ISSN 2278-3091
Volume 9, No.1.4, 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse5291.42020.pdf
https://doi.org/10.30534/ijatcse/2020/5291.42020

Filtration and Segregation of Origin-Destination Pairs for Modal Split Research

Petr Satra[1], Jiri Carsky[2]
[1]Czech Technical University in Prague Faculty of Transportation Sciences, Czech Republic, satrapet@fd.cvut.cz
[2]Czech Technical University in Prague Faculty of Transportation Sciences, Czech Republic, carsky@fd.cvut.cz

ABSTRACT
The goal of the research is to study the mode choice on the Origin-Destination pairs (O-D pairs) between municipalities of the Czech Republic. However, the data set of O-D pairs, which are described by number of travelers commuting to work or school from one municipality to another, is consisting over 50 % of O-D pairs with only one traveler. Modal split on these O-D pairs is then distorted in favor of only one mode of transport, making the whole data set misleading. The paper is presenting and evaluating methods for filtration and removal of O-D pairs of lesser importance from the data set. The options for modal split study are also improved by segregation of the data set into groups of O-D pairs according to their direction characteristics.

Key words: Data pre-procession and filtering, modal split, municipalities, transport modes, Origin-Destination pairs.

1. INTRODUCTION

The subject of our research is modal split on O-D pairs between municipalities of the Czech Republic, especially those on lower levels of governmental hierarchy. The O-D pair is defined as group of travelers, commuting from the municipality of origin to the municipality of destination.The input data were taken from the Czech national census 2011 [1], so far, the last nationwide census in the country. In this census, over 10 million inhabitants of the Czech Republic were asked to describe their commute to work or school. Questions were focused on finding the Commuting frequency, Duration of travel and used Transport modes. Based on the other data collected by the questionnaire, such as home address and location of work or school, it was possible for the Czech Statistical Office to distinguish the inhabitants working or studying in a municipality of their residence from the travelers, who are leaving their municipality of residence, when traveling to work or school. The travelers were then assigned to the O-D pairs between the municipalities. Only outbound travel was researched.

The 94 possible combinations of transport modes used in the census data set were aggregated into 10 representing transport modes. These aggregated modes are described in the Table 1.

The original census data set is consisting O-D pairs between all municipalities of the Czech Republic, having character of big data in transportation [2]. However, our focus are the O-D pairs on the local level, between municipalities on lower levels of governmental hierarchy. As described by [3], the 'Municipalities with Extended Powers' (further abbreviated as MEP) are in the environment of the Czech Republic functioning as natural centers of catchment areas for local commuting. There are 205 MEPs in the country, however, 12 of them are at the same time a regional capital, thus have much broader catchment areas. Keeping in mind the interest in O-D pairs between municipalities of lover levels of governmental hierarchy, the O-D pairs of 12 regional capitals and the Capital city of Prague were omitted from the data set. Intercity O-D pairs between the MEPs were also excluded. After the initial reduction, the starting data set for our research is including O-D pairs from 193 MEPs to municipalities on subordinate levels and vice versa plus O-D pairs between the municipalities on the subordinate levels. This data set will be further referred as "Non-filtered". It includes all the O-D pairs on the required local level. It includes 6 236 municipalities as all 13 capitals and 2 municipalities with no O-D pairs on local level were removed from the data set.All the 193 included MEPs are having status of town, ranging approx. from 2 800 to 76 700 inhabitants with average population size of approx. 16 000 inhabitants. Their catchment areas vary from one to another, but the study of catchment areas of MEPs in the South Bohemian region [3], has shown these are approximately like the actual areas of the administrative districts of MEP. Average area size of administrative district of MEP in the Czech Republic is about 385 km^2.

Table 1: Names, description and abbreviations of aggregated transport modes used in the research

Name of aggregated transport mode	Description	Abbreviation
Bus	Any bus, which is not part of the Urban Public Transport	Bus
Train	Any train service	Train
Urban Public Transport	Metro, tram, trolleybus and bus in urban areas	UPT
Car Driver	Taking a car as a Driver	Driver
Car Passenger	Sharing a car as a Passenger	CarPass
Public Transport +	Using other modes in combination with Bus, Train, UPT	PT+
Bike	Riding a Bike	Bike
Combinations of Public Transport	Combining only Bus, Train, UPT	PTcom
Walk	Walking	Walk
Other Combinations	Combining Car Driver, Car Passenger, Bike, Motorbike	Rest

2. FILTRATION OF O-D PAIRS

The "Nonfiltered" data set is consisted of 126 973 O-D pairs. However, 71 205 of them are O-D pairs with only one traveler, which is over 56 % of all O-D pairs. In other words, in case of 71 205 municipality O-D pairs, there is only one traveler between them in one direction. Whichever transport mode the one traveler choses, will be the only transport mode used on that O-D pair. Then the modal split, describing the distribution of shares of transport modes on this O-D pair, is only consisted of the one used transport mode, which has share of 100 % and all the other modes have share of 0 %. This is an extreme distortion of the modal split in favor of only one transport mode.

2.1 Filtering methodology

Considering the number of O-D pairs with one traveler causing the distortion, one can consider the whole data set to be misleading about the overall modal split. The solution is to remove the O-D pairs of lesser importance from the data set. It is assumed that most of the O-D pairs with one traveler are of lesser importance. Similarly, all other O-D pair with small number of travelers, but the O-D pairs with one traveler were chosen to be the monitored parameter as they cause the greatest distortion.

But how to assess the importance of the O-D pairs? Which O-D pairs must stay in the data set to keep its informative value? Starting with the latter one, the requirements were set as follows:

A. As this research should serve to the municipal and regional level of government, first of the requirements is to sustain all the municipalities in the data set. In other words, there should not be a municipality, whose O-D pairs will be all removed from the data set and so the municipality with them.

B. Second requirement is to maximize the amount of the travelers in the data set. This is driven by the will to conduct a research, which will have an impact to maximum of travelers. Practically, this is about minimizing the loss of travelers from the data set caused by removing of O-D pairs.

Simple removing all O-D pairs with one traveler does not meet the criterion of sustaining all the municipalities in the data set because it results in removing 13 municipalities from it. This is due to the specific municipal structure of the Czech Republic. At the time of the census, there were 6251 municipalities in the country, ranging from 17 to 1 268 796 inhabitants, having on average 1 670 inhabitants. Since the municipalities can have so small population as 17 inhabitants, it is then expectable, the O-D pairs originating or terminating in such municipality will have only one traveler. At the same time, the O-D pair with one traveler is an important O-D pair for this municipality as it was used by over 5 % of population of municipality and perhaps by more than 10 % of the workers and pupils residing in it.

This gives the preview how the importance of the O-D pairs can be assessed. Description, evaluation and comparison of three methods of assessment of O-D pair importance will follow.

2.2 Filtration method "T_{ij}"

Assessment of O-D pair importance in this method is based on the formula (1) presented by Afonso and Venancio [4]:

$$T_{ij} = \frac{C_{ij}}{\min(r_i, r_j)} \tag{1}$$

where:
T_{ij} – strength of commuting tie between two municipalities i and j
C_{ij} – number of travelers (workers and pupils) residing in the municipality i and commuting to municipality j to work or school
r_i – number of all workers and pupils residing in municipality i
r_j – number of all workers and pupils residing in municipality j

The parameter T_{ij} can be used for assessment of the O-D pair importance. Afonso and Venancio recommend considering the O-D pairs with T_{ij} above 0,02 to be "stronger" (important from the perspective of this paper).

For example, if there would be a municipality i, where in total 100 workers and pupils are residing and two travelers (no difference if workers or pupils) are commuting from there to municipality j, where in total 500 workers and pupils are residing. Then, the C_{ij} is 2 and smaller of the two r is r_i, which is 100. By dividing 2/100 we get that the T_{ij} of this O-D pair is 0,02; just below the set criteria, thus it is an unimportant O-D pair sentenced to be removed from the data set.

2.2.1 Experiences from using the "T_{ij}" method

The recommended level of importance of 0,02 did not meet the criterion of sustaining all the municipalities in the data set because it resulted in removing 120 municipalities from it. Less strict level of importance must have been found to keep all the municipalities in. The maximum level of importance, for which all 6 236 municipalities remains in the data set is 0,0042796. The resulting data set was named "**T_{ij} filtered**". Parameters of this data set can be seen in the Table 2, their modal split in the

Table 3. Using the description

Table 2: Parameters of the compared data sets of O-D pairs. Mun's = Municipalities; ANTP = Average Number of Travelers on (one) O-D pair; ANUMT = Average Number of Used Modes of Transport (on one O-D pair); Pw1T = O-D pair with One Traveler; T = Travelers; %↓ = decrease in % compared to the base value; %↑ = increase in % compared to the base value

File \ Parameter	O-D pairs	Mun's	Travelers	%↓	ANTP	%↑	ANUMT	%↑	Pw1T	%↓	Pw1T/T
Nonfiltered	126 973	6 236	769 340	base	6,06	base	1,78	base	71 205	base	9,26%
T_{ij} filtered	69 765	6 236	690 876	-10,2%	9,90	163%	2,29	128%	25 263	-64,5%	3,66%
P_{ij} filtered	19 614	6 236	526 198	-31,6%	26,83	443%	3,78	212%	1 512	-97,9%	0,29%

***P*$_{ij}$ *comp* filtered**	19 850	6 236	532 635	-30,8%	26,83	443%	3,79	213%	1 512	-97,9%	0,28%

Table 3: Modal split in the compared data sets of O-D pairs

File \ Parameter	Bus	Train	UPT	Driver	Car Pass	PT+	Bike	PTcom	Walk	Rest
Nonfiltered	14,2%	5,0%	1,0%	55,1%	7,6%	5,4%	1,7%	4,8%	2,9%	2,3%
***T*$_{ij}$ filtered**	18,1%	3,5%	0,8%	53,1%	8,9%	5,7%	2,3%	3,2%	1,9%	2,7%
***P*$_{ij}$ filtered**	29,9%	3,2%	1,1%	40,5%	10,0%	6,8%	2,2%	2,5%	1,6%	2,2%
***P*$_{ij}$ *comp* filtered**	29,8%	3,2%	1,1%	40,5%	10,0%	6,8%	2,2%	2,5%	1,6%	2,3%

in the caption of the table, the parameters are self-explanatory. The parameter Travelers is a total sum of all C_{ij} over all O-D pairs in the data set.

As can be seen in the Table 2, the number of O-D pairs with one traveler has decreased significantly (by over 64%) after the filtering out the O-D pairs assessed unimportant by the T_{ij} method. However, large amount of them (25 263) remain in
the data set, which is over one third of the O-D pairs in the "T_{ij} filtered" data set. If over one third of the data is still distorted, there is a solid ground to consider the whole resulting data set to be misleading. From this perspective, the filtering method T_{ij} did not meet the expectations and further methods need to be introduced.

2.3 Filtration methods based on "P_{ij}" parameter

The T_{ij} method focuses on the assessing the importance of the O-D pair based on the proportion of the travelers using the O-D pair to the travelers (workers or pupils) residing in the smaller of two municipalities. The drawback is that it does not deal with the proportion of the travelers using the O-D pair to the total amount of travelers leaving the municipality of origin to work or school.

2.3.1 Introduction of the parameter "P_{ij}"

To take into consideration the total number of travelers leaving the municipality, the P_{ij} parameter was introduced, defined by (2):

$$P_{ij} = \frac{C_{ij}}{\sum_{1}^{n} C_{ix}} \tag{2}$$

where:
P_{ij} – share of O-D pair from municipality i to municipality j on all travelers leaving the municipality i
C_{ij} – number of travelers (workers and pupils) residing in the municipality i and commuting to municipality j to work or school
C_{ix} – number of travelers (workers and pupils) residing in the municipality i and commuting to municipality x to work or school, where $x = (1 \dots n)$
1 – the first O-D pair from municipality i
n – the last O-D pair from municipality i

The number of O-D pairs originating in the municipality i or the total amount of travelers leaving the municipality i is not in any sense dependent on these parameters in municipality j, therefore no parameter of municipality j is included in the P_{ij} definition.

2.3.2 Experiences from using the "Pij" method

The maximum level of importance, for which all 6 236 municipalities remains in the data set is 0,06578947. The resulting data set was named "***P*$_{ij}$ filtered**". Parameters of this data set are in the Table 2. One can see there that the number of O-D pairs with one traveler has decreased significantly (by nearly 98%) after removing the O-D pairs assessed unimportant by the P_{ij} method. The remaining number of O-D pairs with one traveler is 1 512, which is less than 8 % of all O-D pairs in the "P_{ij} filtered" data set.

However, the cost for the successful reduction of such a number of O-D pairs with one traveler is the reduction of the number of travelers in the data set. Compared to the "Nonfiltered" data set, over 31 % of travelers was removed from the data set by using the P_{ij} method. To evaluate, whether the loss of nearly one third of travelers is justifiable by removing of nearly 98 % O-D pairs with one traveler, the parameter Pw1T/T was introduced, which is describing the ratio of lost travelers per removed O-D pairs with one traveler. The P_{ij} method loses only 0,29 of traveler per one removed O-D pair with one traveler comparing the 3,66 travelers in case of T_{ij} method, showing better performance in this parameter.

An important drawback of the P_{ij} method was found in the end. It has assessed some O-D pairs with over 100 travelers as unimportant. It was due to fact these O-D pairs are originating in large towns, from which over 2 600 workers and pupils commute and where the originating O-D pair with 100 travelers has share below 6,578947 %. However, any loss of O-D pair with 100 travelers is not desirable, as these are not causing any distortion of the modal split and are used by large number of travelers.

2.3.3 Introducing the "P_{ij} *comp*" method

Therefore, a decision has been made to include a compensation to the P_{ij} formula, which would gradually increase the P_{ij} parameter of the O-D pairs based on their increasing number of travelers. The calculation of the compensation was designed in a way the compensation is 0 for O-D pairs with one traveler and gradually increases to be just as large to keep all O-D pairs with over 100 travelers in the data set. The P_{ij} *comp* method is then defined by (3):

$$P_{ij}comp = \frac{C_{ij}}{\sum_{1}^{n} C_{ix}} + \frac{\left(C_{ij} - 1\right)}{K} \tag{3}$$

where:
P_{ij} *comp* – share of O-D pair from municipality *i* to municipality *j* on all travelers leaving the municipality *i*, increased by a compensation, which is given by the second fraction in the formula
C_{ij}, C_{ix}, 1 and *n* are as described in P_{ij} definition
K – a coefficient, which is given by (4):

$$K = \frac{\left(C_{xy} - 1\right)}{\left(LoI - P_{xy}\right)} \tag{4}$$

where:
C_{xy} – number of travelers (workers and pupils) travelling on an O-D pair with over 100 travelers (between municipalities *x* and *y*), which has the least P_{ij} – which is denoted as P_{xy} for this O-D pair
LoI – level of importance, the same as in P_{ij} method (0,06578947)

2.3.4 Experiences from using the "P_{ij} *comp*" method

Looking back to the Table 2, the key performance indicator, the Pw1T/T, is showing the best value in case of the **"P_{ij} *comp* filtered"** data set. Thus, P_{ij} *comp* method is considered the best performing in filtering the unimportant O-D pairs from the data set.

2.4 Impact of filtration to modal split

Based on the results in the

Table 3, the finding are as follows:

1. The transport modes, which are showing decrease in their share after filtration are Driver and Walk. These individual transport modes are a natural choice of the travelers for O-D pairs, which are so unimportant that there is no public transport service. After removing these unimportant O-D pairs, driving and walking decreases.

2. The usage of modes Bus, Car Passenger and PT+ increases. These are example of mass transport modes, which are flexible in meeting the transport demand, thus easily deployable on important O-D pairs.

3. On the contrary, the decreasing usage of mode Train with decreasing number of O-D pairs with one traveler might pointing at the fact the mode Train is not so flexible in its deployment due to fixed railway network, which might be from historical reasons connecting municipalities, with nowadays mutually unimportant O-D pair.

4. The decrease of the mode PTcom implies the mode Train is more important within this combined mode then the Bus.

3. SEGREGATION OF O-D PAIRS

The next step in the research is a segregation of the O-D pairs. In this phase, the O-D pairs are segregated from the data set into separate files based on their direction characteristics. Each O-D pair between two municipalities always has only one direction, for example from municipality *i* to municipality *j*. If there is an O-D pair in the opposite direction from municipality *j* to municipality *i*, it is a different O-D pair. There is no requirement for a bi-directional connection of two municipalities, so between two connected municipalities, there can exist one or two O-D pairs. Existence of one /**1**/ or two /**2**/ O-D pairs between two municipalities is the first direction characteristics.

The second characteristic is giving into relation the governmental level of the connected municipalities. If the municipality of origin is a MEP and the municipality of destination is one of the municipalities from subordinate governmental level, the direction is marked as a Down /**D**/ direction. The opposite situation is marked as an Up /**U**/ direction. If both connected municipalities are below the MEP level, their mutual O-D pair is marked as a Tangent /**T**/ direction.

The third direction characteristic is further distinguishing the Tangent directions according to the total number of workers and pupils residing in the municipalities. If the municipality of origin is having higher total number of workers and pupils and the municipality of destination smaller, the direction is marked as a Down' /**D'**/ direction. The opposite situation is marked as an Up' /**U'**/. The characteristics /D'/ and /U'/ are a tertial characteristics used only for the distinguishing the Tangent directions.

By combining these direction characteristics, eight different direction groups (files) of O-D pairs can be created. The directions are illustrated on the Figure 1.

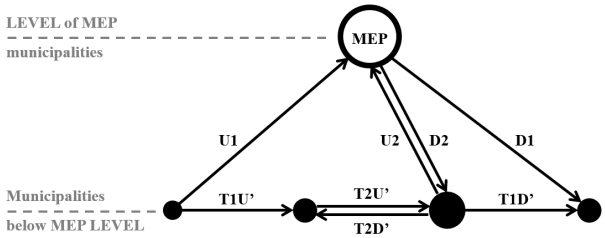

Figure 1: Scheme of different direction groups. The size of municipalities below MEP level reflects the number of residing workers and pupils

The O-D pairs from the data set filtered by the proven P_{ij} *comp* method were segregated into 8 files according to their direction characteristic groups and named accordingly. The O-D pairs from the "Nonfiltered" data set were also segregated for comparison.

3.1 Parameters of segregated files

The first part of comparison of parameters of files segregated from "Nonfiltered" and "P_{ij} *comp* filtered" data sets can be seen in the

Table 4, which is using conditional formatting to highlight the differences between the files and their change. The change from "Nonfiltered" to "P_{ij} *comp* filtered" files is expressed by a ratio calculated by their simple division. The comparison of the segregated files gives us more insights of what happened to the data set during the filtration. The key findings from

Table 4 are as follows (continuing in numbering of findings throughout the paper):

5. The changes the filtration causes to the data set are not evenly distributed over the files segregated by the direction characteristics. Thus, it is important to distinguish these direction groups of O-D pairs as they behave differently.

6. From some pairs of O-D pairs /2/ between two municipalities, one O-D pair was found unimportant, thus the second O-D pair was no longer part of the bi-directional group but has moved to single direction groups /1/ of O-D pairs. Example can be seen by files U1 and T1U', whose number of O-D pairs has dropped less, and their number of travelers even rose. This is because the bi-directional O-D pairs are on average having more travelers than single direction O-D pairs.

Table 4: Comparison of parameters of files of O-D pairs segregated from "Nonfiltered" (N) and "P_{ij} *comp* filtered" (F) data set. ANTP = Average Number of Travelers on (each) O-D pair; ANUMT = Average Number of Used Modes of Transport (on each O-D pair)

File	O-D pairs			Travelers			ANTP			ANUMT		
	N	F	N/F	N	F	N/F	N	F	N/F	N	F	N/F
D1	5414	80	1.48%	8213	3292	40.1%	1.52	41.15	2713%	1.165	5.438	467%
D2	9247	600	6.49%	93315	47401	50.8%	10.09	79.00	783%	2.368	6.498	274%
T1D'	17666	574	3.25%	28648	8846	30.9%	1.62	15.41	950%	1.209	3.294	272%
T1U'	35432	8257	23.30%	74664	88145	118.1%	2.11	10.68	507%	1.345	2.879	214%
T2D'	11346	378	3.33%	51959	10795	20.8%	4.58	28.56	624%	1.893	4.772	252%
T2U'	11360	378	3.33%	89798	8695	9.7%	7.90	23.00	291%	2.401	4.399	183%
U1	27261	8983	32.95%	87058	278231	319.6%	3.19	30.97	970%	1.599	4.180	262%
U2	9247	600	6.49%	335685	87230	26.0%	36.30	145.38	400%	3.992	7.098	178%

7. Parameters ANTP and ANUMT have perfect match in their differences between the files and their change, highlighted by the same color pattern. This high correlation between average number of travelers on relations with average number of modes of transport used on each O-D pair means that more travelers are commuting on the O-D pair, the larger number of transport modes they chose. This implies that increasing transport demand is never satisfied by increased capacity of modes but by a competition of increased number of modes.

The second part of comparison of parameters of files segregated from "Nonfiltered" and "P_{ij} *comp* filtered" can be seen in the Table 5, where the key finding is as follows:

8. The color trends in change of parameters Pw1T a Pw1T/T are similar, which implies, the parameter Pw1T (number of O-D pairs with one traveler) is more significant when calculating the ratio Pw1T/T than the parameter T (Travelers). It implies that the number of O-D pairs with one traveler was reduced due to filtration more than the number of travelers. The trends of change (N/F) differ in the files T2U' and U1, where the number of travelers goes through greatest decrease, increase respectively (see

9. Table 4), which is connected to the finding 6.

Table 5: Comparison of parameters of files of O-D pairs segregated from "Nonfiltered" (N) and "P_{ij} *comp* filtered" (F) data set. Pw1T = O-D pair with One Traveler; T = Travelers

File	Pw1T			Pw1T/T		
	N	F	N/F	N	F	N/F
D1	4385	0	0.00%	53.39%	0.00%	0.00%
D2	3246	0	0.00%	3.48%	0.00%	0.00%
T1D'	13289	35	0.26%	46.39%	0.40%	0.85%
T1U'	24066	993	4.13%	32.23%	1.13%	3.49%
T2D'	4825	1	0.02%	9.29%	0.01%	0.10%
T2U'	3404	12	0.35%	3.79%	0.14%	3.64%
U1	16663	471	2.83%	19.14%	0.17%	0.88%
U2	1327	0	0.00%	0.40%	0.00%	0.00%

3.2 Changes to modal split in segregated files

As the parameters of the segregated files are changing before and after the filtration, so is the modal split in these files. Due to large extend of the modal split tables, only changes to the modal split will be presented. But we will use the opportunity to present the change expressed by two different approaches.

In the first approach, the change is expressed by simple subtraction of shares of transport modes in filtered files from shares of transport modes in nonfiltered files. The resulting differences in the share of transport modes can be either positive – the share of the transport mode in the file has increased after the filtration, or negative – the share has decreased after the filtration. This type of change is depicted in the Table 6. Using the conditional formatting of the table, the increase in share is highlighted by scale of green color and the decrease in scale of red. The key findings from the Table 6 are as follows:

10. The transport mode Driver is decreasing in every segregated file and except one also shows the largest decrease in its absolute magnitude.

11. The exemption is the file representing the direction groups U1, where the large decrease is shown by

modes Train, PTcom and Walk. Special behavior of the files U1 is related to the finding 6.

In the second approach, the change is expressed as a relative difference of shares of the transport modes. The share of transport mode in filtered file is subtracted from share of transport mode in nonfiltered file and divided by the larger of the two. This type of change is depicted in the Table 7, from which the key findings are as follows:

12. The transport modes Driver and PTcom are decreasing in all segregated files. In most of the files, the mode Train is decreasing while the mode PTcom is experiencing the largest decrease relative to size of its share in the files. This relates to findings 1, 3 and 4.
13. The transport modes Bus, Car Passenger and PT+ are increasing in all segregated files. The mode Bus is also among the most increasing modes in every file. This relates to the finding 2.
14. If we neglect the result from the file D1 with only 80 O-D pairs left, the most even increase over all files is experiencing the mode Bike. It implies that the choice of this mode is less dependent on importance of the O-D pairs, which were removed by filtration from the files with uneven manner, see Table 5.

Table 6: Change of modal split in segregated files before and after the filtration, expressed as subtractions of transport mode shares in filtered (F) files and nonfiltered files (N)

File	F-N	Bus	Train	UPT	Driver	Car Pass	PT+	Bike	PTcom	Walk	Rest
D1	Change	17,2%	0,83%	-0,57%	-23,8%	7,61%	4,46%	-0,39%	-0,89%	-6,27%	1,85%
D2	Change	7,64%	-0,56%	2,00%	-16,4%	3,52%	2,30%	1,44%	-0,95%	-0,18%	1,23%
T1D'	Change	15,9%	-1,21%	0,39%	-25,2%	5,70%	0,75%	3,47%	-1,92%	0,86%	1,21%
T1U'	Change	19,1%	-1,46%	0,34%	-19,8%	2,78%	0,56%	1,49%	-2,91%	-0,22%	0,11%
T2D'	Change	10,0%	-0,54%	0,43%	-17,9%	1,59%	1,61%	2,17%	-0,80%	2,25%	1,17%
T2U'	Change	11,0%	-0,72%	0,65%	-16,8%	2,16%	0,86%	1,79%	-1,08%	2,33%	-0,21%
U1	Change	12,4%	-5,50%	-0,40%	-1,63%	3,14%	1,32%	0,39%	-5,43%	-4,52%	0,19%
U2	Change	4,98%	-1,79%	1,92%	-6,68%	1,57%	1,17%	1,28%	-2,68%	-0,10%	0,33%

Table 7: Change of modal split in segregated files before and after the filtration, expressed as a relative difference of shares of the transport modes. The share of transport mode in filtered file (F) is subtracted from share of transport mode in nonfiltered file (N) and divided by the larger of the two, max(F,N)

File	F-N / max(F,N)	Bus	Train	UPT	Driver	Car Pass	PT+	Bike	PTcom	Walk	Rest
D1	Change	71,5%	11,6%	-33,2%	-39,8%	52,6%	51,0%	-64,5%	-16,6%	-96,9%	51,3%
D2	Change	38,8%	-12,4%	69,3%	-27,4%	30,4%	29,6%	47,2%	-32,3%	-9,38%	31,8%
T1D'	Change	68,0%	-47,7%	33,9%	-37,4%	42,3%	15,3%	67,8%	-63,0%	26,4%	30,6%
T1U'	Change	59,6%	-46,0%	32,6%	-32,9%	25,5%	10,4%	46,9%	-72,0%	-10,8%	4,17%
T2D'	Change	45,2%	-21,3%	35,2%	-30,3%	14,4%	25,1%	32,9%	-56,6%	52,8%	25,4%
T2U'	Change	33,7%	-26,7%	36,9%	-33,5%	19,5%	13,4%	34,2%	-51,7%	59,0%	-7,54%
U1	Change	42,8%	-54,7%	-31,0%	-3,83%	36,0%	16,2%	45,0%	-57,7%	-81,8%	11,9%
U2	Change	16,8%	-24,9%	62,2%	-15,7%	18,0%	13,7%	59,2%	-45,7%	-6,24%	17,1%

4. CLUSTERING OF MODAL SPLIT IN SELECTED FILES

The O-D pairs from segregated files U1 were further grouped into clusters. The clustering of the O-D pairs was done by the Mixture model [5], using means of Bayesian statistics [6], which allowed to group the O-D pairs into the clusters based on similarity of their modal split patterns.

For comparison, clustering was done in files from "Nonfiltered" and "P_{ij} *comp* filtered" data set. The files U1 were selected for clustering because they consist high number O-D pairs and travelers. They are also attracting the attention because of the findings 6, 8 and 10.

The clustering script was set to maximize the number of produced clusters. After running the script, 9 significant clusters were obtained for both files. The O-D pairs grouped in each cluster were showing different modal split pattern. In each cluster, the share of some mode or modes has increased in comparison with the file before clustering.

From both U1 files, clusters with increased share of transport mode Train were selected for comparison. The quality of the resulting Train clusters was verified using the Geographic Information System [7]. Each O-D pair from the cluster is represented by the municipality of origin, where the mode choice decision is made.

The mode Train was selected, because its fixed infrastructure makes it the least flexible mode. It is then easier to visualize where the railway infrastructure is present and which municipality has easier access to the Train service. It then effects the travelers in selecting the mode.

The municipalities of origin from the Train cluster of O-D pairs from the nonfiltered file U1 are depicted in the Figure 2. The municipalities in this figure are commonly located in some distance from the railway network, where the access to railway service is limited. This points to potential problem with this cluster as well as the originating file, which did not go through filtering process.

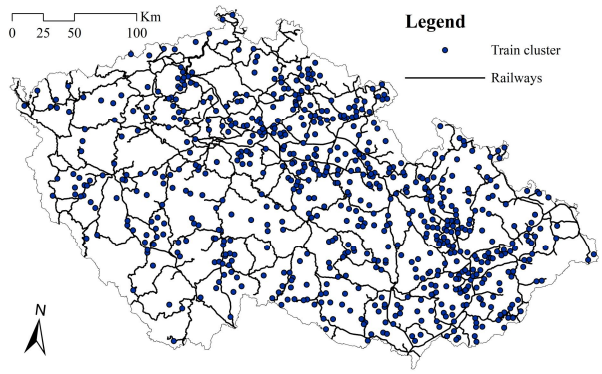

Figure 2: Municipalities of origin from the Train cluster of O-D pairs made from the file U1, originating from "Nonfiltered" data set

The Figure 3 is showing the municipalities of origin from the Train cluster of O-D pairs from the file U1, which was produced from "P_{ij} *comp* filtered" data set. The municipalities in this figure are mostly located in proximity of the railway network, only exceptionally in some distance from the railway network, where the access to railway service is limited. This shows an increase in the quality of produced Train cluster as well as the data set, which was filtered using the P_{ij} *comp* method.

It is necessary to obtain good quality clusters with increased share of the transport modes for the future research. The dependency of usage of any transport mode is best to be studied in a cluster of O-D pairs (municipalities), which are showing the increased share of such mode, thus in clusters, where motivation of the travelers to choose such mode was stimulated by some researchable influence.

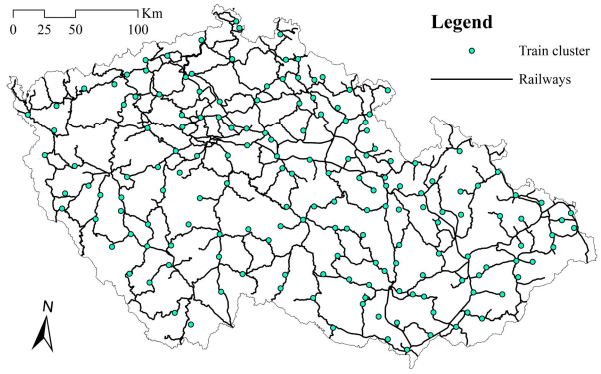

Figure 3: Municipalities of origin from the Train cluster of O-D pairs made from the file U1, originating from "P_{ij} *comp* filtered" data set

5. CONCLUSION

From the presented filtration methods of O-D pairs, the P_{ij} *comp* method was the best performing, because removing nearly 98 % of O-D pairs with one traveler, which are causing a distortion of the modal split statistics of the whole data set, which could negatively affect the reliability of further mode choice research.

The benefit of the filtering was verified by plotting and comparing the clusters of O-D pairs, which were produced from nonfiltered and from filtered data set. The Train cluster from filtered data set is showing higher adherence of O-D pairs to the railway lines.

The presented segregation of the data set into groups of O-D pairs according to 8 direction characteristics has shown the importance of distinguishing these direction groups in every data set and studying them separately.

The following findings have supported the developed filtering method and they are preferable for further research:

A. The share of transport mode Train is decreasing with decreasing number of O-D pairs with one traveler, implying the train service in the Czech Republic is offered on O-D pairs with small transport demand from travelers commuting to work or school on the local level.

B. With increasing average number of travelers on O-D pairs the average number of modes of transport used on each O-D pair also increases. This implies, the increasing transport demand is generally not satisfied by increased capacity of modes but by competition of increased number of modes.

REFERENCES

1. Czech Statistical Office. URL. https://www.czso.cz/csu/sldb/home. Last checked on 3-10-2020.
2. S. El Mendili, Y. El Bouzekri El Idrissi, N. Hmina, **Big Data Processing Platform on Intelligent Transportation Systems**, International Journal of Advanced Trends in Computer Science and Engineering, Vol. 8, No. 4, 2019.
3. J. Cekal. **The South Bohemian region: a regional-geographical analysis of spatial mobility of the population**, Ph.D. dissertation, Masaryk University, 2006.
4. A. Afonso and A. Venancio. **The relevance of commuting zones for regional spending efficiency**, *Applied economics*, Vol. 48, No. 10, pp. 865–877, 2016.
5. I. Nagy, E. Suzdaleva, T. Mlynarova, **Stochastic Systems and Applications**, Prague: Czech Technical University in Prague Faculty of Transportation Sciences, 2012.
6. M. J. Christ, R. N. Permana Tri, W. Chandra, T. Mauritsius, **Lending Club Default Prediction using Naïve Bayes and Decision Tree**, International Journal of Advanced Trends in Computer Science and Engineering, Vol. 8, No. 5, 2019. https://doi.org/10.30534/ijatcse/2019/99852019
7. P. Satra and J. Carsky, **Verification of Bayesian Clustering in Travel Behaviour Research – First Step to Macroanalysis of Travel Behaviour**, *IOP Conference Series: Earth and Environmental Science*, Vol. 140, conference 1, 2018.