Verification of Bayesian Clustering in Travel Behaviour Research – First Step to Macroanalysis of Travel Behaviour

P Satra[1] **and J Carsky**[1]
[1] Department of Transportation Systems, Faculty of Transportation Sciences,
Czech Technical University in Prague, Konviktska 20, Praha 1, CZ – 110 00,
Czech Republic

E-mail: satrapet@fd.cvut.cz, carsky@fd.cvut.cz

Abstract. Our research is looking at the travel behaviour from a macroscopic view, taking one municipality as a basic unit. The travel behaviour of one municipality as a whole is becoming one piece of a data in the research of travel behaviour of a larger area, perhaps a country. A data pre-processing is used to cluster the municipalities in groups, which show similarities in their travel behaviour. Such groups can be then researched for reasons of their prevailing pattern of travel behaviour without any distortion caused by municipalities with a different pattern. This paper deals with actual settings of the clustering process, which is based on Bayesian statistics, particularly the mixture model. An optimization of the settings parameters based on correlation of pointer model parameters and relative number of data in clusters is helpful, however not fully reliable method. Thus, method for graphic representation of clusters needs to be developed in order to check their quality. A training of the setting parameters in 2D has proven to be a beneficial method, because it allows visual control of the produced clusters. The clustering better be applied on separate groups of municipalities, where competition of only identical transport modes can be found.

1. Introduction

The main goal of the research is to depict and analyse the modal split in the municipalities of the Czech Republic on the beginning of 21st century based on the statistical data.

One of the key problems one needs to deal with in the research is the absence of any nation-wide mobility survey in the Czech Republic. [1] A slightly better situation is on lower government levels. Many municipalities and some self-governing regions in the Czech Republic have conducted mobility surveys in order to facilitate a thorough research of the travel behaviour of their inhabitants. However, the methodology was not the same in all the cases, thus limiting the possibility to generalize the findings for the national level or at least for some group of municipalities. [2]

The most convenient available statistical data are possible to collect from the Czech national censuses 2001 a 2011. It includes responses of over 10 million inhabitants of the Czech Republic. Three key questions regarding the travel behaviour are included in the questioner, asking the inhabitants about their Commuting frequency, Duration of travel and Usually used mean of transport.

This data are assigned to every municipality in the Czech Republic. That makes 6 258 data sets, one for each municipality. The travel behaviour of whole municipalities, rather than individuals or households will be analysed in this research. This type of travel behaviour research can be classified as a macroscopic analysis of travel behaviour.

To facilitate the research, the municipalities will be grouped into groups suitable for analyses. A data pre-processing will be used in order to find groups (clusters) of municipalities, which will show similarities in their modal split. The data pre-processing is described in more details further.

2. Practical approach to clustering of municipalities

The clustering of municipalities into clusters showing similar modal split patterns was done by mixture model, using means of Bayesian statistics. [3] This process was theoretically described in our paper explaining advantages of such approach in this type of research, using default settings of mixture model. [4] This paper should be focusing more on the practical usage of mixture model, especially settings of its parameters in relation to the outputs of the model – the clusters of municipalities.

The parameters, which settings were necessary to research and optimize, were as follows: number of used components, initial coordinates of the centres of the components, size of the fixed initial variance of components and time, for which the fixed initial variance of components is used in the model.

The number of used components predetermines how many clusters will be searched in the data by the mixture model. The component is a multi-dimensional Gauss function with initial centre and initial variance. The mixture model uses estimation, resulting in change of the centre and variance of the component according to the particular pieces of data, which are assigned to the component. In the end of the clustering process, the component has position and shape corresponding to the data, which was assigned to it and the resulting cluster is consisting this data. The resulting cluster is a sub-set of data, which the component has found on its way through the whole data set during the clustering process.

It should be noted that some portion of the components might not attract any data at all and can remain empty after the end of the clustering process, especially if larger number of components is used. Thus, the resulting number of clusters can, and often is, smaller than the number of used components.

The initial coordinates of the centres of the components can be set based on some previous knowledge of the data. If starting with the clustering process from a scratch, it is advisable to use randomly set initial coordinates to minimise the interventions to the process based on autonomously running estimation. The initial coordinates of the centres of the components were set randomly in all computation runs of the clustering scripts in the research described in this paper.

In one time step of the model estimation (discrete time model), one piece of data is assigned to some component based on the data already assigned to the component in the previous steps. To allow components to attract some initial data before the estimation process starts to run autonomously, the initial variance of components is set rather large and fixed for several time steps. Otherwise the estimation process could disproportionately reduce the variance of components, in which no data was assigned in the initial steps. Thus, the Fixed Initial Variance of Components (further abbreviated as FIVC) and time t, for which the FIVC is applied (further abbreviated as time t), are deciding on how much are the components pre-filled with the data before the estimation continues to run autonomously, which significantly pre-determines the clustering process. Thus, the values of FIVC and time t is necessary to keep to only necessary minimum to minimise the intervention to the estimation process.

As the used data consist information on usage of six transport modes (Walking, Bicycle, Urban Public Transport, Bus, Train and Passenger Car), a six-dimensional data set was created. Such data set cannot be represented in any graphic, thus possibilities for control of the clustering process are limited. Therefore, the clustering process was divided into several consecutive steps. In some steps, it was then possible to apply visual control of the process combined with optimization of its settings parameters.

2.1. Training of the clustering script in 2D

The first step was training of the clustering script using only two-dimensional data. Results of clustering of such data can be pictured in a simple 2D plot. Based on visual control of the plots, it was possible to find relation between settings of the clustering script and conclusiveness of produced clusters.

In each run of the clustering script, only modal split of two transport modes was processed. For example, one of the two-dimensional data sets was consisted of only utilization rates of transport modes Urban Public Transport and Train per all the municipalities of the Czech Republic.

The four following figures are showing the outputs of two clustering processes with different setting parameters and thus different resulting clusters. Each of the red dots is one municipality, the X coordinate is showing the usage of the transport mode Urban Public Transport in each municipality, while the Y shows the Train usage. Figure 1 and Figure 2 are showing clusters resulting from clustering process in which FIVC = 0.100 and t = 50 were used. The cluster 2 then consists of municipalities, with high share of Urban Public Transport as well as municipalities with high share of Train, thus no clear preference of one transport mode can be seen in this cluster and it is considered inconclusive.

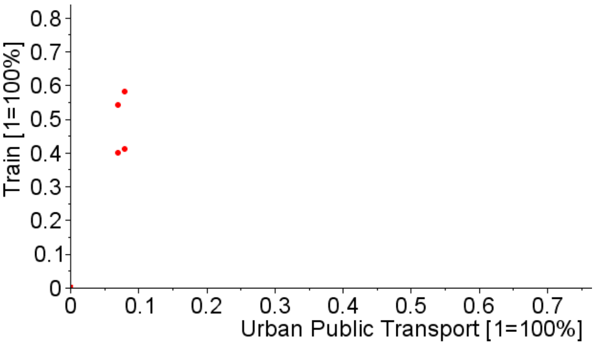


Figure 1. Shares of the transport modes in municipalities grouped in cluster 1 (two used components, FIVC = 0.100 and t = 50)

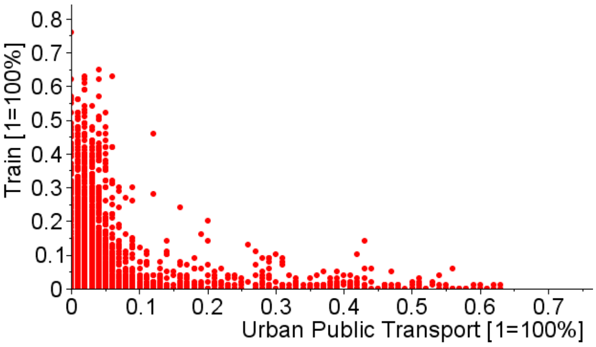


Figure 2. Shares of the transport modes in municipalities grouped in cluster 2 (two used components, FIVC = 0.100 and t = 50)

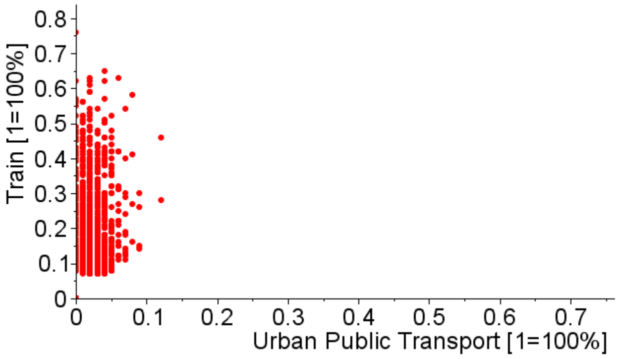


Figure 3. Shares of the transport modes in municipalities grouped in cluster 1 (two used components, FIVC = 0.150 and t = 59)

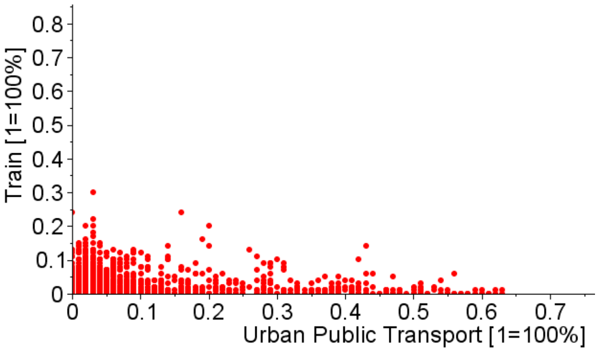


Figure 4. Shares of the transport modes in municipalities grouped in cluster 2 (two used components, FIVC = 0.150 and t = 59)

Whereas Figure 3 and Figure 4 are showing the resulting clusters when FIVC = 0.150 and t = 59 were used. The cluster 1 then consists only municipalities with higher Train usage and cluster 2 only municipalities, where Urban Public Transport is used and usage of Train is low. Such clusters are considered conclusive as they all over show higher usage of one transport mode then the other one has.

This visual quality check of the clustering has shown that settings FIVC = 0.150 and time t = 59 are better. By using this simple method, it was possible to try more variants of settings, to find the best.

From practical point of view, to optimize the time t first has proved to be better. Once finding the least value time t, at which the clustering process produces conclusive clusters, it is then possible to find the least value of FIVC (for already found least time t), at which the clustering process still produces conclusive clusters. The reason for this is that the change in quality of produced clusters is rapid and sudden in case of changing the time t, but it is slow and gradual in case of changing the FIVC.

By this optimization method, the least values of FIVC and time t, producing conclusive cluster in case of transport modes Urban Public Transport and Train were found to be FIVC = 0.071 and time t = 57. This values were verified by clustering of all other possible pairs of transport modes. More than 500 clustering script runs were evaluated. In the end, the least FIVC and time t values, at which 14 out 15 possible pairs produce conclusive clusters are FIVC = 0.072 and time t = 59. In case of the pair Walking and Bus, it was not possible to produce conclusive clusters at any settings, however by using 3 components (instead of two, used in case of the other 14 pairs), the produced clusters got better.

Those 14 pairs producing 2 clusters each and the pair number 15 producing 3 clusters make together 31 clusters, each of them with final coordinates of its centre – mean utilization rates of the two transport modes in each cluster. However, several similarities between those 31 clusters were found. For instance, if one cluster from clustering the transport modes Walking and Bus had final coordinates 0.15 and 0.36 and another cluster from clustering the transport modes Bus and Passenger Car had final coordinates of 0.36 and 0.41, one can see, it is still the same cluster, spreading over more dimensions. By assessing these similarities, it was found out that the data consist (at some level of uncertainty) 13 unique clusters at the most.

As a conclusion of the training process, the following setting were chosen to be the default settings for the next step: 13 used components, FIVC = 0.072 and time t = 59.

2.2. Optimization of the script for clustering of six-dimensional data

The full data set of 6 transport modes and 6258 municipalities was clustered. Since clustering of six-dimensional data does not produce any graphical representation of created clusters, it was necessary to find some parameters, which would serve for comparison of the results of particular runs and for optimization of the settings parameters. For this purpose, the Pointer Model Parameters (further abbreviated as PMP) of each component and the Relative Number of Data in each component (further abbreviated as ND%) were put in relation, creating relative performance indicator. The PMP and ND% play role of key performance indicators describing the results of the clustering process, thus also helping to evaluate, whether the initial setting parameters (e.g. FIVC and time t) were set appropriately.

Pointer model is a part of the mixture model, which points to which particular component should be the actual piece of data assigned in each time step. The data is then assigned to the component with the greatest weight (PMP variable). The total number of data assigned to the component during the clustering process (the final number of municipalities in the resulting cluster), can be than divided by the total data count (6258 municipalities) to get the ND% in each component (cluster respectively).

The cumulative sum of the PMP for each component should be in the end more or less similar to the ND% in the component. Larger discrepancies between these two values would mean that the amount of data is significantly smaller or larger than it should be according to the PMP’s. This relative performance indicator was already monitored as an auxiliary parameter during training of the clustering script – the correlation between this relative performance indicator and quality of produced clusters in terms of their conclusiveness was 76 %.

The previously researched values of settings, 13 used components, FIVC = 0.072 and time t = 59 were used as a first to cluster the whole data set. Seven clusters were empty; six clusters were containing from two pieces of data (the smallest one) to 3835 pieces of data (the largest one). This distribution of data can be considered as a pattern. Patterns are generally defined by the number of clusters per each size category, which are as follows: 10-100 pieces of data, 100-1000 pieces of data, over 1000 pieces of data (municipalities) in a cluster. The clusters with less than 10 pieces of data are considered insignificant and are not classified as any size category. The example of patterns is shown in the Table 1 and Table 2, where different categories are distinguished by colouring of the cells.

After the first run, the setting parameters (FIVC and time t) were alternated to verify, if it was possible to achieve better value of the relative performance indicator with some different settings, while several other patterns were discovered. Two most common patterns were optimized. The Table 1 shows the results of the optimized clustering process, producing the pattern named pattern A13. The setting parameters were FIVC = 0.200 and time t = 98. The relative performance indicator was 10.5 %.

Table 1. Pattern A13. One cluster with over 1000 pieces of data (cluster 12), four clusters with 100 – 1000 pieces of data (clusters 1, 3, 7 and 8), one cluster with 10 – 100 pieces of data (cluster 13).

	1	2	3	4	5	6	7	8	9	10	11	12	13
ND	537	2	640	0	1	0	565	238	0	0	0	4261	14

The following Table 2 shows the results of the optimized clustering process, producing the pattern named pattern B13, as a result of setting parameters FIVC = 0.063 and time t = 120. The relative performance indicator was 10.0 % (smaller the better).

Table 2. Pattern B13. Two clusters with over 1000 pieces of data (clusters 8 and 12), three clusters with 100 – 1000 pieces of data (clusters 1, 3 and 13), two clusters with 10 – 100 pieces of data (clusters 5 and 7).

	1	2	3	4	5	6	7	8	9	10	11	12	13
ND	193	0	547	0	24	0	26	4076	0	0	0	1292	100

3. Verification of the clusters

The next step in the research of the cluttering approach is the broader analysis of the previously produced clusters of selected patterns. For each transport mode in each cluster, the average value, minimum value, maximum value of its utilization were found and compared with the respective values for the whole data set. Based on the comparison, it was possible to define what are the prevailing transport modes in each cluster. An example of analysis of a cluster from optimized patter A13 can be seen in the Table 3, Table 4 and **Error! Reference source not found.**.

Table 3. Average, minimum, maximum and variance values of population count and shares of particular transport modes in the whole data set.

Population		Walking	Bicycle	UPT	BUS	Train	PC		
1635	AVG	0.19	0.08	0.04	0.38	0.05	0.26	AVG	**Whole data set**
7	MIN	0.00	0.00	0.00	0.00	0.00	0.02	MIN	
1169106	MAX	0.71	0.62	0.63	0.83	0.76	0.89	MAX	

Table 4. Average, minimum, maximum and variance values of population count and shares of particular transport modes in Cluster 1 (from pattern A13), which was evaluated as Train cluster.

Population		Walking	Bicycle	UPT	BUS	Train	PC		Cluster 1
624	AVG	0.15	0.08	0.02	0.21	0.29	0.25	AVG	**(Train cluster)**
29	MIN	0.00	0.00	0.00	0.00	0.08	0.07	MIN	
8208	MAX	0.63	0.62	0.58	0.78	0.59	0.76	MAX	

Table 5. Average, minimum, maximum and variance values of population count and shares of particular transport modes in Cluster 7 (from pattern A13), which was evaluated as Bicycle, Walking and Train cluster.

Population		Walking	Bicycle	UPT	BUS	Train	PC		Cluster 7
1433	AVG	0.23	0.20	0.03	0.24	0.10	0.21	AVG	**(Bicycle +Walking +Train)**
45	MIN	0.00	0.00	0.00	0.00	0.00	0.05	MIN	
17506	MAX	0.62	0.62	0.14	0.56	0.33	0.42	MAX	

For comparison, one cluster from an unnamed pattern, which were not going through optimization, is shown in the Table 6. This cluster is a bicycle cluster.

Table 6. Average, minimum, maximum and variance values of population count and shares of particular transport modes in Cluster 8 (from unnamed pattern, which was not result of the optimization process), which was evaluated as Bicycle cluster.

Population		**Walking**	**Bicycle**	**UPT**	**BUS**	**Train**	**PC**		**Cluster 8**
1188	AVG	0.17	0.23	0.02	0.31	0.05	0.22	AVG	**(Bicycle cluster)**
35	MIN	0.00	0.00	0.00	0.03	0.00	0.08	MIN	
16318	MAX	0.62	0.62	0.09	0.74	0.30	0.59	MAX	

An example of analysis of clusters from optimized patter B13 can be seen in the Table 7, Table 8 and Table 9.

Table 7. Average, minimum, maximum and variance values of population count and shares of particular transport modes in the whole data set.

Population		**Walking**	**Bicycle**	**UPT**	**BUS**	**Train**	**PC**		
1635	AVG	0.19	0.08	0.04	0.38	0.05	0.26	AVG	**Whole data set**
7	MIN	0.00	0.00	0.00	0.00	0.00	0.02	MIN	
1169106	MAX	0.71	0.62	0.63	0.83	0.76	0.89	MAX	

Table 8. Average, minimum, maximum and variance values of population count and shares of particular transport modes in Cluster 3 (from pattern B13), which was evaluated as Walking and Urban Public Transport cluster.

Population		**Walking**	**Bicycle**	**UPT**	**BUS**	**Train**	**PC**		**Cluster 3**
6624	AVG	0.38	0.06	0.13	0.18	0.03	0.22	AVG	**(Walking +UPT cluster)**
26	MIN	0.00	0.00	0.00	0.00	0.00	0.07	MIN	
97339	MAX	0.71	0.36	0.56	0.53	0.24	0.52	MAX	

Table 9. Average, minimum, maximum and variance values of population count and shares of particular transport modes in Cluster 13 (from pattern B13), which was evaluated as Urban Public Transport cluster.

Population		**Walking**	**Bicycle**	**UPT**	**BUS**	**Train**	**PC**		**Cluster 13**
28281	AVG	0.18	0.03	0.47	0.09	0.01	0.22	AVG	**(UPT cluster)**
74	MIN	0.03	0.00	0.28	0.01	0.00	0.11	MIN	
1169106	MAX	0.37	0.10	0.63	0.26	0.06	0.37	MAX	

3.1. Verification method

A compliance with such prerequisites is very simple to be shown using the Geographic Information System (further abbreviated as GIS). On the **Error! Reference source not found.**, one can see the Train cluster from the pattern A13. The brown dots are the municipalities in the cluster and the green (electrified) lines and black (non-electrified) lines are the railway lines in the Czech Republic. The municipalities show very strong adherence to the railway lines.

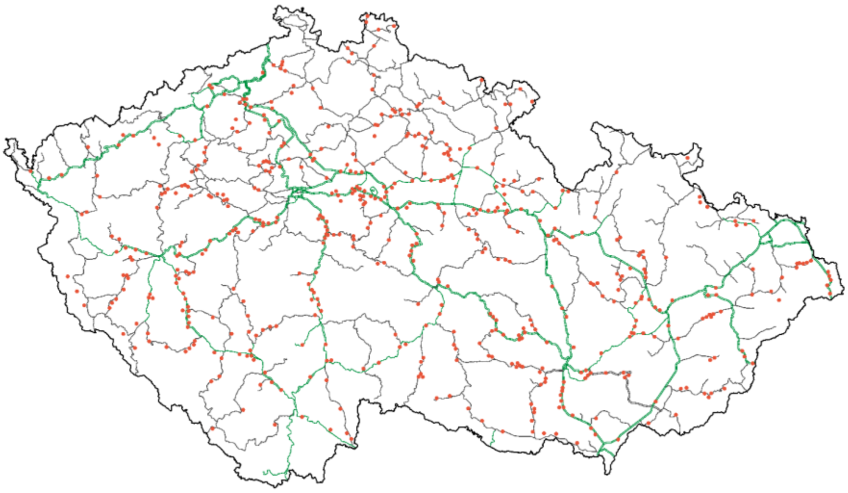
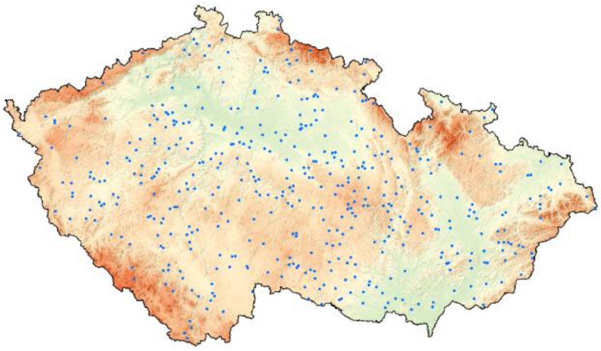


Figure 5. Graphic representation of municipalities (brown dots) clustered in the Train cluster from the pattern A13. The lines are the railway lines in the Czech Republic.

Two following figures are showing comparison of two Bicycle clusters. On the Figure 6 is the cluster 8 from the unnamed pattern, which did not go through optimization. The Figure 7 shows the cluster 7 from the optimized pattern A13. The blue dots are showing the municipalities labelled by the particular clustering process as the ones with significant bicycle usage. It is the cluster 7 on the Figure 7, which shows better adherence of the bicycle municipalities to the low lands of the Czech Republic with the most level terrain.



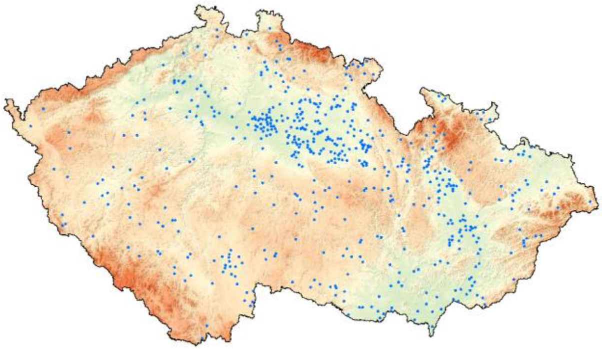


Figure 6. Graphic representation of municipalities (blue dots) clustered in the cluster 8 from the unnamed pattern, which was evaluated as a Bicycle cluster.

Figure 7. Graphic representation of municipalities (blue dots) clustered in the cluster 7 from the pattern A13, which was evaluated as Bicycle, Walking and Train cluster.

Two following figures are showing comparison of two Urban Public Transport clusters. On the Figure **8** is the cluster 3 from optimized pattern B13. The Figure 9 shows the cluster 13 from optimized pattern B13. The big blue dots are representing the towns or cities of the Czech Republic having their own Urban Public Transport. The brown dots are showing the municipalities labelled by the particular clustering process as the ones with significant urban public transport.

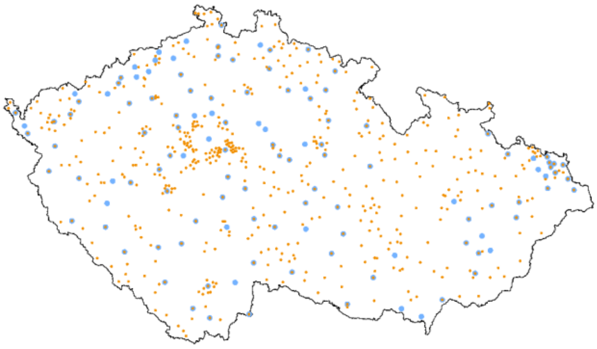


Figure 8. Graphic representation of municipalities (brown dots) clustered in the cluster 3 from the pattern B13, which was evaluated as a Walking and Urban Public Transport cluster.

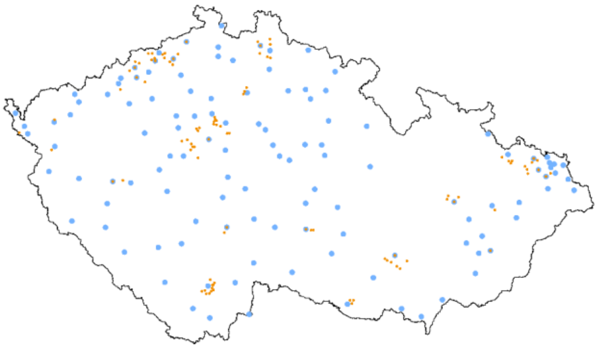


Figure 9. Graphic representation of municipalities (brown dots) clustered in the cluster 13 from the pattern B13, which was evaluated as an Urban Public Transport cluster.

While comparing these two figures, the results of cluster 13 on the Figure 9 are more convincing as the municipalities using much the Urban Public Transport are in the proximity of the ones actually having it, which is not exactly true in case of the cluster 3 represented by the Figure 8. However, in case of the latter one, more towns and cities having their own Urban Public Transport are actually included in the cluster.

4. Conclusions

1) The optimization of clustering process for particular patterns is working with only about 75 % reliability and some verification of the produced clusters is still necessary.
2) The graphic representation in GIS has confirmed that the Train cluster (cluster 1 from the pattern A13) is suitable for the research of dependency of usage of transport mode Train. It is because the Train cluster is consisting municipalities both having high share of the transport mode Train and that also have access to this mean of transport.
3) The graphic representation in GIS has confirmed that the cluster 7 from the optimized pattern A13 (Bicycle, Walking and Train cluster) is more suitable for the research of dependency of usage of transport mode Bicycle than the cluster 8 from the unnamed pattern (Bicycle cluster), which did not go through the optimization process. It is because the cluster 7 is consisting municipalities both having high share of the transport mode Bicycle and that are also advantageously located in areas with level terrain.
4) On the other hand, there are no large cities in the cluster 7 from the optimized pattern A13, however, there are few cities in the Czech Republic well known for high share of Bicycle transport, namely Hradec Kralove a Pardubice.
5) The graphic representation in GIS has shown that the cluster 13 from the pattern B13 (Urban Public Transport cluster) is more suitable for the research of dependency of usage of transport mode Urban Public Transport than the cluster 3 from pattern B13 (Walking and Urban Public Transport cluster). It is because the cluster 13 is consisting municipalities both having high share of the transport mode Urban Public Transport and that also have access to this mean of transport.
6) On the other hand, cluster 3 from pattern B13 (Walking and Urban Public Transport cluster) consists more towns and cities, which actually have the Urban Public Transport of their own. In those towns and cities, the Urban Public Transport is surely significantly used, but still not enough to be clustered into the clear Urban Public Transport cluster.
7) Looking at the conclusions 3) and 5), the fact that some cluster is showing increased usage of a sole transport mode does not automatically means is more suitable for research of travel behaviour concerning that mode. Regarding the mode Bicycle, the cluster showing increased usage of three modes Bicycle, Walking and Train has proven to be more suitable for the research of Bicycle usage than the cluster with increased usage of only the Bicycle transport. It is the opposite

way in case of the Urban Public Transport, where the cluster with sole Urban Public Transport increase is the more suitable one.

8) Based on the conclusions 4), 6) and 7), one can tell that for the further research in macro analysis of travel behaviour of municipalities, it will be necessary to apply the clustering on separate groups of municipalities, where a similar competition of transport modes can be found. Those groups would be:
   a. Municipalities with access to Train transport and Urban Public Transport
   b. Municipalities with access to Train transport
   c. Municipalities with access to Urban Public Transport
   d. Municipalities with access to only Walking, Bicycle, BUS and Passenger Car

   This approach will have its analogy in the classical (micro) analysis of travel behaviour of households, where households are classified based on the fact whether they have access to a Passenger Car or not.

References

[1] Senk P, Kouril P. 2014, Pruzkumy dopravniho chovani v CR a zahranici. Potrebujeme narodni pruzkum, *Dopravni inzenyrstvi* **9**(1) 24

[2] Macejka P. 2014, *Vliv Struktury Osidleni Kraje a Vybranych Sociekonomickych Ukazatelu na Dopravni Poptavku* (Ostrava: VSB – Technical University of Ostrava) p 140

[3] Nagy I. 2012, Stochastic Systems, *Prague: Czech Technical University in Prague Faculty of Transportation Sciences*, p 29-32

[4] Carsky J. and Satra P., 2017, *Proceedings of the 5th Annual International Conference on Architecture and Civil Engineering.*

Acknowledgments

We would like to thank to Mr. Pavel Kanka for his kind assistance with the GIS.