Spatial-temporal analysis of the public transportation network of Zürich in R

Manuel Bär
MSc GIScience, University of Zürich
Winterthurerstrasse 190
CH-8057 Zürich
+41 44 635 51 11
manuel.baer@uzh.ch

Alexandra Kessler
Department of Geography, University of Zurich
Winterthurerstrasse 190
CH-8057 Zurich
+41 44 635 51 11
alexandra.kessler@geo.uzh.ch

ABSTRACT

Public transportation is of growing importance in Switzerland. The steady increase of commuters using this mode of transportation has an impact on the infrastructure and characteristics of the public transportation network. Spatiotemporal problems arise which have a negative influence on the client satisfaction. In this paper we analyzed various aspects of the public transportation system in Zurich, including the daily mean delay times and the detection of problematic nodes. We analyzed a large dataset from OpenData Zürich using R by creating various maps, diagrams and a video. Two different delay propagation types (line following and network spread) are identified and discussed. We detected large intra-daily fluctuations in delay times of public transportation services and large inter-daily similarities. A significant difference in the delay times of public transportation services between weekdays and weekends is noticed. Various societal structure characteristics are identified using the delay times as a proxy and event detection is discussed. Problematic nodes are found in Dietikon, Seebach and the Main Train Station. We suggest further research be done regarding delay propagation through public transport networks and the detection of network spread epicentres including the affected lines.

Categories and Subject Descriptors

D.3.0 [**Programming Languages**]: General – *R Script*; E.1 [**Data Structure**]: Graphs and networks, Tables – *Public Transportation Network, CSV*; G.3 [**Probability and Statistics**] Time series analysis – *Spatial-temporal analysis, Delay times*; H.5.1 [**Multimedia Information Systems**] Animations – *Mean delay video output*; I.1.4 [**Applications**]: *R, Excel*;

General Terms

Public Transportation Network Analysis, Advanced Spatial Algorithms, R scripting, Spatial temporal analysis, Geographic Information Science

Keywords

Network, PTN, Public Transportation Network, Transport, Zurich, Analysis

University of Zürich, January 15, 2016, Zürich, Switzerland.

1. INTRODUCTION

In an increasingly interconnected and urbanised world, modes of transportation have become significantly more important. In particular commuting services made possible by various public transport companies and the performance thereof have gained increasing interest [1]. It has been argued that the mode of transport can have major impacts on various dimensions of the quality of urban life, including "emission of pollutant and greenhouses gases, road congestion, accidents, and energy consumption" [2]. Various papers [3]–[5] have discussed possible push and pull factors regarding the use of public transport compared to the use of private vehicles. Quality and efficiency play a major role in the attractiveness of public transportation services [1], [3], [6] and it has been argued [6]–[8] that availability, accessibility, time, trustworthiness, frequency, maximum load, comfort and information and support facilities are key proxies to determine the public transportation service general performance. In a study conducted by [9], line reliability, bus punctuality and service frequency where among the aspects relevant to user satisfaction. With an increasing number of users for a given infrastructure, spatial or temporal problems will arise, decreasing user satisfaction and acting as a push factor to change the mode of transport.

The Federal Department of Statistics of Switzerland recorded a total of 4'258'557 commuters in 2013, accounting for 52.3% of the inhabitants. The absolute number of commuting citizens has also seen an increase in the period from 2010 to 2013 from 4'095'042 to 4'258'557 commuters, with an ongoing average positive trend. This highlights the importance of the public transport sector in Switzerland. There has also been a trend to further commuting distances. In 1990 only 12% inter-cantonal commuters were recorded, which has grown to 20% by the year 2013. Interregional commuters moving within the canton they are living in has not seen much change and local commuters commuting within the municipality of living has shrunk from 41% to 30%. This can be attributed to the "changing nature of society and lifestyle patterns which generate diversified travel needs" [3] which as [3] argue must "be considered in decision making concerning transport".

In this paper we analyse, discuss and visualise various aspects of the public transportation network of Zurich. The ZVV system consists of all trams, busses and local trains in the canton of Zurich. The available data consists of all the busses and trams in Zurich. Special attention is given to the detection of potential problematic nodes or routes in the public transportation network of Zurich, to be able to make empirically verified statements of possible improvements or circumnavigation of mentioned problematic nodes.

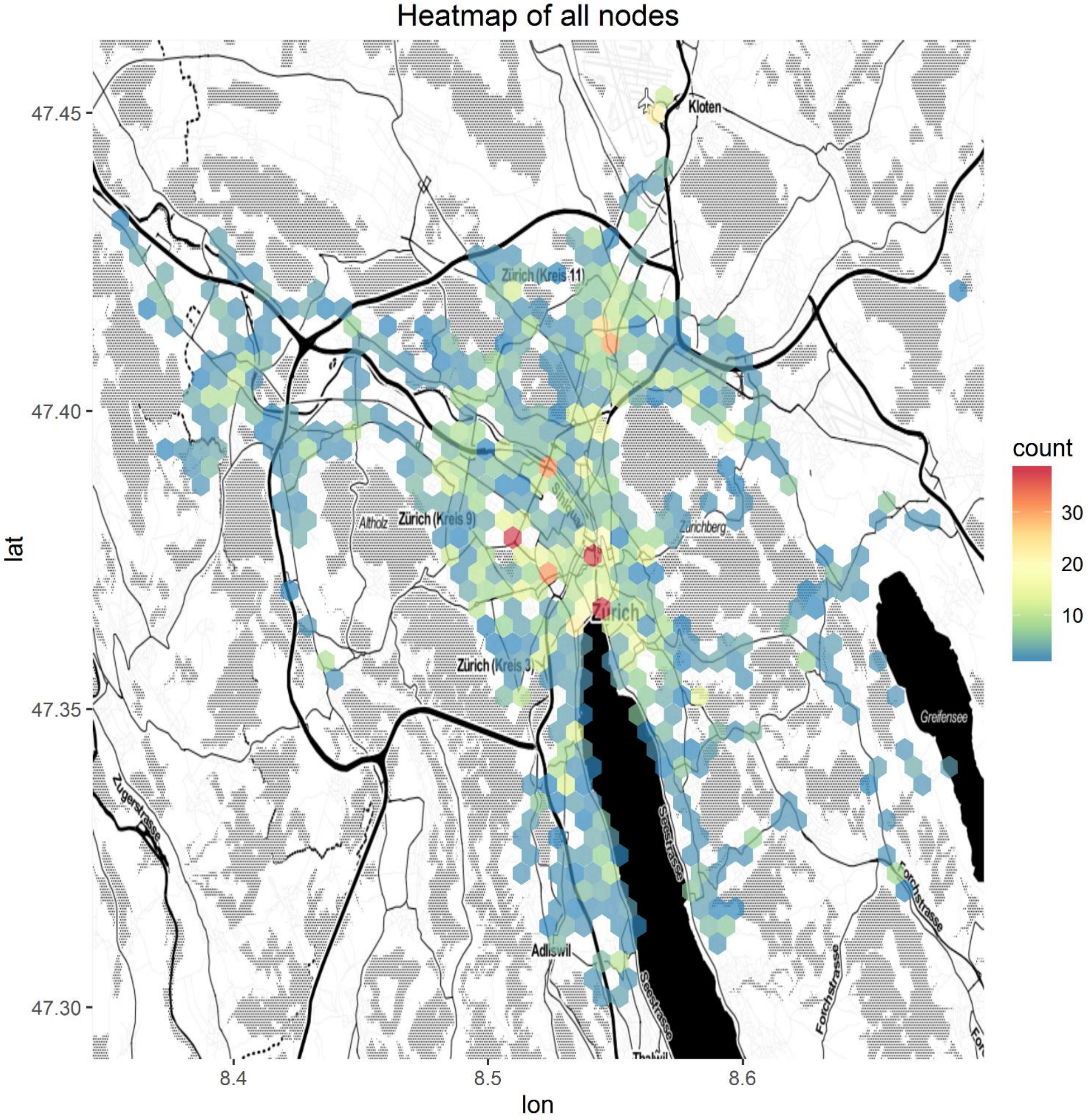

Figure 1: Hexagonal binned heat map of the node density of VBZ public transportation nodes

This paper revolves around the following hypothesis':

- Spatial and temporal hotspots of delayed arrivals and departures in the public transportation network of Zürich can be identified
- Nodes are directly or indirectly linked to other nodes enabling a traversal spread of delayed arrivals and departures
- Nodes particularly susceptible to delayed arrivals and departures and nodes with major influence on the public transport network can be identified

In the first section we have presented an introduction to this topic, which is followed by the second section the methodology. The results are found in the third section after which the discussion can be found in the fourth, the limitations in the fifth and the conclusion in the sixth section.

2. METHODOLOGY

2.1 Data

Our data origins from Open Data Zurich[1], where the "Verkehrsbetriebe Zürich" (VBZ) has made its database public.

[1] https://data.stadt-zuerich.ch/dataset?page=2

The data is made available on the OpenData Zürich web platform. New datasets are added in irregular intervals. Therefore, we had to regularly download the data. Due to the large size of the individual datasets (~230mb) only a limited amount of data (when writing this paper three months' worth of data) is made available. Older data is removed. In this paper we analysed the data from September 14. to the 31. December. The total dataset size is 3.63gb.

The datasets are composed of rows including the date, the should and actual arrival and departure times in seconds from midnight of the given date, the vehicle and line number and a node reference key to be able to join the node coordinates and details to each row. Seeing that every second of deviation in the course schedule is recorded, we had to aggregate our data in time. The time format is not in the UTC+1, so we first had to find the correct time format. Then we were able to aggregate our data into half hourly intervals. Moreover, one day for public transportation does not start and end at midnight, but the earliest departure of a vehicle is 4:32 and the latest is 2:00. Some of the trams leave the stations early, especially in the morning. We choose to only include delays less than -240 seconds in our analysis and declared everything less as outliers. Which means that leaving sooner than 4 minutes before the scheduled departure time was excluded.

It was noticed that some of the station nodes have multiple coordinates, as the coordinates for a station node are precisely for one platform in one direction. We aggregated the station node coordinates by taking the mean in the x and y axis.

We use the packages "Ggmap" and "Ggplot" to plot the maps and diagrams. "Ggmap" uses the map tile server stamen with the map type toner. We additionally use the package "Lubridate" to deal with date time formats, the package "Reshape2" to restructure our data to a more compatible structure and the package "RColorBrewer" to be able to use the ColorBrewer color pallets.

2.2 Computation

We run our analysis script in R on a standard home computer. At the time of writing this paper, the R script was not multicore enabled and due to the immense dataset size, led to an exceptionally long computation time of over four hours. The script allows setting various options including the input and output paths, analysis timeframe, interval, if arrival or departure times should be analysed, if only specific nodes or lines should be analysed, if maps, line diagrams or general statistics should be analysed and if a heat map of the node count per area should be produced. These options were introduced to allow different datasets to be analysed.

The script loads all files found in the specified input data directory and creates one large data frame. Once all files have been loaded into the data frame, rows not containing specified nodes or lines are dumped. Filtering early drastically reduces computation time for further analysis. The columns are then formatted to be compatible with the written script. Column names are changed and the time format is adapted. Finally, the dataset is joined with the node table to attach node coordinates to each row.

The analyser section, is then performing tasks set in the configuration section of the script. We perform a general visual analysis of the output maps, by generating a video file[2]. The video is composed of all maps creating a time lapse of all recorded delays, where every 150 milliseconds of video corresponds to an aggregated half hour of data. Out of the video visualisation we are able to see how the delays propagate through the transportation system and we can detect delay time differences according to the time of day and hotspots of nodes with high delay times. We create a hexagonal binned heat map [Figure 1] of all the analysed nodes. This gives a general overview of the density distribution of public transportation nodes. Further we create a line diagram plot of the mean delay times ranging from 14. September to 31 December [Figure 2]. We also create a diagram of the mean of all delay times according to the time of day [Figure 3] and grouped by week days [Figure 4], giving further information of temporal clustering of delay times. We performed the analysis using the delay times of arrivals and ran the script for the whole public transport network of Zürich.

3. Results

The produced heat map [Figure 1] shows an uneven distribution of public transport nodes with clusters near the main train station, Bellevue and Oerlikon. It shows the highest density of nodes is found in the inner city and generally decreases towards the outer boundaries of the public transport network of Zürich. The vicinity of the lake as well as the main transit axis' have a good overall cover of public transportation nodes. A few areas can be identified where the cover of VBZ nodes is only minimal or non-existent. These areas are mostly found at medium to large distances from the inner city and the majority is shown as forest area.

The created video shows high daily temporal fluctuations of delay times. One major peak is in the morning between 6:00 and 8:00 and another in the evening between 17:00 and 19:00. Generally, no delay times are registered between 1:30 and 4:30. Clusters of high delay times are observed in the inner city of Zürich near the main train station and Central and on the north-west axis connecting the inner city of Zürich to the Limmattal.

The video also visualises the propagation of delays through the public transport network. Two general delay propagation behaviours are identified: line following and network spread. Line following is what we identify as major delays following a specific public transport line, having minimal effect on other lines and having a rather short lifespan. We define network spread behaviour as major delay times of which a specific epicentre can be determined. The delays then spread out through the network affecting a large number of different public transport lines and having a longer lifespan then the line following behaviour. Hotspots of network spread epicentres are identified in the video as the inner-city and near Dietikon. A special combination can be observed where a line following or network spread behaviour in Dietikon traverses the public transport network into the inner-city causing a further network spread behaviour in the inner-city (e.g. 29.10.2015 6:30 – 9:00; video: 5:17 – 5:19). The two described behaviours are mostly observed between the mentioned peak hours.

High daily and monthly similarity can be observed in the line diagram plot of the mean delay times [Figure 2]. Major minus delay times (actual time of arrival < should time of arrival) are recorded at the beginning and slight minus delay times at the end of a given day. The majority of days are accompanied by two distinct peaks, of which the second is generally higher. A small number of outliers can also be distinguished, not fitting into the observed general pattern. The means of all stations in the public transport network of the VBZ averaged, without regarding the vehicle frequency.

[2] https://youtu.be/B5zUARiTe5o

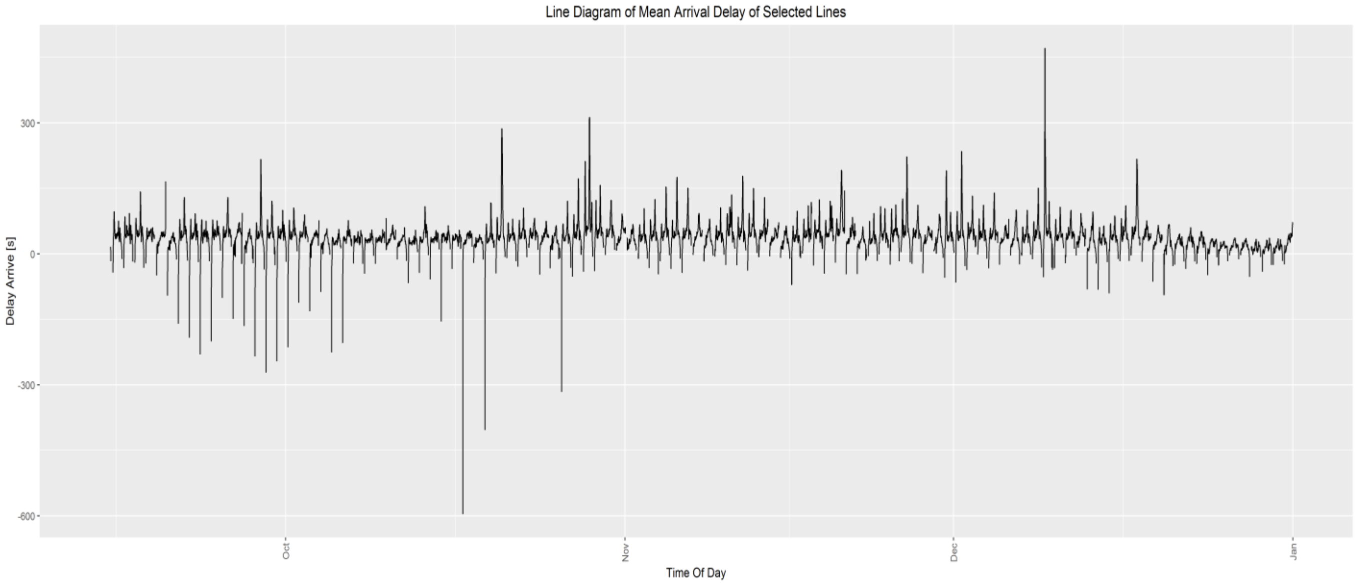

Figure 2: Line diagram plot of the mean delay times ranging from 14. September to 31 December

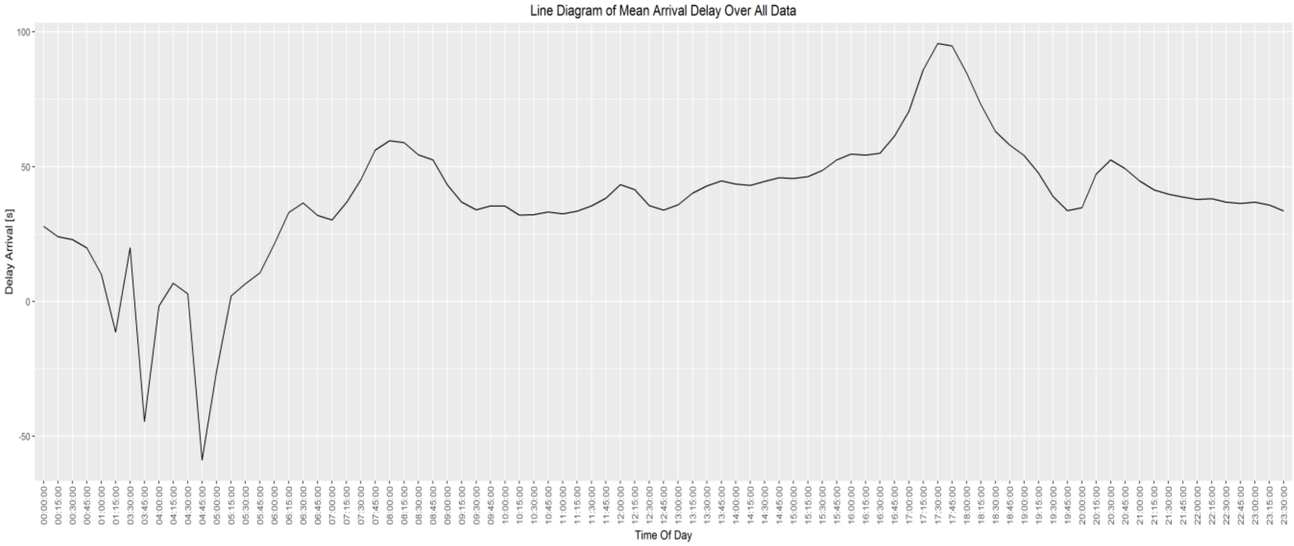

Figure 3: Mean of all delay times according to the time of day

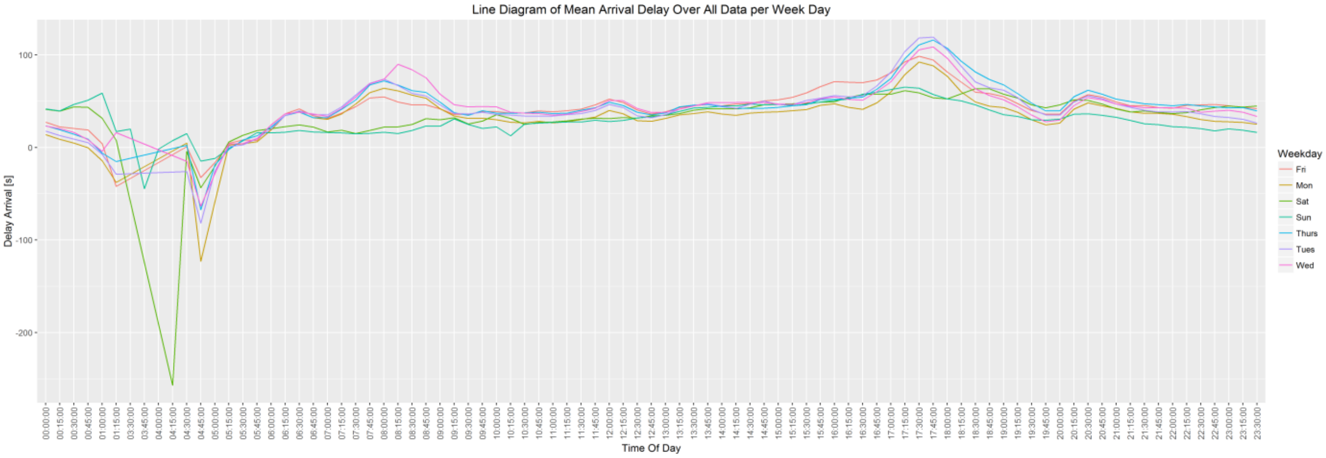

Figure 4: Mean of all delay times according to the time of day per week day

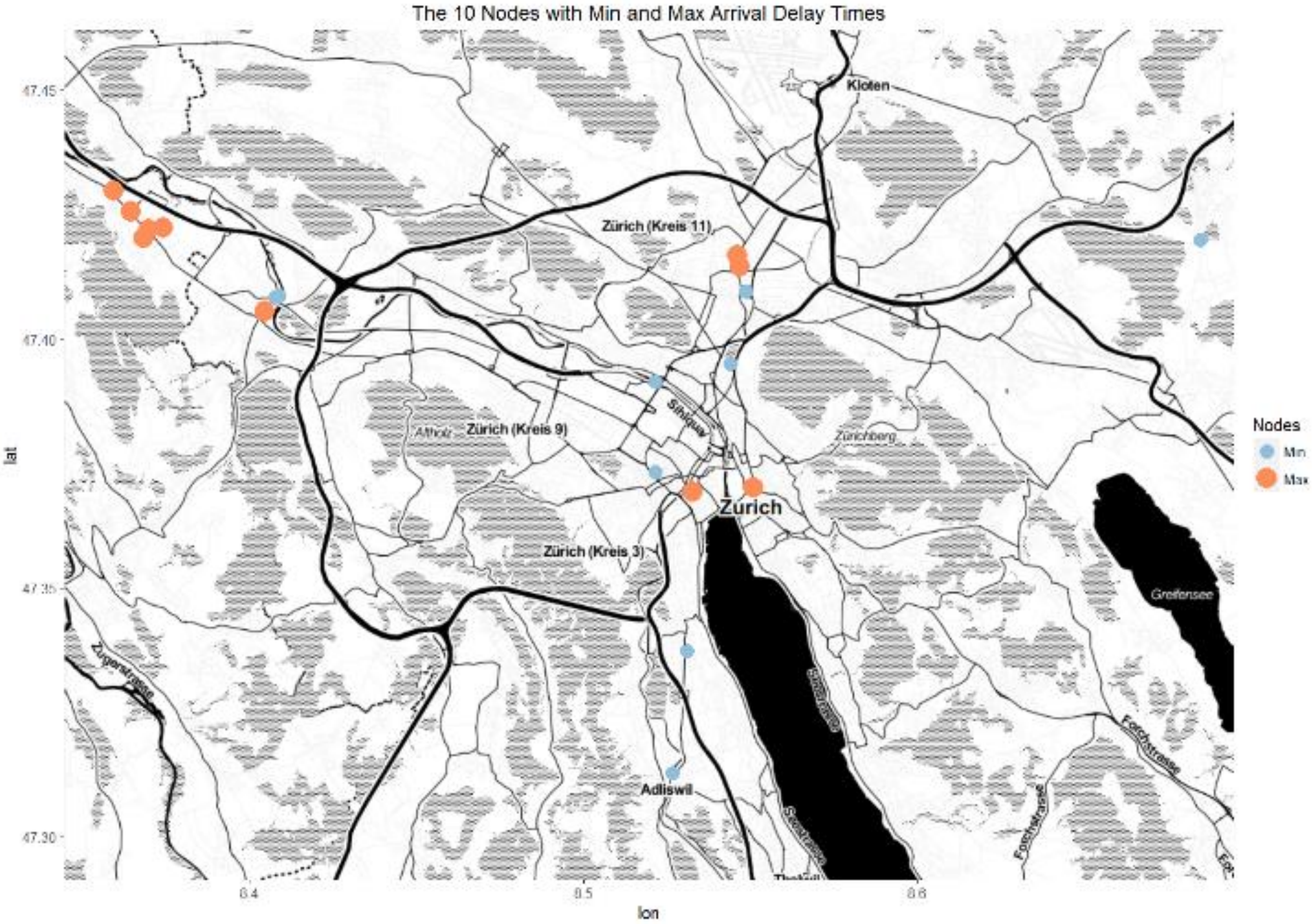

Figure 5: Map showing nodes with minimum and maximum mean delay times

The observed peaks are highlighted in the line diagram showing the mean of all delay times according to the time of day [Figure 3]. The first distinct peak is observed between 7:45 and 8:45, which is followed by a period of smaller mean delay times between 8:45 and 11:45. A small peak is present between 11:45 and 12:15, which is followed by an almost exponential increase in delay times up to the peak between 17:30 and 18:30. A small valley is detected between 19:15 and 20:15, which is directly followed by a minor peak between 20:30 and 21:00. The delay times then decrease linearly. Substantial noise is present between 0:30 and 4:45 and the scale is irregular. It must be pointed out that this is due to lack of data or highly irregular public transportation schedule.

According to the line diagram showing the mean of all delay times according to the time of day per week day [Figure 4], major differences are found between week days and weekend days. Week days have similar temporal characteristics of delay times, whilst weekend days do not show the two distinct peaks. Wednesdays show the highest mean delay times in the morning, whereas Fridays the lowest. In the evening Mondays show the lowest mean delay times and Tuesdays the highest. Saturdays show the highest mean delay times in the evening and Sundays have the lowest overall mean delay times. Fridays have an irregular peak in the evening forming with a first spike between 16:00 and 16:15 and a second between 17:15 and 18:15.

The generated map of the 10 nodes with minimum and maximum mean delay times [Figure 5] approximately correlates to the observed temporal characteristics of major and minor delay times in the mentioned video. The Limmattal in the north-west of Zürich is highly represented with 5 out of 10 of the nodes with the highest mean delay times. The inner city of Zürich and Oerlikon have two nodes each. The majority of nodes featuring low mean delay times are found on a diagonal axis from Adliswil to Oerlikon. Diedikon shows nodes with maximum, as well as minimum mean delay times.

4. DISCUSSION

It was mentioned in the introduction, that public transportation and the stability and flexibility of the networks thereof are of growing importance. We have created an analysis script in R and run various analysis on the available data of the public transportation network of Zürich. The results show various characteristics not only of a functioning public transportation network, but it can be argued that various characteristics of the general societal structure can be observed using delay times of a public transportation network as a proxy.

The diagram depicting the mean of all delay times according to the time of day [Figure 3] shows an increase of delay times during the rush hours around 8:00 and 17:30. The increase in number of passengers and vehicles on the street strains the network and causes higher delays. The delays registered in the morning peak are however not as distinct as in the evening. This finding correlates with [10], who showed how the number of people travelling by car increases in the morning and afternoon to similar times of the day as our observed peaks. They found that two thirds of all commuters travel with the public transportation system. The delay peaks with cars correlate with the found intra-daily delay peaks in the public transportation system. Similarly, they showed a stronger peak in the afternoon, than in the morning, but in contrast to our findings the authors describe a distinctive peak around 12:00. Our findings only show a small increase of the mean delay time of about 10 seconds, meaning the public transportation system can better cope with the lunchtime peak. It must be pointed out that the diagram [Figure 3] shows the mean of all delay times including the weekend. If only weekdays were averaged, the peaks would be higher.

4.1 Problematic node detection

Using our generated video and the map showing nodes with maximum and minimum delay times, our approach is able to detect potential problematic nodes. The maximum delays in arrival times occur at the train station, at the terminus of tram number 14 in Seebach, in Dietikon centre and around the shopping centre in Spreitenbach. All of these station either cope with many people, have a lot of possibilities to switch to other public transportation or both. The station in Seebach is probably an exception. Although some people switch from tram to bus and vice versa, the streets to and from the station were under construction at the time of data acquisition and the bus stations were relocated [11], causing delays for busses. The detected potential problematic nodes [Figure 5] correlate well with the observed delay behaviour in the video. These detected problematic nodes also correspond with the density of nodes [Figure 1], meaning an autocorrelation might be found if analysed further.

Although traversal delay relationships between nodes was not explicitly analysed, visually analysing the video suggests that Dietikon and the inner-city near the main train station and the node “Central” could potentially be catalysts for delay types with a network spread behaviour. The two mentioned areas are subject to the majority of network spread epicentres, hinting that these areas are highly connected with a large number of public transport lines and have large traversal impacts on neighbouring nodes. The video makes it evident, that delay propagations with network spread behaviours have a high negative influence on the whole network, making the epicentres of network spread a high priority target for improvements or potential delay mitigation. If network spread epicentre nodes and the varying impacts of different lines can be distinguished in space and time, one could write an algorithm to forecast the delay traversal throughout the network. Offering a such real time delay forecast system could improve user satisfaction and might also have a positive feedback on decreasing delay times (e.g. If a user sees high forecasted delay times, the user might decide to take a later vehicle, thus reducing the stress on the public transportation system by decreasing the number of users).

4.2 Inter- and Intra-Daily Analysis

The line diagram plot of the mean delay times ranging from 14. September to 31 December [Figure 3] shows large daily similarity regarding the day of the week. This affect is illustrated further in the diagram displaying the mean of all delay times according to the time of day per week day [Figure 4]. Various societal phenomenon can be read out of the presented diagrams. The peaks of the delay times are present from Monday through Friday, but not on Saturday or Sunday. We argue this correlates to the general structure of a week in this society, where for the majority of people Monday through Friday are workdays with fixed hours and Saturday and Sunday are weekend days. Fridays have the smallest mean delay times of all week days in the morning and Mondays in the evening. Mondays generally have the lowest recorded mean delay times. This could be caused by less people using public transport services due to not working on Mondays or Fridays (e.g. part time employment) or people working from home (e.g. home office). We also argue that Friday evening has an irregular two-stepped peak, because of many companies having working hours till 16:00 or 16:30 opposed to the regular 17:00 or 17:30. The observed valley after the evening peak time is present on all week days and the highest mean delay times thereof are observed on Saturdays. We also account this to the societal structure of a day, where the majority of people eat dinner between 18:45 and 20:00 and are thus not using public transportation. It can be argued that the found anomaly of Saturday is due to people going out on Saturday and having a different daily structure, including dinner time. The discussed valley is followed by another minor peak on all week days. We believe this peak is a result of people commuting home after completing some post work activities (e.g. sport; eating out; evening school). The highest recorded mean delay times of this minor peak are observed on a Thursday, hinting at after work socializing behaviour (e.g. after work beer; going out). Over all week days including the weekend days, Sundays show a minimum in mean delay times. This can be attributed to the fact that most stores and companies are closed on Sundays and people use the public transport network for leisure.

4.3 Event detection

During our analysis we noticed the viability of our approach to detect irregular events. The extraordinary peak on the 9. December [Figure 2] is due to a power cut affecting large areas of the inner-city of Zürich. All electricity dependant vehicles were affected. This event can also be seen in our created video starting at 10:13 and ending at 10:15 [footnote 2]. Another infrastructure related event can be seen on the 28. October [Figure 2], where the second highest mean delay times are registered. [12] reported a tram collision having severe impacts on the punctuality of various public transportation lines. These resulting high delay times also affect our created diagrams showing the mean of all delay times according to the time of day, distorting the calculated means to higher values. Another important event that can vaguely be observed is the change of the public transportation schedule on the 13. December 2015. The overall mean delay times seem to decline slightly and the outliers are drastically reduced. This could also be caused by the start of the holidays and more data would be needed to verify our presented assumption.

5. LIMITATIONS

The conducted analysis is accompanied by numerous limitations including not regarding stop frequency, not differentiating between buses and trams, problematic propagation nodes, problematic end nodes, events distorting mean delay times and autocorrelation of the node density with the potential problematic nodes.

In disregarding the stop frequency per node, a major aspect of the public transportation network is lost. Delays at nodes with a high frequency are not as severe as delays at nodes with low frequency (e.g. If the stop frequency is the same as the delay time, only the first vehicle is affected. Passengers taking later vehicles will not notice a delay, but the vehicle will report a delay). Furthermore, if a major delay is recorded on a low frequented node, it will drastically increase the mean delay of a day, in comparison to a major delay on a high frequented node.

Not differentiating between buses and trams might have an influence on the produced data. Buses and trams have different needs in regard to infrastructure. Buses use normal roads, which they mostly share with other private vehicles. Trams on the other hand need power lines and tracks. Trams therefore cannot take a route not in the network to circumnavigate a temporary obstacle, but on the other hand, trams sometimes have separate lanes. As trams and buses are analysed at the same time, delay times due to rush hours might be distorted Trams are less influenced by congestions on the street, whereas busses are highly dependent on road conditions. We argue that influences on the bus delays due to cars might be evened out by tram delays. The differences of trams and buses should be taken into account in a further step.

Noteworthy is the limitation regarding the problematic propagation nodes. Seeing that our analysis script only takes single nodes with fixed spatial and varying temporal components into account, the traversing of delays is not detected. This traversing of nodes could be of particular interest when detecting problematic nodes. It can be argued that problematic nodes are not the nodes with the largest mean delay times, but the nodes causing the largest delay times in the whole network. This would mean the effects of nodes on neighbouring nodes has to be analysed and nodes with the greatest negative effect on the network should be detected as problematic nodes.

Another limitation comes from using the same calculations on end nodes. As can be observed in the video, a majority of line following and network spread behaviours of delays are caught by the end nodes. End nodes often have a given temporal buffer to catch mentioned delays and should thus be treated differently.

Events and irregularities can have major impacts on the calculated mean delay times. Outliers like the discussed loss of power in Zürich in the morning of the 9. December has a major impact on the mean delay times on Wednesday. The peak on Wednesday morning [Figure 4] is a relic of that event. To increase the quality of the analysis, noise and outliers should be eliminated.

The last mentioned limitation is the autocorrelation of node density with potential problematic nodes. With an increasing density of nodes, the probability of detecting a problematic node within said area increases. To circumnavigate this limitation, the delay time per area could be calculated to be able to make assumptions regarding problematic areas and not nodes.

6. CONCLUSION

Our approach proved to shed light on various aspects of the public transportation network of the VBZ. We were able to distinguish various characteristics of the network in accordance with the societal structure of Zürich. The detection of problematic nodes was successfully conducted revealing controversial results. Overall the results of our analysis seem to correlate with empirically witnessed characteristics of the public transportation network. Especially noteworthy are the unexpected findings of only a minor peak around 12:00 and the detected valley around 19:00. The presented analysis can also aid in deciding at what days and times delays are to be expected. Although we have analysed the whole network and all the nodes, we believe this script could prove to be a worthy contribution to the daily analysis of the quality of specific lines or nodes.

7. FURTHER RESEARCH

Seeing the growing importance of public transportation and the increasing number of users thereof, we call for further research to be done in the domain of public transport analysis. We would particularly encourage research to be done regarding delay propagation through public transport networks and the detection of network spread epicentres and the affected lines. We argue this is of increasing importance especially for delay forecasting and making a public transport network able to optimally cope with large delay times and external influences.

8. ACKNOWLEDGMENTS

Our thanks to OpenData Zürich for making the VBZ data available and to ACM SIGCHI for allowing us to modify templates they had developed.

9. REFERENCES

[1] G. Georgiadis, I. Politis, and P. Papaioannou, “Measuring and improving the efficiency and effectiveness of bus public transport systems,” *Res. Transp. Econ.*, vol. 48, pp. 84–91, 2014.

[2] M. Bruglieri, F. Bruschi, A. Colorni, A. Luè, R. Nocerino, and V. Rana, “A Real-time Information System for Public Transport in Case of Delays and Service Disruptions,” *Transp. Res. Procedia*, vol. 10, no. July, pp. 493–502, 2015.

[3] G. Beirão and J. a. Sarsfield Cabral, “Understanding attitudes towards public transport and private car: A qualitative study,” *Transp. Policy*, vol. 14, no. 6, pp. 478–489, 2007.

[4] N. Paulley, R. Balcombe, R. Mackett, H. Titheridge, J. Preston, M. Wardman, J. Shires, and P. White, “The demand for public transport: The effects of fares, quality of service, income and car ownership,” *Transp. Policy*, vol. 13, pp. 295–306, 2006.

[5] D. a. Hensher, “The imbalance between car and public transport use in urban Australia: why does it exist?,” *Transp. Policy*, vol. 5, no. 4, pp. 193–204, 1998.

[6] B. R. Sampaio, O. L. Neto, and Y. Sampaio, “Efficiency analysis of public transport systems: Lessons for institutional planning,” *47th Annu. Transp. Res. Forum 2006*, vol. 1, pp. 371–385, 2006.

[7] R. Anderson, N. Findlay, R. Brage-ardao, and H. Li, “Measuring and Valuing Convenience and Service Quality,” 2013.

[8] R. Imam, “Measuring Public Transport Satisfaction from User Surveys,” *Int. J. Bus. Manag.*, vol. 9, no. 6, pp. 106–114, 2014.

[9] J. M. Del Castillo and F. G. Benitez, “A Methodology for Modeling and Identifying Users Satisfaction Issues in Public Transport Systems Based on Users Surveys,” *Procedia - Soc. Behav. Sci.*, vol. 54, pp. 1104–1114, 2012.

[10] C. Greiner, “Analyse: Arbeiter in Bewegung,” Zürich, 2013.

[11] R. Stäheli, R. Savoy, G. Markwalder, and A. Delle Donne, “BaustellenINFO,” Zürich, 2015.

[12] 20min, “20 Minuten - Tram kracht in Tram – Störung behoben - Zuerich.” [Online]. Available: http://www.20min.ch/schweiz/zuerich/story/Tram-kracht-in-Tram---Stoerung-behoben-23072142. [Accessed: 15-Jan-2016].