get the following diagram (Figure 4). The diagram is strongly growing in the environment of 14 clusters, because more and more different kind of stop points group together from here.

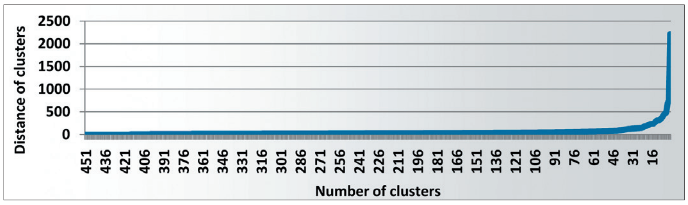

Fig. 4. The distances of clusters

Results

We illustrate the results of the clustering method with the help of PTV VISUM transport planning system. Figure 5 represents how the 14 different kinds of stop points locate in the city. The different stop points locate not exactly how we expect, but the map shows similarity to our previous imagination.

Based on the above we had not rejected the original idea, but we took into account not only the time series distances, but the physical distances too. The traffic model contains the exact coordinates of the stop points, so we counted the Euclidean distances for each pair of stop points. After that,

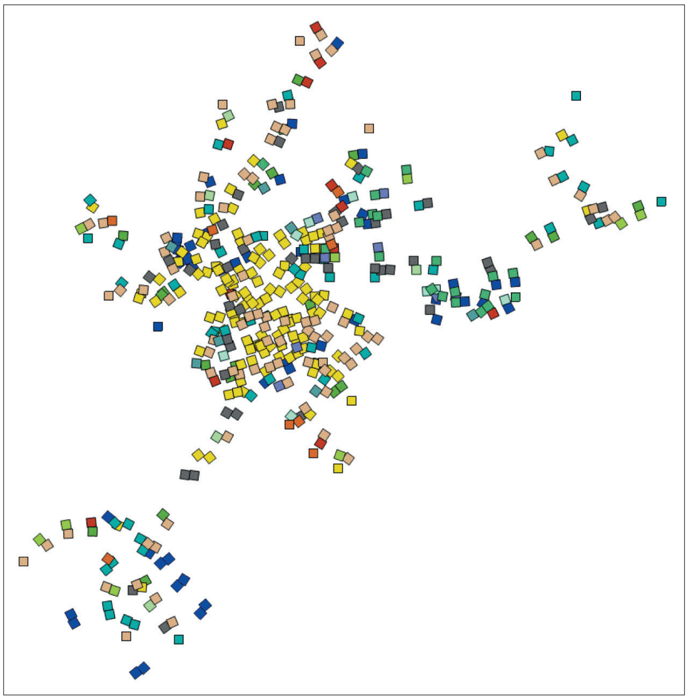

Fig. 5. The positions of different stop points

we created factors from the distances, and corrigated the original time series distances with them. The stop points which are close to each other and shows similarity grouped together, while the remote and dissimilar ones have not.

We were thinking about what would be the results, if we took into account only the physical distances and made the clusters that way. The results show, that however there are essential differences between the two methods, the effects of physical distances are too strong. These effects distort the attribute parameters and the results, so we rejected this method in this form.

As it can be seen on the figure the coverage of different colours and planning zones (drawn on the basis of local knowledge) are quite good, although the colouring was made just with mathematical methods without any local knowledge as shown above.

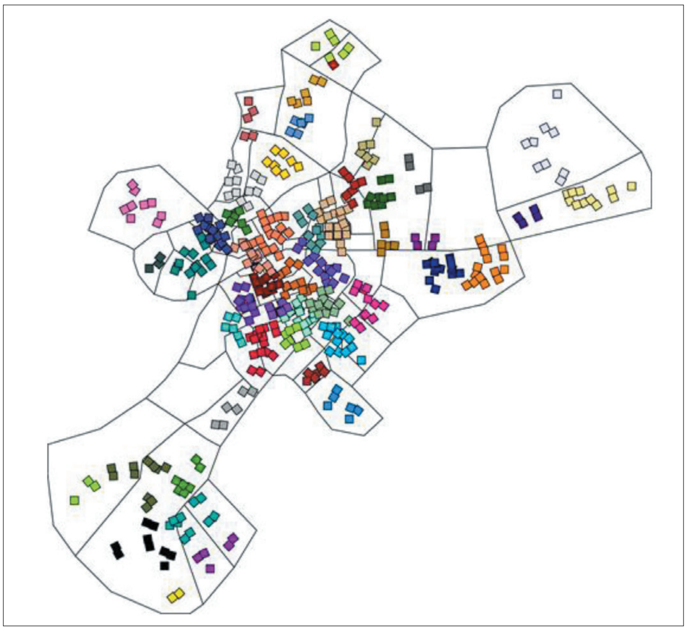

Fig. 6. The clusters of stop points, corrugated with physical distances

Conclusion

Continually developing world of today, the amounts of different kind of data are growing in every aspect of life. In this paper we showed an opportunity to examine and use this data, which could be generated in public transport every day. We clustered the stop points into groups, based on one day's data of passengers boarding. In a city, where the smart cards are in daily usage, our investigation would show a more accurate picture, thanks to the bigger data set. In the examination we used the Euclidean distance method for time series similarity measuring and the Ward method for clustering the stop points. The result did not show the perfect shape of the zones, but definitely gives us a reason for further investigation. For the accurate picture it is necessary to make the correct calibration so in the further work we want to deal with these issues. We also want to take into account the alighting number of passengers and parallel to this to try other calculating methods.

References

1. Csiszár Cs., Földes D., *Analysis and Modelling Methods of Urban Integrated Information System of Transportation*. Smart Cities Symposium, 24-25 June 2015. Prague, Czech Republic, pp. 1-10 DOI:10.1109/SCSP.2015.7181574, ISBN: 978-1-4673-6727-1 http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnum ber =7181574
2. http://psycho.unideb.hu/statisztikav1.0/pages/p_5_10.xml (2016.01.06)
3. Pang-Ning Tan, Michael Steinbach, Vipin Kumar, *Bevezetés az adatbányászatba*, Panem Kft., Magyarország, 2011.
4. Buza K., *Fusion Methods for Time-Series Classification*, University of Hildesheim, 2011.
5. Bodon F., Buza K., *Adatbányászat*, 2014.
6. Abonyi J. (szerk), *Adatbányászat a hatékonyság eszköze*, ComputerBooks, Budapest, 2006.
7. Obádovics C., *Klaszteranalízi*s, Eszterházy Károly Főiskola, Eger, 2009.
8. *A language and environment for statistical computing*, Foundation for Statistical Computing, Bécs, 2008.

Dokończenie tekstu ze strony 8

And the third one – the value of the array data forms the big number of degrees of freedom that is why the most probable states of the O-D matrix can be defined only in some confidence interval but not by point representation.

References

1. Ortuzar J. de D., Willumsen L. G. , *Modelling transport. 4rd ed.*, John Wiley & Sons Ltd, Santiago, 2011.
2. Winston C., Small K. A., *The demand for transportation: Models and applications*, University of California, California, 1998.
3. Fratar T. J., *Vehicular Trip Distribution by Successive Approximation*, Traffic Quarterly", 1954, No. 8.
4. Quarmby D. A. , *Choice of travel mode for the journey to work*, "Journal of Transport Economics and Policy", 1967, Vol. 1, No. 3.
5. Barbier M., Merlin P., *Le futur réseau de transport en région de Paris*, Cahiers de 1'I. A.U.R.P., Paris, 1966.
6. Jones I. S., *Gravity models and Generated Traffic*, "Journal of Transport Economics and Policy", 1970, Vol. 6, No. 2.
7. Preston J., *Demand forecasting for new local rail stations and services*, "Journal of Transport Economics and Policy", 1991, Vol. 25, No. 2.
8. Shuman H., Harding J., *Prejudice and the Norm of Rationality*, *"Sociometry"*, 1963, Vol. 27.
9. Brown L. A., Moore, E. G., *The intra-urban migration process. A perspective*, „General System", 1970,Vol. 15.
10. Moore E. G., Brown L. A., *Urban acquaintance fields: an evolution of a spatial model,* „Environment and Planning",1970, No. 4.
11. Mokhtarian P. L., Bagley M. N., *Modeling Employees' Perceptions and Proportional Preferences of Work Locations: The Regular Workplace and Telecommuting Alternatives*, "Transportation Research", 2000, Part A, No. 34.
12. Weise G., *Die Anwendung mathematisch-statistischer Berechnungsmethoden für die Bearbeitung der Prognose des individuellen Strassenverkehrs*, „Die Strasse", 1965.
13. McCarthy T. R., Izraeli O., *Variations in travel distance, travel time and modal choice among SMSAs*, "Journal of Transport Economics and Policy", 1985, Vol. 19, No 2.
14. Shelejhovskij H. *The composition of the urban plan as the transport problem*, State Institute for Urban Design, Moscow, *1946.*