**1, Introduction**

**1,1 Intro**

The Hungarian capital, Budapest, has a growing population of close to 2 million people, who require an extensive public transport network. The purpose if this paper is to analyse the operation of Budapest’s public transport by exploring the openly accessible GTFS (General Transit Feed Specification) data format. The analysis will further investigate public bus routes as a key part of the urban transportation mix. Discovering key similarities among bus routes will help identify clusters of bus services with different characteristics across the city. Improving our understanding of the Budapest bus network could enhance the efficiency of schedules to better serve the population’s transportation requirements as well as reduce congestions, optimise bus lane arrangements, and decrease air pollution.

Budapest’s residents have significantly different transportation needs during the weekdays and weekends. Therefore, paper will differentiate between a regular weekday’s and weekend’s transportation schedule to provide comparison between the two contrasting operation mode.

**1,2 Technical requirements**

For the analysis, the following Python libraries will be used. The aim was to use standard libraries to showcase how regular Pandas and Geopandas can be sufficient to work with GTFS format. For the clustering, Sklearn machine learning library is imported. To create the data visualisations Seaborn, Contextily and Folium packages are utilised.

**Literature Review**

Public transportation is a key component of urban development and planning. In most major cities, significant public spending is allocated towards enhancing transportation networks, as they impact the socio-economic development of areas, through providing (or limiting) access to education, employment, healthcare, and leisure (Deka, 2004). Access to various forms of transportation has long been seen as a source of economic. Today, numerous studies suggest that lower income households are drawn towards the more impoverished, inner-city areas due to cheaper access to transportation than in the suburbs (LeRoy and Sonstelie, 1983; Glaeser et al., 2008). In addition, transportation networks can have substantial environmental impacts. As the global rate of automobiles per capita continues to rise, countries and particularly large cities can expect a continued rise in traffic congestions and related pollution (via emissions) unless significant action is taken (May, 2013). As such, enhancement opportunities in urban transportation planning are highly sought-after, particularly in the developed world, where countries have means for funding such projects.

There are various modes of transport that define a particular cities urban transportation structure. In most cities cars and public busses tend to be the primary form of transportation, however many cities have popular alternatives, such as the underground in London (Guo and Wilson, 2011), or cycling network in Amsterdam (Buehler and Pucher, 2009). For Budapest in particular, the public transportation mix is fairly complex, with numerous metro lines, buses, trams, trolleybuses, and overground systems collectively operating to sustain the population’s transportation needs. The following essay will focus on the 229 bus services with 710 routes across Budapest and their characteristics.

This analysis will only use GTFS feed data. GTFS is becoming more and more accessible across the world. The largest data store can be found on transitfeeds.com, with currently 677 location’s GTFS feed provided by 1327 provider. Analysing and understanding these feeds could enhance our understanding the transit of cities and enhance public transport network schedules.

**Research Question:**

How can we use clustering methods on Budapest's public bus network schedule to identify improvement solutions on the city's transportation system?

**3. Methodology**

The accessed GTFS data contains an entire month (November 2020) of public transport schedules. This dataset was selected for the purpose of excluding Covid-19 related anomalies in the transportation schedule, as the schedules were published in October 2020, when there were no Covid related restrictions in Budapest. According to the operator company the schedule is defined monthly, therefore generally the same days of the weeks have essentially the same routes and timetables.

For computer memory efficiency reasons, the busiest weekend (Saturday or Sunday) and the busiest weekday (Monday, Tuesday, Wednesday, Thursday, Friday) in the month will be selected as samples for the analysis and clustering.

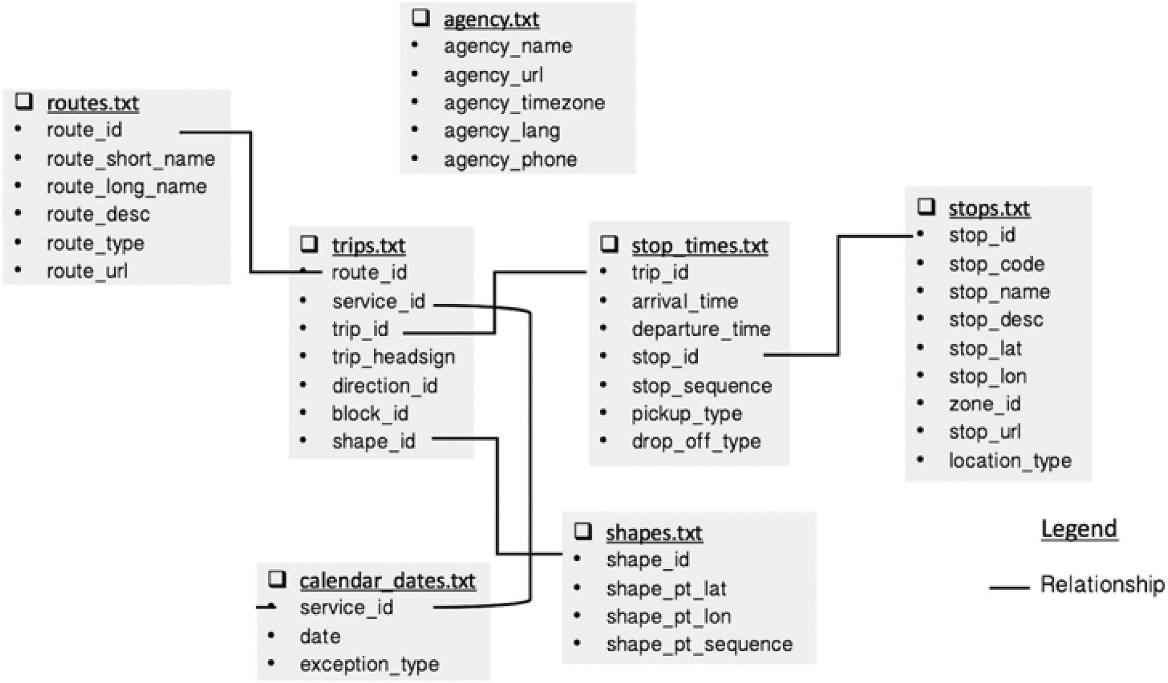
This workpaper will provide step-by-step instructions from reading the raw data, cleaning as well as preparations for the explanatory analysis and bus route clustering. The explanatory analysis will showcase the information that can be gathered from the rich GTFS file format. In addition to highlighting the general usefulness of the GTFS public transport data, we will also gain a better understanding of Budapest public transportation system. Finally, this workpaper was specifically designed to facilitate further research, outside of the scope of this essay, for example by using another month of Budapest’s GTFS data or through using a different city’s GTFS dataset.

This analysis will be using GTFS Static and not the real-time extensions of GTFS. GTFS contains multiple txt files which has complex relationships, see chart below. In the case of Budapest it contains the following files:

* **agency.txt** – It contains information about the service agency (BKK) of the feed, and time zone of the city where the transit operates and contact information.
* **stops.txt** –It contains the transit station/stop names, stop ID, and geo-locations (latitude, longitude).
* **trips.txt** –It contains the directions headed of each vehicle movement. Each trip has a service ID which specifies the days it operates on.
* **route.txt** –It contains the information on how trips are grouped into single services. Vehicle type of transportation, name/number of the route,
* **stop\_times.txt** –It contains the information on the arrival and departure time of each transit station/stop, as well as station sequence and corresponding trip IDs. This is the largest file in the folder
* **calendar\_dates.txt** –It contains the service id and their related dates
* **shapes.txt** –It contains geographical polylines representing the routes that a transit vehicle takes.

(Kunama et al., 2017)

The relationship among the tables is illustrated below.



(Modified from source: Prommaharaj et al., 2020)

In order to execute the clustering of bus routes, K-means classification method will be utilised. K-means is an unsupervised machine learning algorithm, therefore by definition it seeks to uncover hidden characteristics of the data fed into it. (Allahyari et al., 2017). No training dataset is required, and the algorithm can be used on numerical and categorical data attributes. The following numerical attributes will be calculated in order to create the clusters:

1. Average speed of the route/shape – defined in km/h
2. Stop Density of route/shape - stops per kilometre across the route
3. Trip activity level - number of trips have been taken on the route in a day

For defining the K – number of clusters for our classifications, we’ll be using the Elbow Method. The Elbow Method is a popular test performed to identify the appropriate number of clusters that the data requires.

4, Data Processing

4,1 Data Gathering

4,2 Data Cleaning, Preparation

5, Analysis

5,1 Descriptive Stats

5,1,1 Vehicle types

5,1,2 Traffic of stops

5,1,3 Visualise a daily changes of the public transport traffic

Weekday:

The following couple of figures will explore the four identified clusters in the bus routes.

* **Cluster 0:** the group of routes with the lowest average speed and the highest stop density represents the bus routes which connect the core transportation hubs of the inner-city rather than provide quick commuting solutions into/from the city centre. The map below shows that this cluster covers most of the city with a high density in Inner Budapest and certain north suburb areas.
* **Cluster 1:** the category that contains the most frequent bus schedules (count\_trips) with 95.7 average trips per day. It transits slightly faster than Cluster 0 and has fewer stops on average than Cluster 0. Cluster 1 consists of the essential bus routes for the city’s population crossing the Danube bridges connecting the two parts of the city Buda and Pest.
* **Cluster 2:** this group has the higher speed, occasional buses with only 9.7 trips per day. A lot of these routes are the night bus routes, with low stop density and higher speed due to the lack of traffic at night. This Cluster covers most of the main squares and streets of the city.
* **Cluster 3:** This is the cluster of fastest bus routes, providing longer-distance commutes to-and from the outer-cities/suburbs. These busses are able to travel at higher speed due to the low stop density. Based on the map, these routes only reach the edges of the city-centres, where commuters are likely to change transportation lines to reach their final destinations. These routes are designed to avoid congested roads (such as bridges) and take fewer stops between their destinations. For example, this cluster contains bus 200E which connects the city centre with Budapest Liszt Ferenc international airport.

Weekend clusters resulted similar outcome

Similar characteristics of the cluster categories can be observed for the weekend travel schedules. Slight differences are the distribution of routes in the 4 clusters as there are far fewer ‘commuter’ fast routes connecting the suburbs to the city. In addition, all weekend clusters have higher average speed attributes within each cluster, which can be due to the decreased automobile traffic.

6. Discussion and Conclusion

Our analysis has focused on exploring the different public transportation modes of Budapest in order to be utilised for city-planning decision making. There are a variety of ways in which the above analysis and defined clusters can be interpreted. As a part of the exploratory analysis, we have revealed the ‘pulse’ of the city, visualising the peak hours of an average weekday. These visualisations highlight which parts of the city are more active throughout the day, which can be used to facilitate congestion management analysis. In addition, the comparison between weekday and weekend data has revealed the difference in distribution of public transport activity between workdays and non-workdays, highlighting the essential differences in the public’s use of these transport systems at various points in the week.

The bus route clustering can be used to identify routes with similar characteristics in order to support transportation management decision making across multiple routes. For example, the bus routes in Cluster0-weekday and Cluster2-weekend contain the routes which are mostly concentrated in the city with high-stop densities. As such, the routes in the cluster could be used to identify bus routes which are most suited to be converted to e-busses, in order to reduce inner-city pollution levels. In addition, the clusters can also be used to guide decisions regarding the planning of bus lanes, where priority could be given to bus routes in Cluster1-weekday & Cluster3-weekday. These routes have fewer stops and are aimed at connecting longer distance destinations, where the speed between the infrequent stops is a key factor, compared to routes with more frequent stops or night-bus routes (Cluster2-weekday).

In conclusion, the analysis can be utilised to support a variety of urban system enhancement projects in addition to the suggestions outlined above. The research could be extended to incorporate different months of transportation data for Budapest, for example to be used in a seasonal comparison of public transportation activity. Finally, with slight modifications, the workpaper could also be adjusted to explore the GTFS data of different global cities and provide valuable insight into public transportation activities around the world.

References

Buehler, R. and Pucher, J., 2009. Cycling to sustainability in Amsterdam.

Deka, D., 2004. Social and environmental justice issues in urban transportation.

Glaeser, E.L., Kahn, M.E. and Rappaport, J., 2008. Why do the poor live in cities? The role of public transportation. *Journal of urban Economics*, *63*(1), pp.1-24.

Guo, Z. and Wilson, N.H., 2011. Assessing the cost of transfer inconvenience in public transport systems: A case study of the London Underground. *Transportation Research Part A: Policy and Practice*, *45*(2), pp.91-104.

Kunama, N., Worapan, M., Phithakkitnukoon, S., Demissie, M., 2017. GTFS-Viz: tool for preprocessing and visualizing GTFS data, in: Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers, UbiComp ’17. Association for Computing Machinery, New York, NY, USA, pp. 388–396. <https://doi.org/10.1145/3123024.3124415>

LeRoy S., Sonstelie J., Paradise lost and regained: Transportation innovation, income and residential location, Jour-nal of Urban Economics 13 (1983) 67–89.

May, A.D., 2013. Urban transport and sustainability: The key challenges. *International journal of sustainable transportation*, *7*(3), pp.170-185.

Prommaharaj, P., Phithakkitnukoon, S., Demissie, M.G., Kattan, L., Ratti, C., 2020. Visualizing public transit system operation with GTFS data: A case study of Calgary, Canada. Heliyon 6, e03729. <https://doi.org/10.1016/j.heliyon.2020.e03729>

Not used

Kadir, R.A., Shima, Y., Sulaiman, R. and Ali, F., 2018, March. Clustering of public transport operation using k-means. In *2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA)* (pp. 427-432). IEEE.