TRRafic jam! Krakow public transport analysis using R software

Sebastian Bysiak[1] and Dominik Kasperski[1]

[1]Facluty of Physics and Applied Computer Science, AGH UST

March 13, 2018

1. Introduction

1.1. Reflection

Punctuality is one of the most important aspect - from the passenger point of view - of public transport. No one would trust car or train if would know that it fault would be the reason why he or she is late on exam, date, dinner with friends or any other important meeting. Moreover, reliability can promote one mean of transport among the other ones. For example, Japanese high-speed trains are so popular and used, since any delay is unforgettable. Punctuality is the closest thing to be achieved, for Krakow trams and high-speed trains to have something in common.

1.2. Trams in Krakow

There are about 193 km of single track, 160 stops and 27 lines ([1]). Network operator is MPK Krakow (*Miejskie Przedsiębiorstwo Komunikacji.* For all other decisions (stops location, development, lines routing etc.) , municipal institution ZIKIT (*Zarząd Infrastruktury Komunalnej i Transportu*) is responsible ([2]).

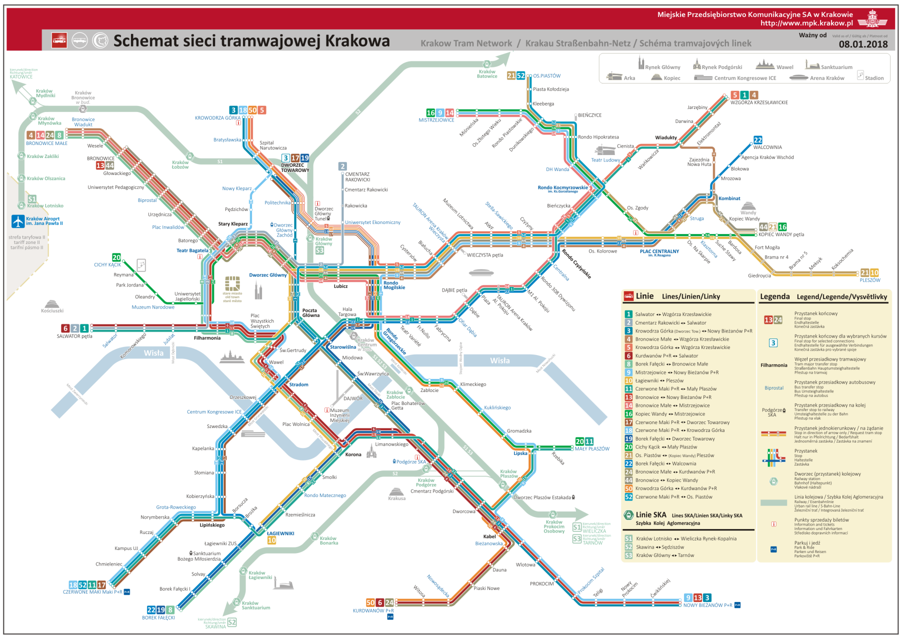


Figure. 1: Tramway network diagram [3]

Tram Traffic Supervision System (*TTSS [4]*) is used in Krakow to monitor tramways, provide data for Passenger Information System (*PIS*) ([5]). It consists of computers and various detectors installed in vehicles, which communicate with server ([6]). Solutions are present in 196 tramways (more than 85% of all). Then information is processed and displayed at 110 displays, next to tram stops. It is also present at www.ttss.krakow.pl. At www.mpk.jacekk.net/map by Jacek Kowalski ([7]) one can watch it operation. It shows vehicles' position, estimated times of arrival etc.

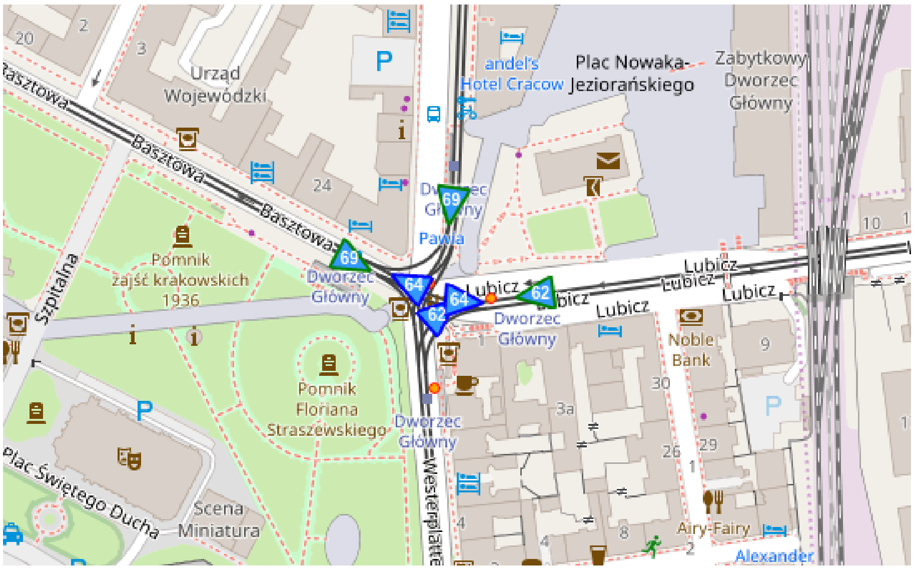
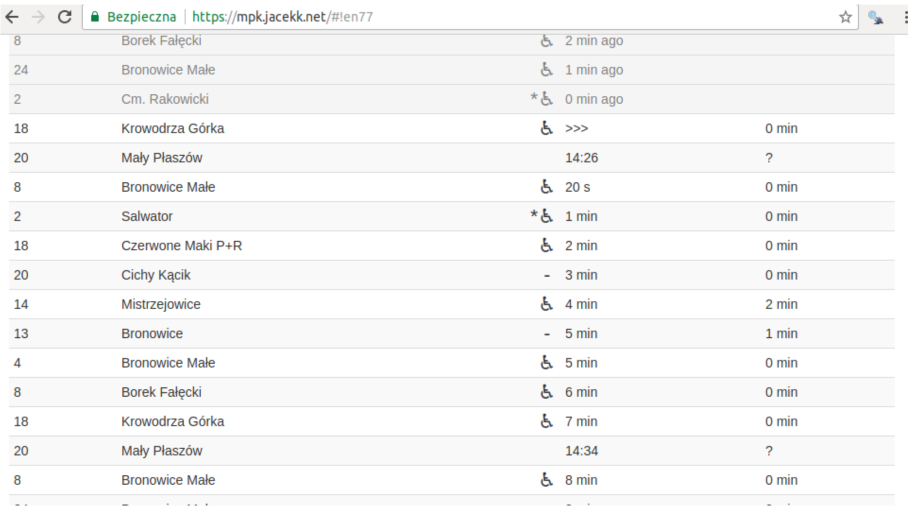


Figure. 2: www.mpk.jacekk.net/map [7]

1.3. Web scraping

Lot of data is streamed online in good-looking sites, what makes it reachable for many people. It could be good source of information to be analyzed. Web scrapping is a way of computer processing information directly from the internet sites. Downloading page as a view from browser and then processing it as a text or XML documents seems easy, but it gets more complicated when one is dealing with internet techniques as JS and other tools which change content dynamically, without reloading ([8]).



8	Borek Fałęcki	♿	2 min ago	
24	Bronowice Małe	♿	1 min ago	
2	Cm. Rakowicki	*♿	0 min ago	
18	Krowodrza Górka	♿	>>>	0 min
20	Mały Płaszów		14:26	?
8	Bronowice Małe	♿	20 s	0 min
2	Salwator	*♿	1 min	0 min
18	Czerwone Maki P+R	♿	2 min	0 min
20	Cichy Kącik	-	3 min	0 min
14	Mistrzejowice	♿	4 min	2 min
13	Bronowice	-	5 min	1 min
4	Bronowice Małe	♿	5 min	0 min
8	Borek Fałęcki	♿	6 min	0 min
18	Krowodrza Górka	♿	7 min	0 min
20	Mały Płaszów		14:34	?
8	Bronowice Małe	♿	8 min	0 min

Figure. 3: www.mpk.jacekk.net

2. Method

2.1. How to measure punctuality?

TTSS provides information about all trams which would arrive at the stop. So if one simply take that information, it could happen that one vehicle generates many delays. First solution proposed and applied was to count tramways which will arrive in 3 minute period after measurement. Moreover, spatial distribution was made to be more sure that delayed tram is "visible" at only one stop at a time. It means that not all stops were measured - every at city center, but at long streets toward loops one in three or four.

Then it was figured out how to get unique arrivals. For example, during fire at Królewska Street 6/02/2018 around 5 AM, vehicles were standing at place where *TTSS* gave them less than 5 minutes to arrive at Plac Inwalidów, but they were standing due to action of firefighters etc. Multiple times during that event we recorded increasing delay. Thanks to the fact that timetable was also in collected data, the longest delay for unique timetable hour in a day was only processed.

Spatial distribution of stops measured was also made to assure relevance of the data. Also it could happen if some area was omitted that all delays are caused by that "black box". Every loop is measured since they "consume" delay. If delayed tramway arrives, it has shorter brake and goes back to operation without delay in opposite direction. So loops are good point of reference, kind of "ground".

Previously aim of research was to check if rush hours could be stated by analysis of public transport's delays time sequence. Due to short period of data acquisition, aim was not reached. Map of dealys was made and diagrams showing time distribution.

Data were collected between Sunday and Wednesday evenings, from 4/02/2018 to 7/02/2018. It consisted 111387 unique arrivals. Although, some data were lost, especially about few tramways in directions such as Dworzec Towarowy, Os. Piastów Cmentarz Rakowicki and Plac Centralny im. R. Regana. Also if vehicle changes its route, *TTSS* do not provide information about timetable on stops which are normally not covered by that line. It also may be the reason for enormous delays: if vehicle goes it is measured on the first regular stop. But it can happen that tramway change its normal loop, and goes back with different direction: then *TTSS* matches it with timetable for stops en route, and such measurement shows delay but it is not proportional to the disruption of functioning public transport. Moreover, thanks to *PIS* which also display information about such changes (in e.g. "Lines 4, 44, 24, 8 directed towards Cichy Kacik") it is not delay which causes discomfort for passengers ([5]). *PIS* shows also real times of arrivals, and thanks to the frequency (fast lines (50, 52) have one vehicle in 5 miunutes, other one in 7-8, or one in 10, or one in 15) ([9]), such event do not cause public transport to fail. It causes a disruption which affects measured delays, but not in a way as it affects passengers. It can be stated that high measured delay affect punctuality of public transport, but only the short ones are causing passenger discomfort.

Also from timetable point of view there is no difference if vehicle arrives at 12:55:00 and 12:55:59. Of course time spent on stop affects the delay and should be minimized in order to increase speed of public transport - travelling with average speed of 20 km/h is not sufficient. But since that basic unit of measurements is one minute.

2.2. Web scrapping in R – options

2.2.1. Package `'rvest'`

([8]) Site www.mpk.jacekk.info was chosen to be scrapped. It presents dynamic data about each stop (line, direction, ETA in minutes, delay in minutes), and author provides source code on GitHub. Dynamic effect are generated using JS, so basic `read_html()` command from `'xml2'` [10] package returns only an empty frame without data one wants.

Package `'rvest'` [11] was intended to wrap `'xml2'` and `'httr'` packages. It provides functions which can parse XML to nodes, and then a node to data frame. But before that one need to use headless webkit such as Phantom JS. It is a kind of GUI-less web browser. With some help of JavaScript it can provide an html document with all scripts etc. evaluated. Such page contains all information, and could be easily parsed using ’rvest. Package*’stringr’*([12]) was used as well. Relevant part of such script is shown below:

```
library(’rvest’)
library("XML")

# scrape_final.js:
#var url =’https://mpk.jacekk.net/#!pl126’;
#var page = new WebPage()
#var fs = require(’fs’);
#
#page.open(url, function (status) {
#        just_wait();
#});
#
#function just_wait() {
#    setTimeout(function() {
#             fs.write(’1.html’, page.content, ’w’);
#          phantom.exit();
#    }, 2500);
#}
#end of script


url <- paste0(’https://mpk.jacekk.net/#!pl’,to.url[i])
lines <- readLines("scrape_final.js")
lines[1] <- paste0("var url =’", url ,"’;")
writeLines(lines, "scrape_final.js")

# Download website
system("phantomjs scrape_final.js")

# rvest job
pg <- read_html("1.html", encoding = "UTF-8")
node<-html_node(pg,’#times-table’)
node.xml <- xmlParse(node)
df <- xmlToDataFrame(node.xml)

# and that’s it!
```

It works, but one stop is parsed in 4-5 seconds. Due to that fact firstly 5-minute interval and limited number of stops was chosen.

2.2.2. Package ’jsonlite’

Improvement of described above method utilizes HTTP requests, which enables to received pure data – without any HTML tags which needlessly (from our point of view) increases size of the files downloaded and forces us to parse them in search for useful data. Usage of HTTP requests is possible thanks to service build in https://mpk.jacekk.net/ which gives user access to data in JSON format (through: https://mpk.jacekk.net/proxy.php/services/passageInfo/stopPassages/stop?stop=77&mode=departure)

With such an input, processing in R is trivial: function read_json() from package jsonlite converts JSON into dataframe, moreover it’s possible to pass url instead of file as an argument - one doesn’t even need to save JSON file on local drive - only after filtering out unneeded data we were saving dataframe to a *.csv file.

2.3. Final method

Method described in 2.2.2 was *sufficiently* fast - we could basically save data for every tram stop with intervals below 2 min. We didn’t do that due to a few reasons:

- we didn’t wanted our data to rise beyond reasonable size, which could be analyzed fast on laptop
- there was no need for checking specific tram stop more often than every 3 – 5 min

- there was no need for checking every tram stop – cases where there are many stops on a single street without larger intersection – nothing interesting happens between them so by skipping every second or third one does not lose much interesting data

Our final choice was to *visit* each from 70 selected stops every 3 minutes (artificial latency had been added). As with every request we were given data for about 20 following minutes, we had to cut off all trams with expected time of arrival (ETA) greater than 3 minutes – in order to avoid multiple saving of particular tram. Data returned by every request was saved in separate file.

Later on, better filtering technique was developed – if a record for specific line, direction and timetable arrival time was found in multiple files, only the latest one was kept. It solves issue with a tram stopped for a longer time in the nearby of the stop – it happened e.g. during the fire on Królewska st. when some trams were immobilized for about an hour and they were logged more than 10 times – each time with larger delay but the same ETA (below 3 min.).

2.3.1. Further data processing

After collecting data, we rewrited it into form easier to analyze, e.g. we gathered all records concerning specific lines and directions.

In further analysis we utilized R graphics package `lattice` [13] and `osmar - OpenStreetMapAndR` [14], which provides an interface for accessing data available online on OpenStreetMap as well as easy on–the–map plotting.

3. Results

Red line along the plots means level of one minute delay

Summary of collected data is shown in table (1). About half of tramways arrived on time, mean delay is about 1 minute 20 seconds. If delayed, vehicle has delay about 2 and a half minute. When one take out maximum and mimimum value, will obtain that about a quarter of tramways is delayed for more than two minutes which is consisted with 3rd quantile of all data.

Table 1: Summary

Type	Min.	1st Q	Median	Mean	3th Q	Max.	Entries	Percent of all
All	0.000	0.000	1.000	1.351	2.000	88.000	111387	100%
Delayed	1.000	1.000	2.000	2.674	3.000	88.000	56271	50.5%
Without max and min	2.000	2.000	3.000	4.149	4.000	54.000	29888	26.8%

Plots (4) and (5) are showing two time series. First of them depicted (4) is for all data, including punctual tramways. Although some data are corrupted, some patterns are visible. For example little peaks showing morning and afternoon rush hours. Impact of large stop in traffic at Krolewska around 6/02/2018 18:00 is visible. Second one, at (5) is only more lifted and more sharp in few places. Both of them are under influence of large delays. Plots are made using packages *'ggplot2'*([15] and *'latticeExtra'*([16].

Top 3 punctual lines (2) are those which are short or operates on track separated from street. 21 due to renovation of track to Pleszew is operating to loop Kopiec Wandy, and its length is about 6 km mainly thouogh separate track out of city centre. Line 2 from Salwator to Cmentarz Rakowicki is about 4,9 kilometers long, on not so attended streets in city centre. Line 11 is similar to 21 - it operates between Czerwone Maki P+R to Maly Plaszow, on separate track, through suburbs.

Table (3) shows summaries for most unreliable lines. Number 6 operates between Salwator and Kurdwanów, dividing track with fast line 50, with the most punctual 2 and 11. But is using the oldest vehicles ([17]) and with low frequency - one in every 15 minutes. So in data collected it is most affected by other events such as fire of tramway track at Jubilat etc. Line 17 is working only during rush hours through heavily congested areas. Line 24 partially uses oldest vehicles and goes through Krolewska street where massive and long stop occured. Interesting fact is that both 24 and

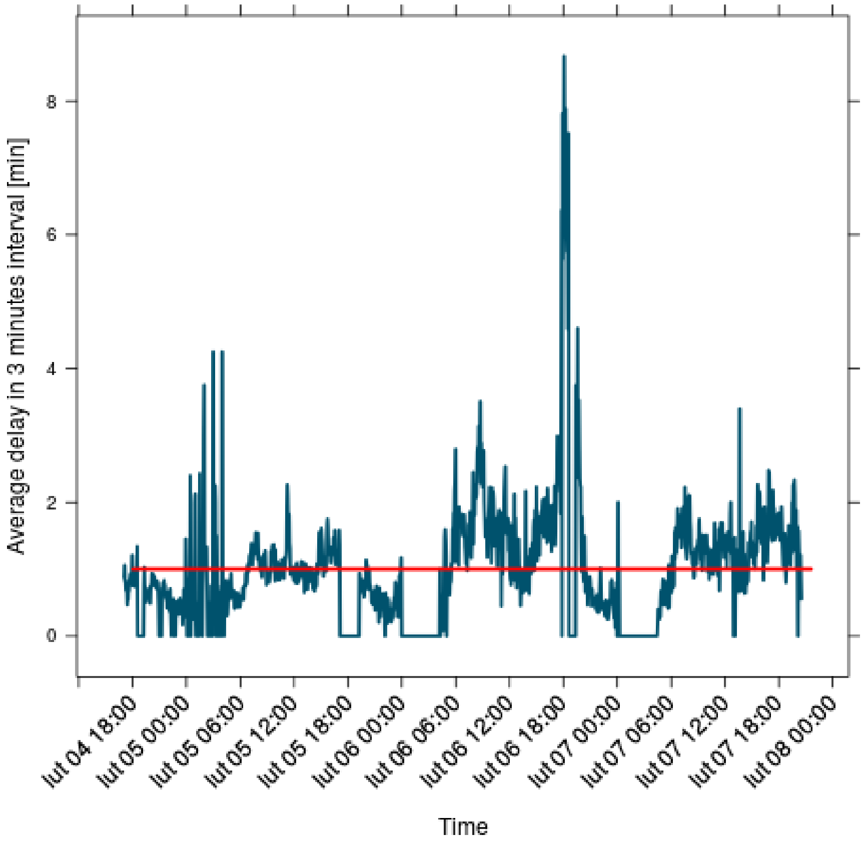


Figure. 4: Average delays.

Table 2: Top 3 punctual lines (lowest mean delay).

Line	21	2	11
Mean [min]	0.22	0.54	0.62
Standard deviation [min]	0.69	1.1	1.3
Maximum [min]	4	8	16

6 divide part of its route with Krakow fast tram 50. Lines 52 and 50 are designed to be the fastest and have frequency of 1 tramway in 5 minutes. All data about the lines are shown in table (4.

Table 3: Worst lines (highest mean delay).

Line	6	17	24
Mean [min]	5.2	1.9	1.8
Standard deviation [min]	10.9	3.7	3.7
Maximum [min]	47	33	44

Table (5) shows most punctual stops. Three of them are loops. Half of arrivals is assumed to be punctual. Only one line stops at Cichy kacik - 20.

Table (6) shows most punctual stops which are not loops. Results are quite surprising. Uniwersytet Ekonomiczny is attended only by 1 line, Wiadukty are near loop Wzgorza Krzeslawickie. Nowy Kleparz is attended by one line, 18 and it is at one end of Dluga Street which is famous mainly from stopping of tramways by cars standing outside of given space. It shows how good MPK is responding to such event.

Table (7) shows stops with highest mean delay. Surprisingly the worst is Salwator which is a

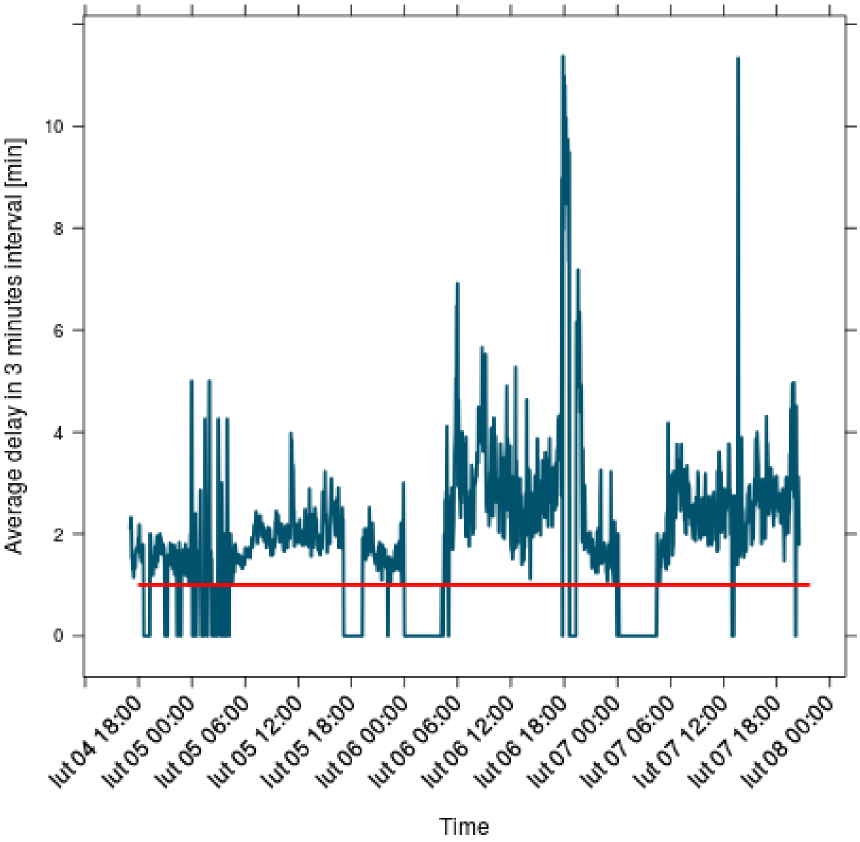


Figure. 5: Average delays without punctual trams.

loop. Next is Jubilat, which is next stop en route investigated. 5/02/2018 fire of tramway track happen at Jubilat Another is Limanowskiego in congested area.

In that part of investigation, delays of a few trams on stops registered on 6/02/2018 is considered. At (6) and (7) diagrams depicting line 18 are shown. As one can see 18 goes through Stary Kleparz - Nowy Kleparz, at Dluga. If it goes in direction of Czerwone Maki, only few delays occure. At Stary Kleparz it is punctual, but on Nowy Kleparz few delays are visible. As facebook page of passengers with on-line information about stoppings and other events "Platforma Komunikacyjna Krakowa - PKK" informes ([18]), there was 10 minute stopping at Dluga around 5 AM. After fire at Krolewska many vehicles were directed on other loops than Bronowice or Bronowice Male causing "tramway jam" what is also visible at line 18. Also in opposite direction delays on Dluga tends to increase, but mainly due to the fact that this is a cozy street without separated traffic. Plots are made using packages *'ggplot2'*([15] and *'latticeExtra'*([16].

Diagram (8) shows line 2 towards Salwator. As one can see it is not frequent. It gains some delay between Lubicz and Filharmonia. Data from Stary Kleparz - it is possible for tramway to gain some time there since the track is separated. That area is congested with tramways, about 7 lines cover them what makes it is the most accesilbe part of the city. So as line 2 shows, dense tramway transport may cause delays. Surprisingly line 2 reduces delay between Filharmonia, Jubilat and Salwator. Plots are made using packages *'ggplot2'*([15] and *'latticeExtra'*([16].

Doubts about results for line number 6 and Jubilat and Salwator stops are supported by diagrams (9) and (10). It seems that one tram simply runs half an hour after it should and that due to old vehicles - line is not covered by the *TTSS* ([5]). Plots are made using packages *'ggplot2'*([15] and *'latticeExtra'*([16].

Figure (11) shows correlations of time sequences of mean delay in 10 minute window between each stops. They should be neighbouring ones, and it most cases they are. Figure clearly shows, that delays defined in that way are uncorrelated. That cannot be true, and those measure is improper,

Table 4: Full data about lines delays

Line	Mean delay	Maximum Delay	Standard deviation
21	0.2159624	4	0.6731569
2	0.5376059	8	1.0682412
11	0.6234973	16	1.3468231
16	0.6710526	4	0.9001949
18	0.7541102	88	2.4109798
62	0.7849462	5	1.2670529
64	0.7931034	5	1.3221671
20	0.8207780	24	1.6055775
69	0.8415842	5	1.1217672
5	0.8474253	18	1.5140105
1	0.8613808	22	1.5523201
9	0.9399745	7	1.2254901
4	0.9810848	42	2.8222580
3	1.0122592	13	1.3303886
50	1.1232566	24	1.7798215
22	1.1840574	37	1.7357053
10	1.2172564	17	1.9932565
19	1.2926125	10	1.5224087
52	1.3047371	13	1.7795984
8	1.4207224	49	3.6104089
44	1.5546411	48	5.0478144
14	1.5659717	47	3.9901735
13	1.7751660	52	3.6630723
24	1.8084939	44	3.7031147
17	1.9326241	33	3.6528597
6	5.1711943	47	10.9454339

Table 5: Top 3 punctual stops (lowest mean delay).

Stop	Walcownia	Cichy Kacik	Os. Piastow
Mean [min]	0.013	0.017	0.022
Standard deviation [min]	0.113	0.237	0.229
Maximum [min]	1	4	3

Table 6: Top 3 punctual stops which are not loops (lowest mean delay).

Stop	Uniwersytet Ekonomiczny	Wiadukty	Nowy Kleparz
Mean [min]	0.078	0.301	0.676
Standard deviation [min]	0.332	1.12	1.34
Maximum [min]	3	22	10

since some lines have frequency of one vehicle in more than 10 minutes. So defining correlation between stops is hard and is out of scope of this research.

Figure (12) shows map of Cracow with mean delays measured on the tram stops marked with numbers and red circles, which size denotes the delay.

Table 7: 3 worst stops (highest mean delay)

Stop	Salwator	Jubilat	Limanowskiego
Mean [min]	3.09	2.47	2.43
Standard deviation [min]	7.71	6.97	6.43
Maximum [min]	46	52	45

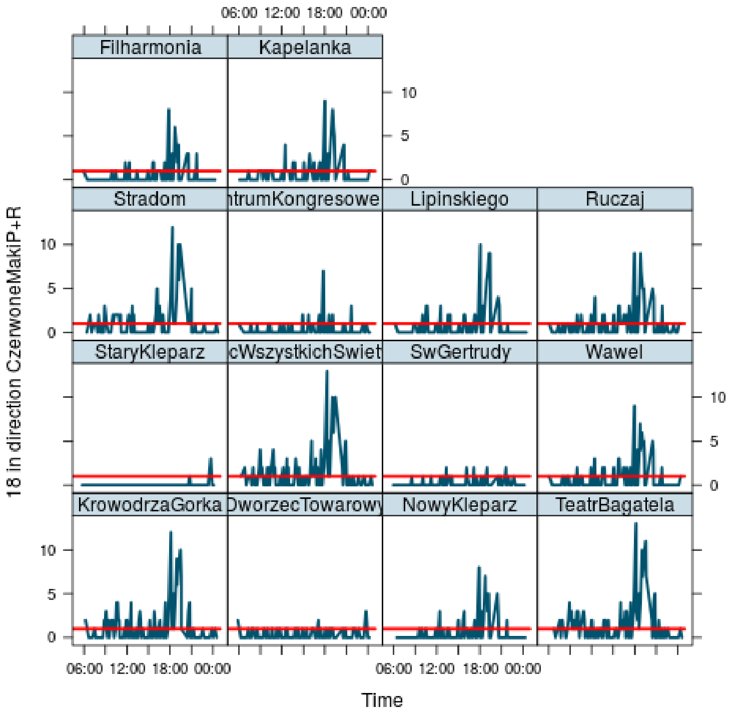


Figure. 6: Delays of a line in given direction on each stop

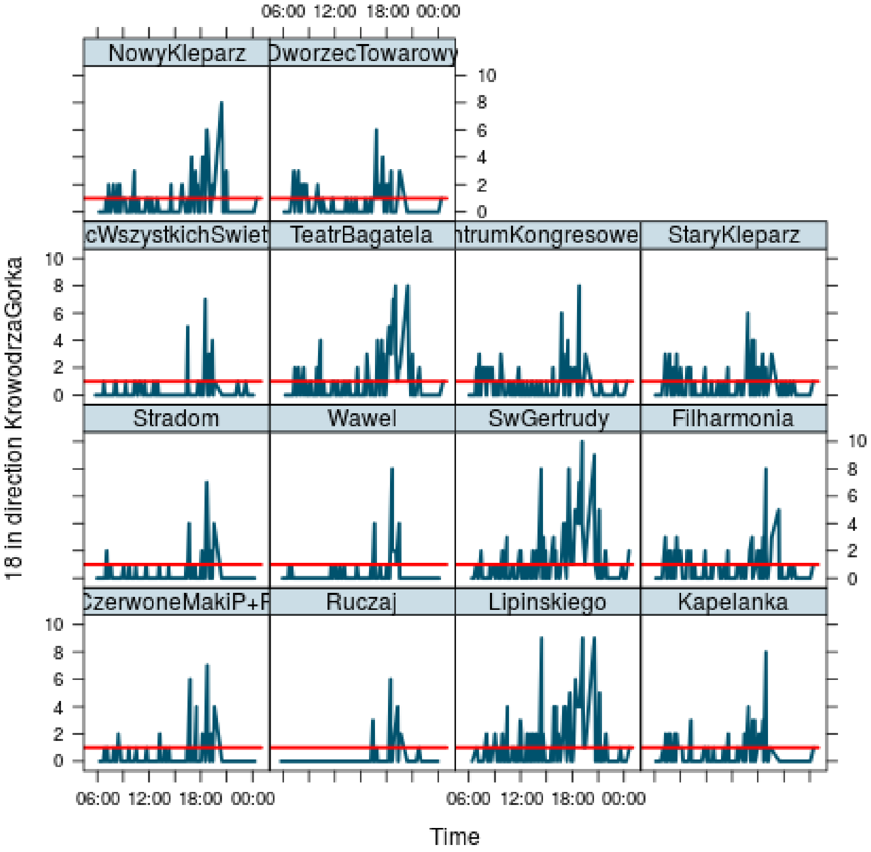


Figure. 7: Delays of a line in given direction on each stop

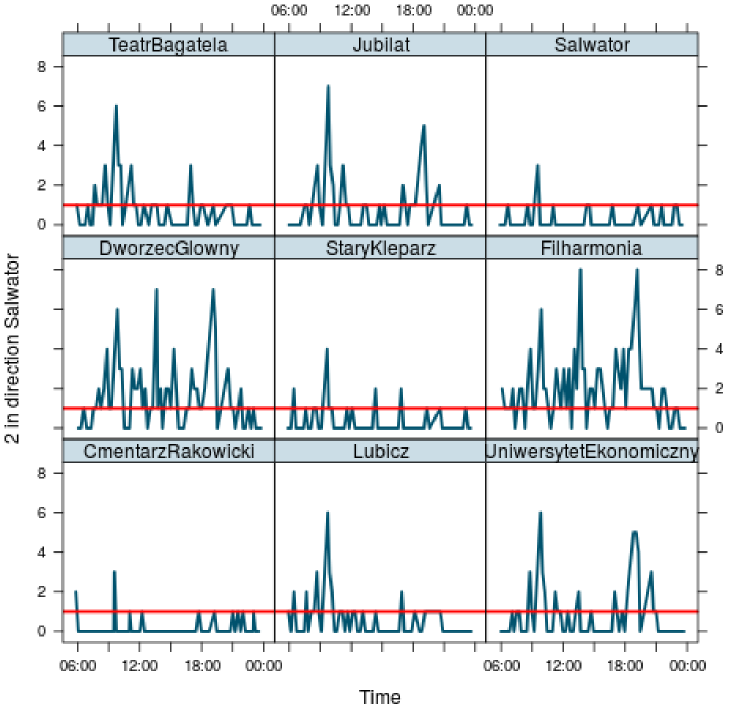


Figure. 8: Ddelays of a line in given direction on each stop

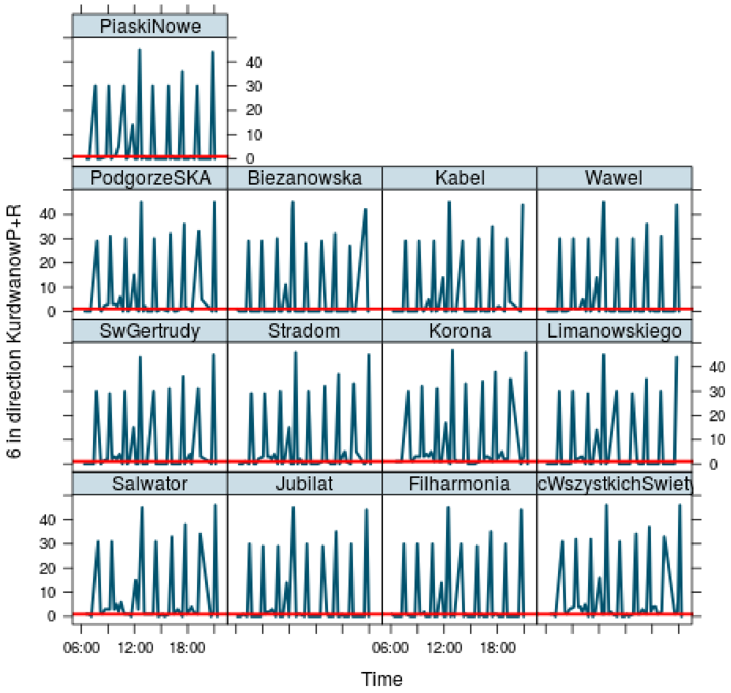


Figure. 9: Delays of a line in given direction on each stop

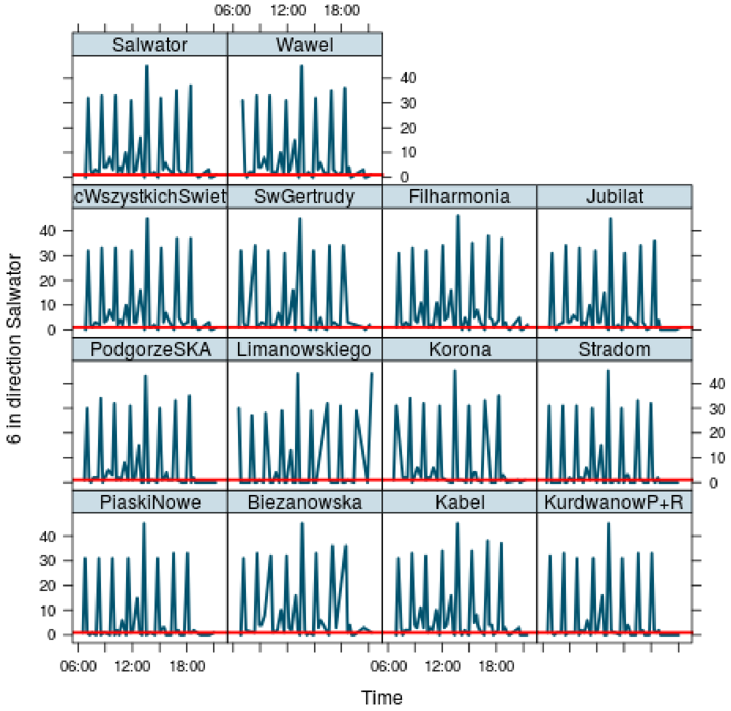


Figure. 10: Delays of a line in given direction on each stop

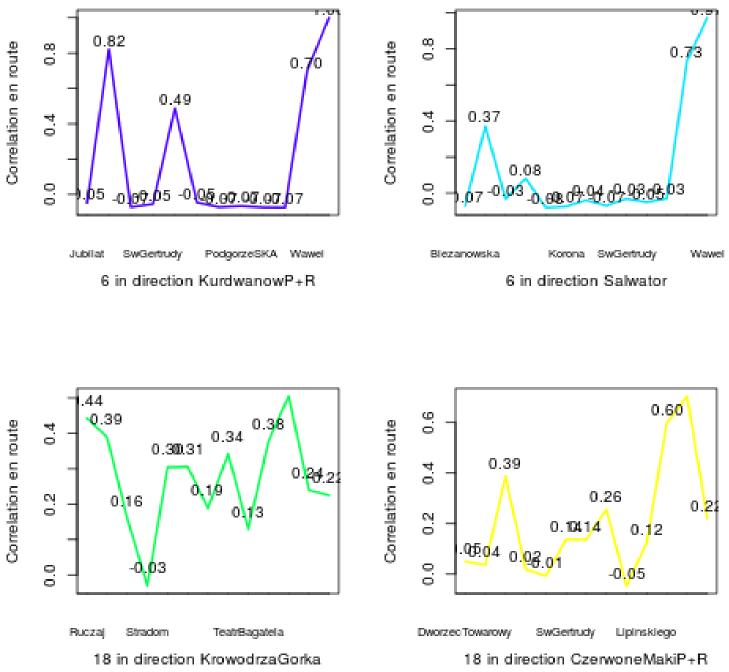


Figure. 11: Correlation of delays between each stops.

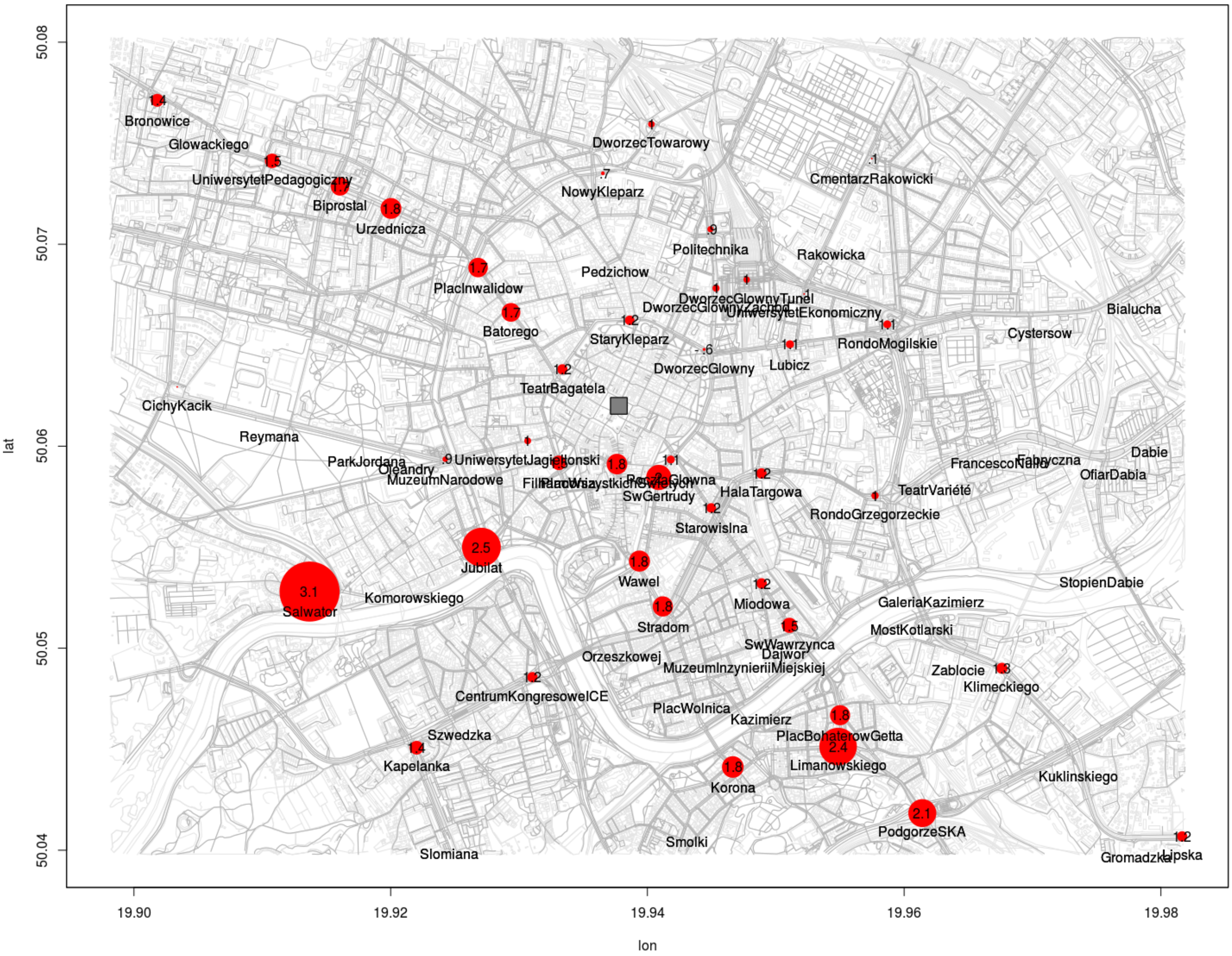


Figure. 12: Map of Cracow with trams stops and mean delay on specific stops marked as number in red circle [min], which size is proportional to the delay. Created using `osmar` package.

4. Conclusion

Tramways in Krakow are punctual in terms of data provided by the following research. Mean delay is 1.3 minute what in terms of timetable precision and other factors affecting passenger's comfort is a good result. Długa St. is confirmed to increase delays of tramways. Some effects of not all lines to be covered with TTSS are shown, as false line 6 delays. Also lines which operate on the oldest vehicels are the slowest. Effects on tramway congestion is catch.
An attempt to describe quantitatively delays of public transport was made. Although methods have some limitations, general trends are consistent with reality. Despite the fact that unique arrivals were filtered out of the data, they were still blurred. To do it better one needs to rethink some measures - as use of correlation. Data were collected on too short period of time to catch general patterns. Also problem of missing data should be solved, as well as some measurement of tramways out of its usual route should be implemented. It will help catch the congestion on tracks, which as it was shown, also affects tramways punctuality Wide possibilities of R software are shown, from various ways of web scrapping through data analysis to plotting diagrams. The preparation of the project proved (the authors) that it's a very mature tool with outstanding capabilities to perform complex tasks easily.

References

[1] URL: http://www.dziennikpolski24.pl/region/wiadomosci-krakow/a/krakow-plona-szyny-tory-sa-w-fatalnym-stanie-czekaja-nas-lata-remontow,12674582/.

[2] URL: http://www.gazetakrakowska.pl/artykul/522005,po-co-w-krakowie-jest-zikit,id,t.html.

[3] URL: http://www.mpk.krakow.pl/pl/mapki-komunikacyjne/.

[4] URL: http://www.ttss.krakow.pl/.

[5] URL: http://kmkrakow.pl/informacje-o-systemie-kmk/infrastruktura/136-dynamiczna-informacja-pasazerska.html.

[6] URL: http://kmkrakow.pl/informacje-o-systemie-kmk/infrastruktura/138-system-sterowania-ruchem.html.

[7] URL: https://mpk.jacekk.net/map.html.

[8] URL: https://www.r-bloggers.com/web-scraping-javascript-rendered-sites/.

[9] URL: http://www.gazetakrakowska.pl/wiadomosci/krakow/a/krakow-maksymalizacja-czyli-od-8-stycznia-tramwaje-i-autobusy-pojada-czesciej,12751069/.

[10] Duncan Temple Lang and the CRAN Team. *XML: Tools for Parsing and Generating XML Within R and S-Plus.* R package version 3.98-1.9. 2017. URL: https://CRAN.R-project.org/package=XML.

[11] Hadley Wickham. *rvest: Easily Harvest (Scrape) Web Pages.* R package version 0.3.2. 2016. URL: https://CRAN.R-project.org/package=rvest.

[12] Hadley Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations.* R package version 1.2.0. 2017. URL: https://CRAN.R-project.org/package=stringr.

[13] Deepayan Sarkar. *Lattice: Multivariate Data Visualization with R.* ISBN 978-0-387-75968-5. New York: Springer, 2008. URL: http://lmdvr.r-forge.r-project.org.

[14] Manuel J. A. Eugster and Thomas Schlesinger. "osmar: OpenStreetMap and R". In: *R Journal* (2010). Accepted for publication on 2012-08-14. URL: http://osmar.r-forge.r-project.org/RJpreprint.pdf.

[15] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2009. ISBN: 978-0-387-98140-6. URL: http://ggplot2.org.

[16] Deepayan Sarkar and Felix Andrews. *latticeExtra: Extra Graphical Utilities Based on Lattice.* R package version 0.6-28. 2016. URL: https://CRAN.R-project.org/package=latticeExtra.

[17] URL: http://kmkrakow.pl/informacje-o-systemie-kmk/tabor-naszych-operatorow/57-tabor-tramwajowy.html.

[18] URL: https://www.facebook.com/PKKinfo.