Extracting railway passenger demand patterns from origin-destination data

Renate van der Knaap
Department of Transport and Planning, Delft University of Technology

It is well known that the passenger demand for railway services fluctuates over the day. During the peak hours there is a high demand for transport, while in the middle of the day or in the evening this demand is much lower. The demand also fluctuates over the week: not all (week)days have the same demand. Due to the COVID-19 pandemic, these differences are expected to increase. Several studies show that more people will partly work from home after the pandemic (see e.g. Ton, et al. (2022)).

Nevertheless, many European countries including the Netherlands have cyclic railway timetables. These timetables provide regular interval services throughout the day and are therefore also easy to remember for passengers. One of the disadvantages however is the fact that the cyclic timetables are usually based on the peak hour demand, and are therefore not tailored for other periods outside the peak hours. This is both in terms of volume, as well as in terms of the structure of the demand. People who travel outside of the peak hours are likely to have different travel motives and hence destinations.

Better matching the train services (line plan and timetable) to the travel demand can have multiple benefits for both the passengers as well as the railway undertaking (RU). Passengers will have a better travel experience with faster trips and less transfers. Furthermore, the RU can save money on energy, rolling stock and personnel by only operating those trains for which there is demand.

Before being able to better match their train services to the demand, the RU first needs to have a good overview of how the demand patterns change throughout the day and week. Therefore, there are two questions that need to be answered:

- What homogeneous periods do we see during the day?
- What characteristics does the demand have in different homogeneous periods?

Note that the answer to these questions may vary when looking at different days. We would like to answer these questions using revealed preference data in the form of origin-destination (OD) data, which is usually readily available at RUs.

Data & method

For this research, we use realized origin-destination (OD) data of 2019 from the Netherlands Railways (NS). NS is the principal passenger railway undertaking in the Netherlands and operates trains to and from 253 stations in the Netherlands. In this study, we aim to look at passenger demand patterns for a normal workday. Therefore, all weekend days, (school) holidays, and days with major disturbances (according to NS' Annual Statement 2019 (NS, 2019)) are taken out of the dataset. From the data of the remaining days, the median is taken to create an OD matrix for every half hour between 6:00AM and 23:59PM of every workday. These matrices are the input data for the data analysis.

Hierarchical clustering is used to determine homogeneous periods during the day, by clustering the OD matrices of a day together. This is done in two different ways. First we cluster the normalized OD matrices following the method described by Ji, Mishalani, McCord, & Goel (2011). They use the squared Hellinger Distance as the distance between two normalized OD matrices and complete linkage to determine the distance between two clusters. Furthermore, they require that the formed clusters must consist of OD matrices that are contiguous in time. By only looking at the normalized OD matrices, where each cell is divided by the total sum of the matrix, only the structure of the demand is taken into account. The changes in volume during the day are completely cancelled out. However, besides the demand structure, the volume of the demand is also important when constructing train services such as line plans and timetables. Therefore, for the second method of clustering we use the non-normalized OD matrices and the Euclidean distance as the distance between two matrices. Similar as in the previous step, we use complete linkage for determining the distance between two clusters and require that the formed clusters must be contiguous in time. Both steps are performed per workday as the demand patterns per workday may be different. Furthermore, the quality of the clusters is assessed using the Silhouette coefficient (Rousseeuw, 1987), which looks at both the compactness of the clusters as well as the closeness to other clusters.

Results

The results of the hierarchical clustering method can be visualized in a dendrogram plot, which shows the grouping of the (normalized) OD matrices. The dendrograms of the normalized and non-normalized OD matrices of the median Tuesday in 2019 are given in Figures 1 and 2, respectively. The dendrograms should be read as follows. The horizontal axis displays the time, where for example 600 denotes the OD matrix containing all trips that start between 6:00AM and 6:29AM and 1430 denotes the OD matrix containing all trips that start between 2:30PM and 2:59PM. All matrices start in separate clusters (as can be seen at the bottom of the dendrogram) and in every step two matrices are clustered together until all matrices are in a single cluster (as can be seen at the top of the dendrogram). The combining of two clusters is denoted in the dendrogram by a horizontal line connecting the two clusters. The vertical axis denotes the distance between two clusters at the moment they are combined.

When looking at the clustering results of the normalized OD matrices (see Figure 1 for the results of the median Tuesday), the clearest pattern that emerges is that each workday is divided into two parts: the morning hours on the one side and the afternoon and evening hours on the other side. The transition from morning to afternoon is usually at 12:00PM, but is earlier on Wednesday (at 10:30AM) and later on Tuesday (at 2:00PM). The result with only two clusters is also the result that gets the highest Silhouette coefficient, which tells us that this result is the most appropriate. We hypothesize that the division between morning and afternoon/evening mainly comes from the fact

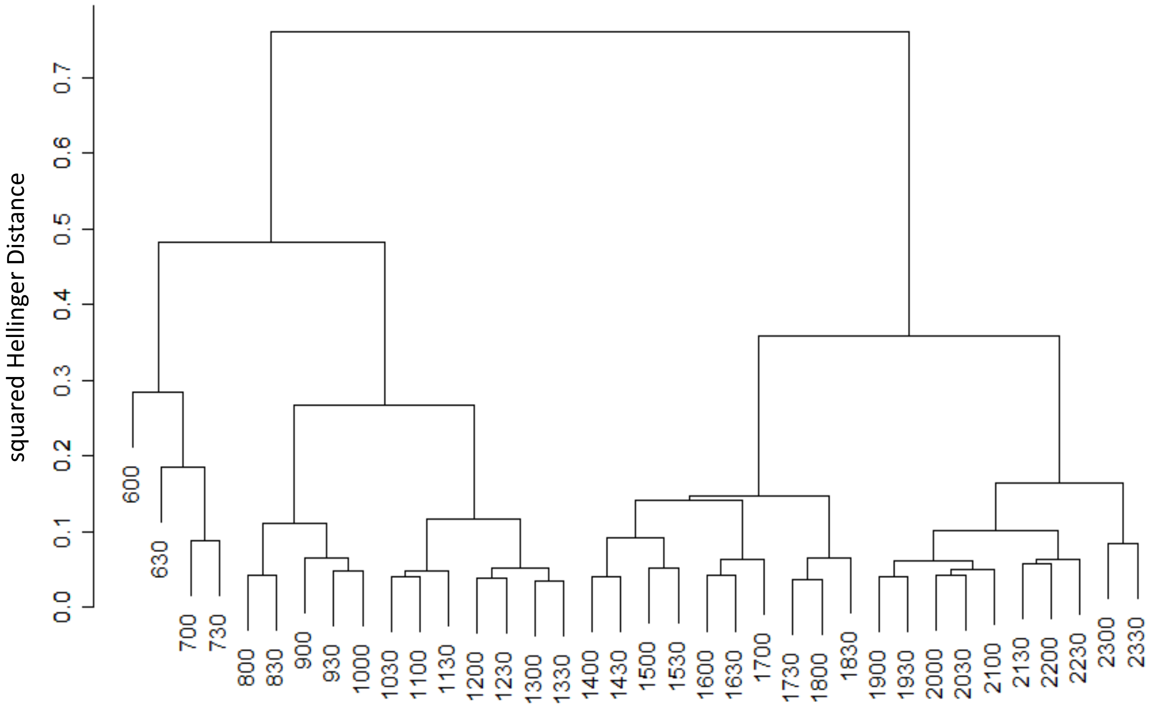

Figure 1 Dendrogram of cluster result using normalized data of the median Tuesday and the squared Hellinger Distance

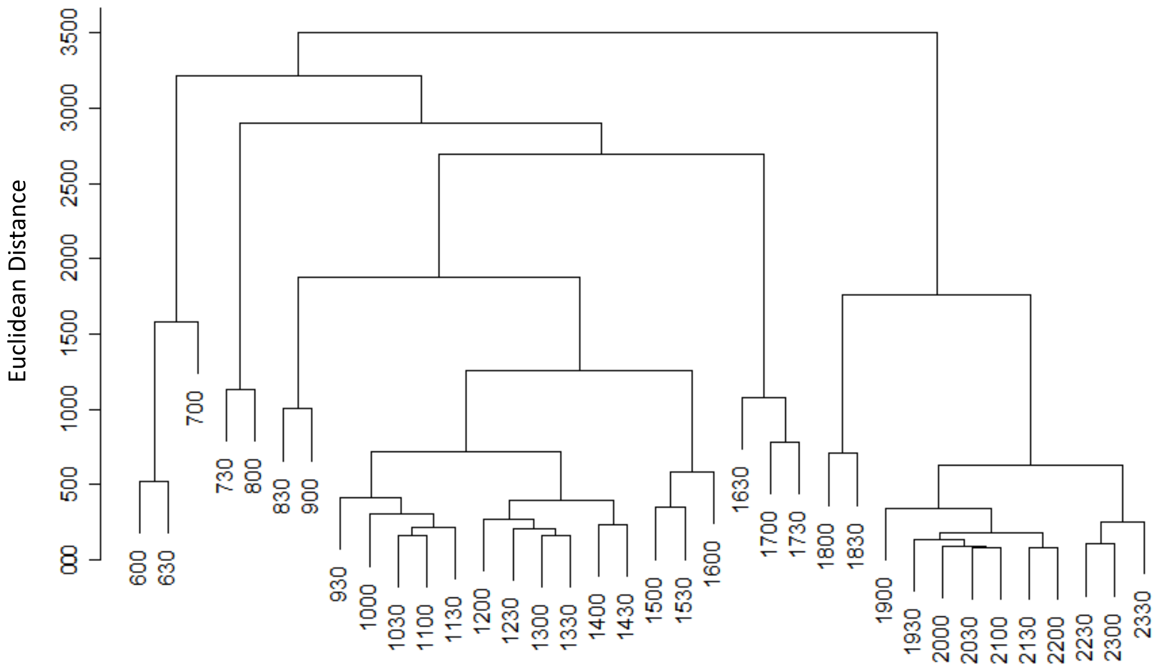

Figure 2 Dendrogram of cluster result using non-normalized data of the median Tuesday and the Euclidean distance

that people usually go somewhere in the morning and return back to where they came from in the afternoon or evening of the same day. We test this hypothesis by comparing the distance between the morning OD matrix and the afternoon/evening OD matrix to the distance between the morning OD matrix and the transpose of the afternoon/evening OD matrix. The results show that the morning matrix is between 5 and 15 times closer to the transposed matrix than to the normal OD matrix of the afternoon/evening. Hence, this supports the hypothesis.

When we take the volume into account by looking at the non-normalized OD matrices (see Figure 2 for the results of the median Tuesday), the morning versus afternoon/evening pattern does not emerge. Also the recommended number of clusters (based on the silhouette coefficient) is much higher: between 9 and 10 depending on the day. In the results, we see larger clusters during the day (from 9:30AM until approximately 3:00PM) and in the evening (from 7:00/7:30PM until 12:00AM). Around and during the peak hours we see many small clusters with between one and three matrices per cluster.

In the next steps of this research we aim to determine which of these two clustering results provides the most useful information when determining homogeneous periods for adjusting the railway services. We propose to do this by looking at the average trip length and changes in the relative popularity of the stations within and between the identified periods. When appropriate clusters are made, we expect small differences in average trip length and relative popularity of stations within the identified periods and large differences between periods.

Acknowledgements

This research was supported by NS, the largest passenger railway operator in the Netherlands.

References

Ji, Y., Mishalani, R. G., McCord, M. R., & Goel, P. K. (2011). Identifying homogeneous periods in bus route origin-destination passenger flow patterns from automatic passenger counter data. *Transportation Research Record*(2216), 42-50.

NS. (2019). *NS Annual Report 2019.* Retrieved June 14, 2022, from https://2019.nsjaarverslag.nl/FbContent.ashx/pub_1000/downloads/v200416100752/NS_annualreport_2019.pdf

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65.

Ton, D., Arendsen, K., de Bruyn, M., Severens, V., van Hagen, M., van Oort, N., & Duives, D. (2022). Teleworking during COVID-19 in the Netherlands: Understanding behaviour, attitudes, and future intentions of train travellers. *Transportation Research Part A: Policy and Practice, 159*, 55-73.