Public transport flow networks in the city of Amsterdam

Sophie Ensing

March 15, 2019

1 Personal details

My email mailto:sophie_ensing@hotmail.com

My supervisors email mailto:c.amrit@uva.nl

The wiki on my github account github.com/sophieensing/thesis

2 Introduction

The municipality of Amsterdam has a large dataset from the public transport company GVB. This data set contains traveller data from all public transport lines in the city of Amsterdam. In a big city there are often malfunctions, delays or constructions that could lead to problems in public transport. In these situations travellers will often have to take different routes, which can result in crowdedness in different places than usual.

This research is aimed at visualising traveller flows in the city of Amsterdam in the form of flow networks. With these networks it is possible to analyse current flows through public transport in the city, and it also gives the possibility to see the effects of possible malfunctions. The main research question is defined in section 3.1.

3 Research question

3.1 What are the effects of public transport malfunctions on traveller flows in Amsterdam?

In order to answer the main research question, two sub-questions will have to be answered. The first sub-question will focus on the visualisation of traveller flows and the second sub-question is aimed at predicting alternative route choices.

3.1.1 How can traveller flows be visualised in a flow network?

To assess the effect of any changes in public transport, it's essential to visualise the general traveller flows correctly first as a baseline. From this baseline, the influence of malfunctions in public transport can be shown in the flow network.

When creating the baseline, it is important to also take the different seasons and times of the week into account. There are probably quite some differences between winter/summer and weekdays/weekend days. These visualisations have to be evaluated to ensure they are a good representation of the actual data.

3.1.2 How can alternative route choices be predicted?

When a certain line fails, the flows through that line will have to go elsewhere. After the establishment of a baseline, the models can be used to show the (predicted) alternative routes people will take. In some cases travellers might take the shortest alternative route, but for some people this might not be the case. This sub-question will tackle the problem of predicting the alternative route choices that people will take. When this is done the results and alternatives flows that will emerge can be discussed and evaluated to answer the main research question.

4 Related Literature

A lot of research has been done on route choice behaviour and traveller preferences. Since there are huge differences in transport networks in different types of cities it's hard to take these findings and apply them to the city of Amsterdam. The comfort, prices and speed of certain types of transport vary a lot across different cities and even more across countries. There are some features that appear more often. Research shows that time in transport, cost, personal preference and waiting time are important factors when making these choices [1]. Earlier research often uses the shortest path approach, because it is assumed this is also the fastest and cheapest route [3]. A problem with this approach arises when there are common lines. If there are two lines between a certain origin-destination pair, other factors also play a part. Waiting time becomes more important here for example.

Visualising the transport flows and predicting route choices are the main focus of this research, but the results also have to be evaluated. The predictions can be evaluated with test data, but the network itself can be analysed and evaluated as well. Previous research shows methods to analyse the flow distribution and other properties of the network [2]. Both methods of evaluation should be applied.

5 Methodology

5.1 Data processing

The GVB data is quite organised, with traveller streams per hour from all origin and destination pairs. For this project a full year of GVB data will be available. Figure 1 shows an example of travel streams from Weesperplein to Central Station per hour. The stations have a unique code (see third column: *VertrekHalteCode*) from which it's possible to derive whether it's a bus, train, metro or tram station.

Datum	UurgroepOmschrijving (van vertrek)	VertrekHalteCode	VertrekHalteNaam	AankomstHalteCode	AankomstHalteNaam	AantalRitten
2018-12-24	06:00 - 06:59	WPP	Weesperplein	CS	Centraal Station	22.0
2018-12-24	07:00 - 07:59	WPP	Weesperplein	CS	Centraal Station	43.0
2018-12-24	08:00 - 08:59	WPP	Weesperplein	CS	Centraal Station	48.0
2018-12-24	09:00 - 09:59	WPP	Weesperplein	CS	Centraal Station	82.0
2018-12-24	10:00 - 10:59	WPP	Weesperplein	CS	Centraal Station	95.0

Figure 1: Data sample

There are some side notes to the data. For streams smaller than 10 travellers per hour the stations are set to *Other* because of privacy. The second exception is streams with the same origin and destination. A lot of stations can only be entered after checking in with a public transport chip card. Some people want to access shops within the station or use the route through the station because it's faster. These people can just check in and check out without any costs to do so, but since these are not transport flows they will be removed as well.

This data needs to be matched to malfunctions from the past year. This way the data can be split into regular travel behaviour and travel behaviour that is a reaction to malfunctions. Support vector machines and random forests have been researched in relation to route choice behaviour and are considered suitable options for prediction [1]. Both methods will be considered for this research.

5.2 Visualising the flow network

The python package networkx will be used to visualise all traveller streams in the form of a flow network. The nodes represent the stations and the edges represent the bus, train, metro or tram lines. The flow over these edges are the traveller streams per hour for the particular lines. The visualisation should be made in an interactive way so a certain line can be selected as a malfunctioning line. The flows through the selected edge have to go elsewhere which should be visualised as a reaction to the malfunction. Figure 2 shows this in a very simplified manner. If the line from Central Station to Science Park Station is malfunctioning, this flow will go through other lines. In this case they are added to the other possible route via Amstel Station. In reality there are more options and factors to consider of course.

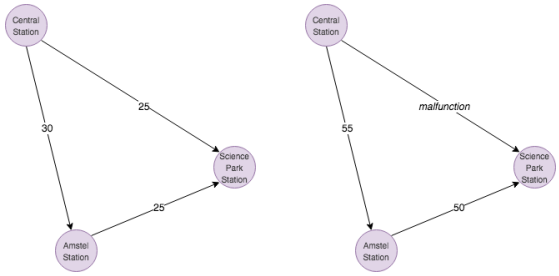

Figure 2: Flow network

5.3 Predicting route choices

The first step to predict route choices is to talk to domain experts at the municipality and GVB. The domain knowledge is very valuable to improve the research and explainability of the model that will be created. To predict the route choices, the second step is to use data on previous malfunctions. This data has to be split up into train en test data for the training and evaluation of the model.

As mentioned in section 4, there are several factors that play a part in route choice behaviour. The possibility of adding more data sources to add more features to the model, like waiting time and costs for example, should be considered as well in this section.

6 Risk assessment

6.1 Malfunction data

While developing the model the data on the actual malfunctions pose a risk to this research. It might not be extensive enough to make correct predictions of route choices, because there are many factors that play a role here. Another issue to keep in mind that the data is probably very skewed towards certain lines in the network and the data can not be easily generalised over all lines. In some places there might be only one alternative, but in other places there might be many more. In these two situations the amount of options and external factors and preferences that play a role make it very hard to generalise certain results. Therefore it might be necessary to alter the scope of the network based on available malfunction data.

If it turns out that the data on malfunctions is not extensive enough, alternative methods can be used for prediction. With a basis of domain knowledge and extensive literature research certain features can be engineered as well. For example: some literature is focused on fastest path choices in public transport and if it is possible to back this up with the GVB dataset this can be used as a predictive feature in the model. The evaluation and analysis of this approach is a bit more difficult because in this case there is no test data. Evaluation should be more focused on comparing different approaches to predict route choices.

7 Project plan

Week (date)	**Task**
1 (1/04 - 7/04)	Related work and literature study.
2 (8/04 - 14/04)	Pre-processing data and create flow network
3 (15/04 - 21/04)	Pre-processing data and create flow network
4 (22/04 - 28/04)	Evaluate flow network
5 (29/04 - 05/05)	Predict routes in flow network
6 (06/05 - 12/05)	Predict routes in flow network
7 (13/05 - 19/05)	Predict routes in flow network
8 (20/05 - 26/05)	Analysis and evaluation of results
9 (27/05 - 02/06)	Analysis and evaluation of results
10 (03/06 - 9/06)	Write thesis and hand in complete draft
11 (10/06 - 16/06)	Write thesis and work on feedback
12 (17/06 - 23/06)	Finish thesis and prepare for defence
13 (24/06 - 30/06)	Defend thesis

References

[1] Xinjun Lai, Hui Fu, Jun Li, and Zhiren Sha. Understanding drivers' route choice behaviours in the urban network with machine learning models. *IET Intelligent Transport Systems*, 2018.

[2] Keumsook Lee, Woo-Sung Jung, Jong Soo Park, and MY Choi. Statistical analysis of the metropolitan seoul subway system: Network structure and passenger flows. *Physica A: Statistical Mechanics and its Applications*, 387(24):6231–6234, 2008.

[3] Yulin Liu, Jonathan Bunker, and Luis Ferreira. Transit users' route-choice modelling in transit assignment: A review. *Transport Reviews*, 30(6):753–769, 2010.