

Problem description

Metro lines in big cities are frequently used for commuting purposes or by traveling tourists who are sightseeing. This is relevant especially for cities like Berlin, which has 3.7 million inhabitants and in addition to that a very high number of tourists. In 2019, Berlin was visited by approximately 13.7 million guests who spent 34.1 million guest-nights. The task for city officials is to run and maintain the transportation service in a way that there are no congestions in the transport capacity but also keeping the system cost-efficient.

Intuitively one might think that commuting traffic is highest on workdays, especially on mornings and evenings because that are the times when most workers travel to their working place or home from there. Also, public transport numbers for tourist purposes may be higher during the weekend than on weekdays due to day-travelers or short weekend vacations. So, it would be interesting to know where in the city the transport is used for what main purposes. Therefore, we might be able to cluster the stations of the public transportation net to learn more about where to expect what kind of traffic and draw conclusions on the expected load of the stations or lines.

Another use of this analysis might be found in the pricing/ticket system for public transport. If a cluster is found in the stations which is geographically compact and for example is interpreted as a mainly touristic cluster, city marketing might think of a special tourist ticket for this part of the city. With these tickets tourists maybe could travel within the designated cluster for a whole weekend (short trip) instead of buying all-day tickets for the whole city for every day of their stay.

Main problem of this analysis is to cluster the stations of a public transportation net to gain insights on the characteristics of the usage of the transportation net.