ECOLE POLYTECHNIQUE DE BRUXELLES

Hack My Ride

Assessment of Quality of Service

Hakim Amri (000459153), Rania Baguia (000459242),
Mehdi Jdaoudi (000457507), Abdelmoumen Oumahi (000494018)

Project submitted under the supervision of

Professor : Mahmoud Sakr

In the context of the course

INFO-H423 Data Mining

Academic year

2022-2023

Today's Agenda

- Business Purpose and Problem Statement
- Data Description
- Data Exploration
- Tasks Data Mining Solutions
- Interactive Dashboard

Business Purpose and Problem Statement

STIB Vision and Mission

"*Together, let's transport Brussels of tomorrow!*"

Mission: To provide the most suitable travel solution with a competitive travel time at an affordable price in very good comfort and safety conditions

Key Fact

To determine which approach to use, it is assumed that regularity is more relevant to assess QoS than punctuality when the headway is lower than 12 min

Problem Statement

Assessment of quality of service in terms of service reliability.

Service reliability is defined as having no excess waiting time for the travelers.

- Punctuality: delays with the scheduled timetable
- Regularity: irregularity of headways

ECOLE POLYTECHNIQUE DE BRUXELLES

An End-to-End Data Project

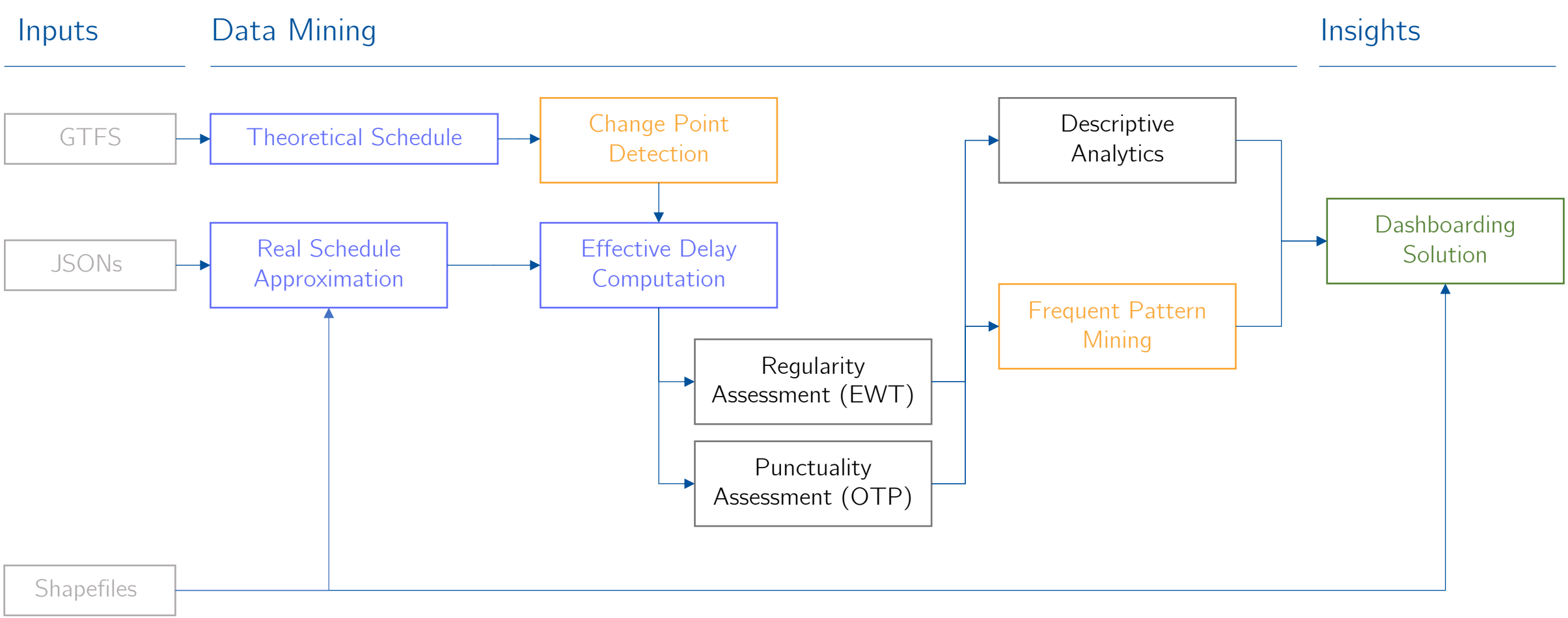


Figure – Methodology Flow chart

ECOLE POLYTECHNIQUE DE BRUXELLES

— Data Description

Key input data

ECOLE POLYTECHNIQUE DE BRUXELLES

— GTFS and JSON are key files to answer the problem

General Transit Feed Specification files (03/09)

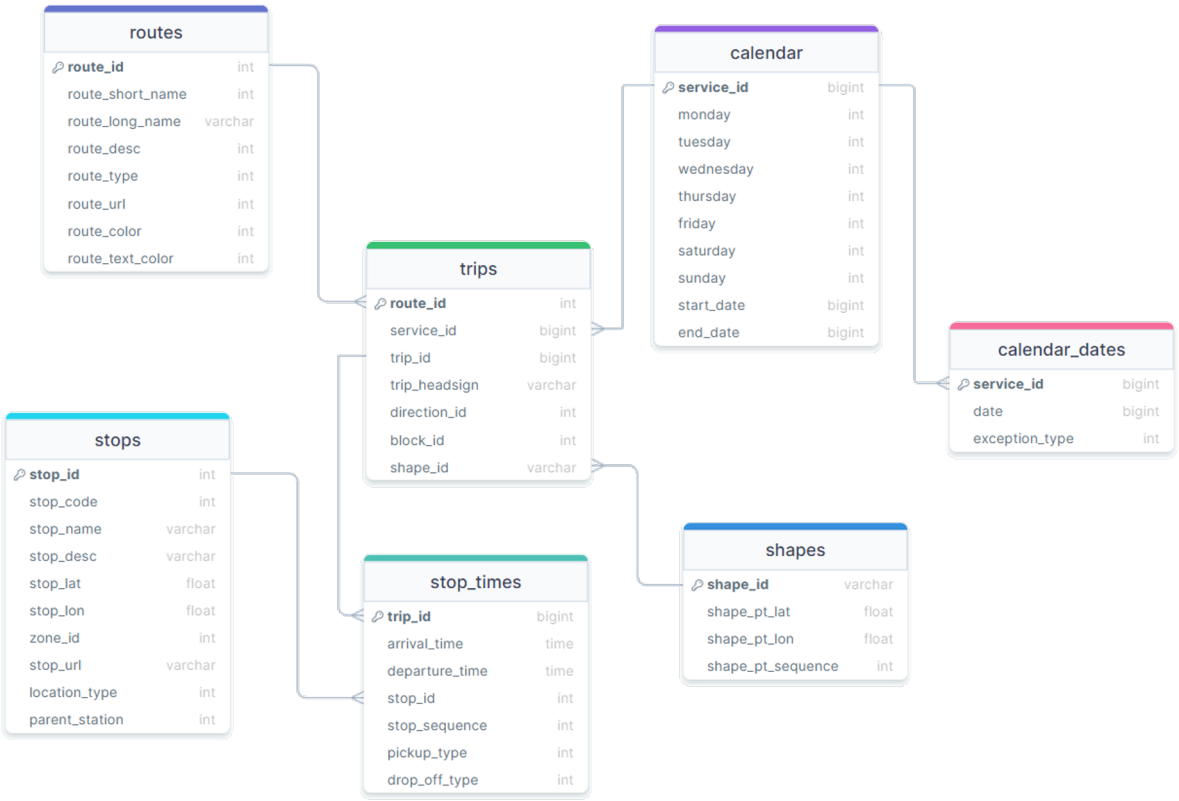


Figure – Entity Relationship Diagram

JSON files (06/09 – 21/09)

	date	time	route_short_name	distance_from_point	stop_id	stop_id_terminus
0	19-09-2021	10:33:18	1	1	8022	8161
1	19-09-2021	10:33:18	1	1	8732	8161
2	19-09-2021	10:33:18	1	1	8071	8731
3	19-09-2021	10:33:18	1	0	8152	8161
4	19-09-2021	10:33:18	1	1	8271	8731
...	...	...	...	...	...	...
19421878	14-09-2021	12:25:54	97	232	6109	6361
19421879	14-09-2021	12:25:54	98	0	3810	1951
19421880	14-09-2021	12:25:54	98	0	1951	2382
19421881	14-09-2021	12:25:54	98	270	3815	2382
19421882	14-09-2021	12:25:54	98	0	1951	2382

19421883 rows × 6 columns

Table – Flattened JSON: From 06/09/21 until 21/09/21

Shape files : Stops & Lines

	Code_Ligne	Variante	succession	stop_id	descr_fr	descr_nl	alpha_fr	alpha_nl	coord_x	coord_y	mode	numero_lig	terminus	geometry
0	012b	1	1	9600B	BRUSSELS AIRPORT	BRUSSELS AIRPORT	Brussels Airport	Brussels Airport	157950.0	176429.0	B	12	BRUSSELS CITY	POINT (157950.000 176429.000)
1	012b	1	2	3017	BOURGET	BOURGET	Bourget	Bourget	154334.0	174200.0	B	12	BRUSSELS CITY	POINT (154334.000 174200.000)
2	012b	1	3	5048	DA VINCI	DA VINCI	Da Vinci	Da Vinci	152934.0	173976.0	B	12	BRUSSELS CITY	POINT (152934.000 173976.000)
3	012b	1	4	2695	GENEVE	GENEVE	Genève	Genève	152428.0	172606.0	B	12	BRUSSELS CITY	POINT (152428.000 172606.000)

Table – Stops and lines spatial description

ECOLE POLYTECHNIQUE DE BRUXELLES

— Data Exploration

Key data processing steps

ECOLE POLYTECHNIQUE DE BRUXELLES

Flattening the GTFS for clarity

Merge GTFS tables

Step 1: Cleaning

- Stops: reformatting the *stop_ id* in order that it is equivalent in all the tables
- Lines: keeping only the used lines that appear in the observed data

Step 2: Merging the tables to obtain one single dataframe containing the relevant features

Key Point

→ No straightforward interconnection between the observed data and the theoretical schedule

Table illustration

	trip_id	arrival_time	departure_time	stop_id	stop_sequence	pickup_type	drop_off_type	stop_code	stop_name	stop_de
0	112947202236269500	07:30:56	07:30:56	8733	1	0	0	NaN	GARE DE L'OUEST	N
1	112947202236269500	07:31:53	07:31:53	8742	2	0	0	NaN	BEEKKANT	N
2	112947202236269500	07:33:33	07:33:33	8292	3	0	0	NaN	ETANGS NOIRS	N
3	112947202236269500	07:34:44	07:34:44	8282	4	0	0	NaN	COMTE DE FLANDRE	N
4	112947202236269500	07:36:08	07:36:08	8272	5	0	0	NaN	SAINTE-CATHERINE	N
...	...	...	...	...	...	...	...	...	...	
5686370	113354575237586501	15:26:44	15:26:44	5916	9	0	0	NaN	EGIDE VAN OPHEM	N
5686371	113354575237586501	15:27:53	15:27:53	5917	10	0	0	NaN	WAGON	N
5686372	113354575237586501	15:30:00	15:30:00	1985B	11	0	0	NaN	GLOBE	N
5686373	113354575237586501	15:32:28	15:32:28	5058	12	0	0	NaN	HEROS	N
5686374	113354575237586501	15:33:00	15:33:00	1951	13	0	0	NaN	HEROS	N

5686375 rows × 53 columns

Table – De-normalized GTFS

ECOLE POLYTECHNIQUE DE BRUXELLES

Inferring moving vehicles from observations

Observed Schedule challenge

The positions of the vehicles every 30 seconds do not provide the arrival and departure time at stops.

Extraction Heuristics

Extract the arrival and departure times of a vehicle at a defined stop, line, direction and day using

- Arrival time when distance from point at t-1 is superior to t
- Departure time when distance from point at t+1 is superior to t
- Time between t and t-1 is higher than 2 min
- Adjusting t based on the average speed of the transport type and the distance between the stops (STIB 2021 annual report)

Table illustration

	date	time	route_short_name	distance_from_point	stop_id__current	stop_id__terminus
8159118	19-09-2021	00:33:58	7	524	5770F	9051F
8159333	19-09-2021	00:34:29	7	904	5770F	9051F
8180741	19-09-2021	04:58:06	7	0	5770F	9051F
8180802	19-09-2021	04:58:37	7	0	5770F	9051F
8180870	19-09-2021	04:59:08	7	0	5770F	9051F
8180922	19-09-2021	04:59:40	7	159	5770F	9051F
8180994	19-09-2021	05:00:10	7	159	5770F	9051F
8181069	19-09-2021	05:00:41	7	509	5770F	9051F
8181115	19-09-2021	05:01:13	7	870	5770F	9051F
8184154	19-09-2021	05:18:19	7	4	5770F	9051F
8184275	19-09-2021	05:18:50	7	104	5770F	9051F
8184389	19-09-2021	05:19:21	7	195	5770F	9051F
8184516	19-09-2021	05:19:52	7	366	5770F	9051F
8184637	19-09-2021	05:20:23	7	693	5770F	9051F
8184766	19-09-2021	05:20:55	7	851	5770F	9051F
8188390	19-09-2021	05:33:47	7	0	5770F	9051F
8188568	19-09-2021	05:34:17	7	82	5770F	9051F

Table – Extracted observed schedule from JSON for Tram 7 at stop Van Praet on the 19th of September

STIB, Statistics 3, 2021.

ECOLE POLYTECHNIQUE DE BRUXELLES

Matching theoretical to observed data

Matching function blueprint

Match based on day, line, direction, stop

Schedule Search : $\mathrm{Min}(t_{obs}, t_{sched}) < headway_{cluster}$

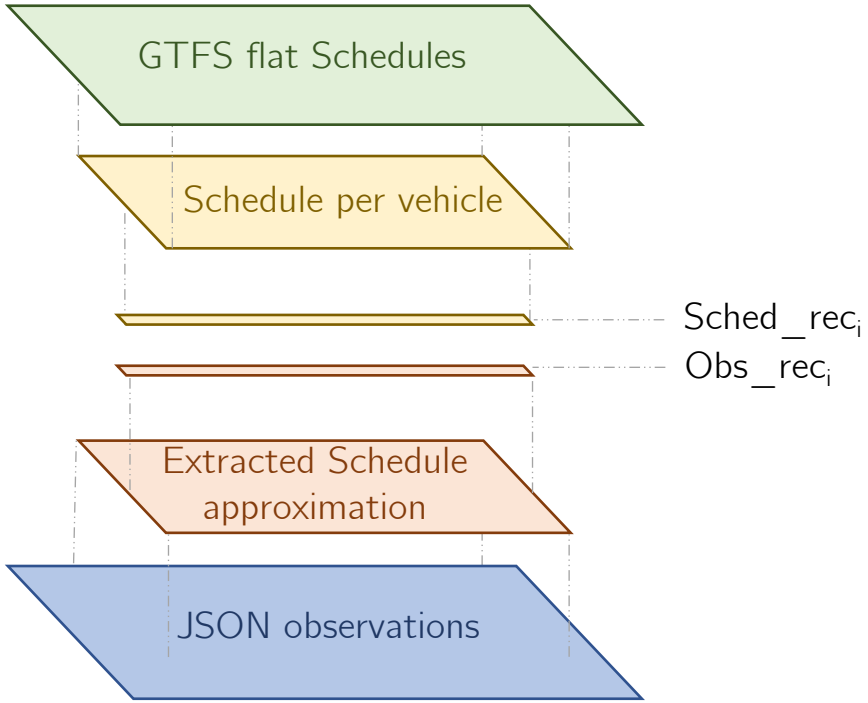


Table illustration

date_normalized	route_short_name	stop_name	departure_time	adjusted_arrival_time
09/09/2021	7	BOILEAU	05:53:02	05:55:15
09/09/2021	7	PETILLON	05:54:00	05:56:41
09/09/2021	7	ARSENAL	05:55:52	05:58:17
09/09/2021	7	VUB	05:57:09	05:59:03
09/09/2021	7	ETTERBEEK GARE	05:58:46	05:59:51
09/09/2021	7	ROFFIAEN	06:01:00	06:02:55
09/09/2021	7	BUYL	06:03:00	06:05:01
09/09/2021	7	CAMBRE-ETOILE	06:04:06	06:06:03
09/09/2021	7	LEGRAND	06:06:00	06:07:43
09/09/2021	7	BASCULE	06:07:05	06:08:07
09/09/2021	7	LONGCHAMP	06:09:00	06:10:16
09/09/2021	7	GOSSART	06:10:15	06:11:17
09/09/2021	7	CAVELL	06:11:19	06:12:43
09/09/2021	7	CHURCHILL	06:11:47	06:13:16
09/09/2021	7	CHURCHILL	06:12:00	06:13:48
09/09/2021	7	CHURCHILL	06:12:10	06:14:11
09/09/2021	7	VANDERKINDERE	06:13:00	06:15:24

Table – Matched schedule for Tram 7 for the first vehicle of the day on 9th of September

ECOLE POLYTECHNIQUE DE BRUXELLES

— Data mining observations

Defining time groups to assess regularity against punctuality and analysing problematic stops within the STIB network.

ECOLE POLYTECHNIQUE DE BRUXELLES

A time series approach to assess headway levels

Key challenge

- Traditional clustering methods won't best suit a linear data (headway as a function of time)

Pruned Extract Linear Time

Change Point Detection – PELT

- Unsupervised learning: groups unlabeled data according to their similarities or differences.
- Purpose: estimating the point at which the statistical properties of a sequence of observations change
- Parameter: (<penalty:int>) penalizes the inclusion of each additional change point by adding a penalty value to the cost of each segment and finding the segmentation with the smallest penalized cost among all possible segmentations.

Approach choice

For each obtained cluster

- Computing the headway median
- Threshold: 12 min
  - Median > 12 min : Punctuality
  - Median < 12 min : Regularity

Killick, R., Fearnhead, P., and Eckley, I.A. (2012). "Optimal Detection of Changepoints With a Linear Computational Cost." Journal of the American Statistical Association. 107. 1590-1598. https://dx.doi.org/10.1080/01621459.2012.737745

PELT, fast and reliable linear cluster detection

Clusters Detection and Metric Assignment Evaluation

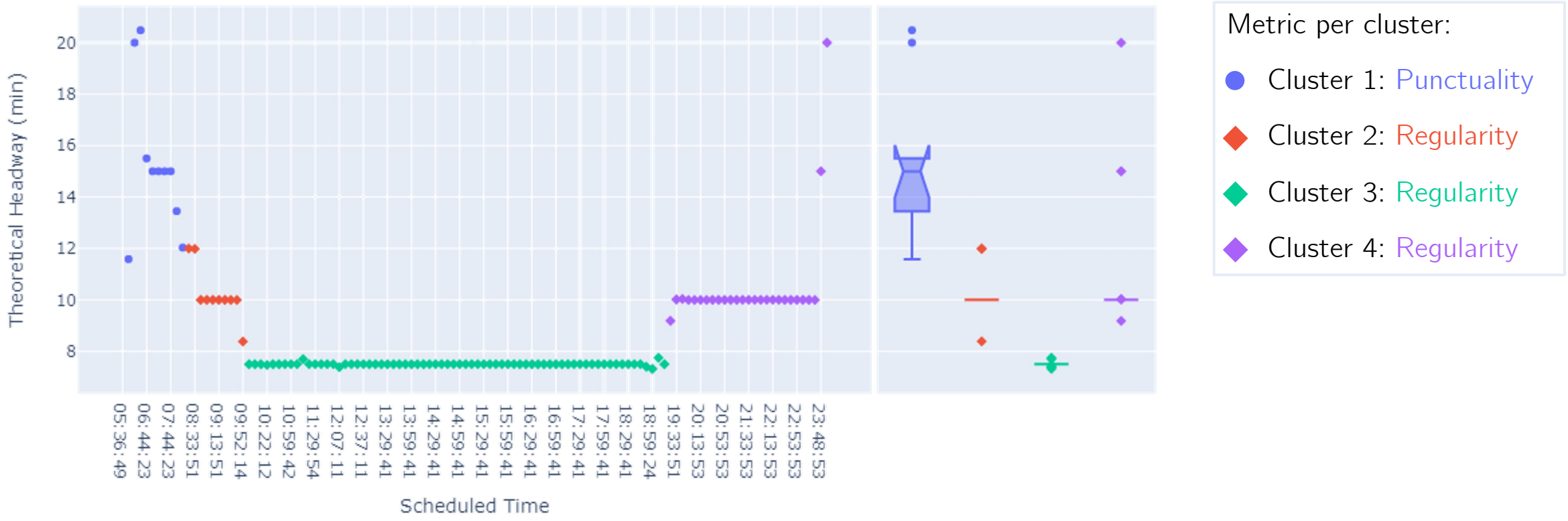


Figure – Headways for Metro 1 at stop de Brouckère in direction of Gare de l'Ouest on the weekend

OTP, the punctuality metric to evaluate

Delays

Delay computation

- For each punctuality clusters,
- Compute the delay between the scheduled arrival time and the real arrival time

Key Insight

- OTP: % observations that are labelled as on time according to their threshold function

In advance, On time or Late

Three threshold definitions

- Strict: the minute of the arrival time is different than in the scheduled timetable
- Moderate: |delay| is greater than 1 min
- Loose: |delay| is greater than 2 min

	ETA	TA	Strict	Moderate	Loose
Vehicule_1	10:32:45	10:33:01	1	0	0
Vehicule_2	10:35:45	10:36:46	1	1	0
Vehicule_3	10:40:45	10:42:46	1	1	1
OTP			0%	33%	66%

Table – Threshold impact on OTP illustration

ECOLE POLYTECHNIQUE DE BRUXELLES

Buses are worst performers, metros the best

OTP (%) per transport type and hour of the day

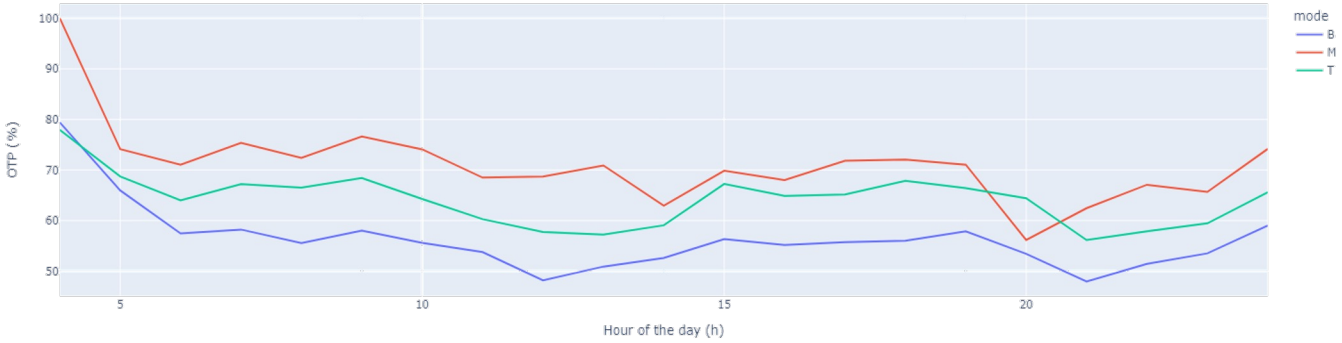


Figure – OTP per transport type throughout the day for loose definition of the delay

Key Insight

Transport type specific

- The Bus has the worst ratio whatever the hour in the day with an OTP median of 53%
- The Metro is performing the best with an OTP median of 65%
- The Tram follows a quite similar trend that the bus but with a slight better performance with an OTP median of 61%

Hourly Specific

- Wide dip of OTP median during the day and the evening

The delay is negatively correlated with OTP (%)

Insights

Figure – OTP and median delays for all lines

ECOLE POLYTECHNIQUE DE BRUXELLES

EWT, the regularity metric used

Excess Waiting Time (EWT)

EWT computation

- For each regularity clusters:
  - Compute the real headways
  - Compute the schedule waiting time
  - Compute the actual waiting time
  - Measure the EWT

A cluster is defined as irregular when the EWT is greater than 1 min

EWT Evolution through a line

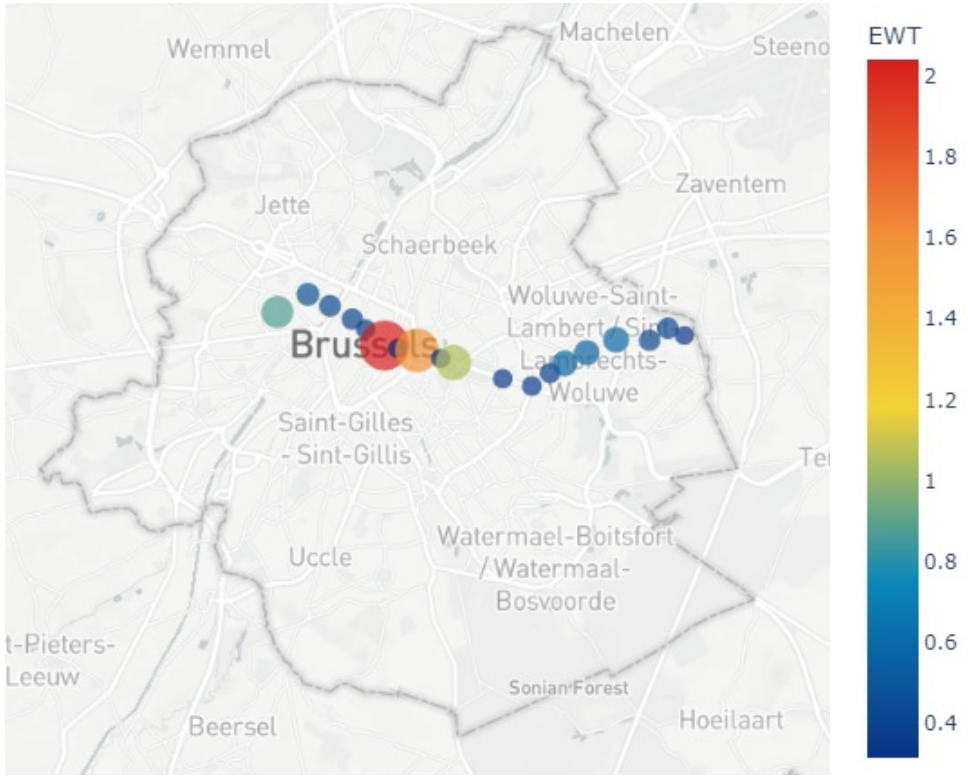


Figure – Median EWT at each stop for Metro 1 in direction of Stockel on saturdays

The reality is fairly different

Theoretical and observed headways comparison

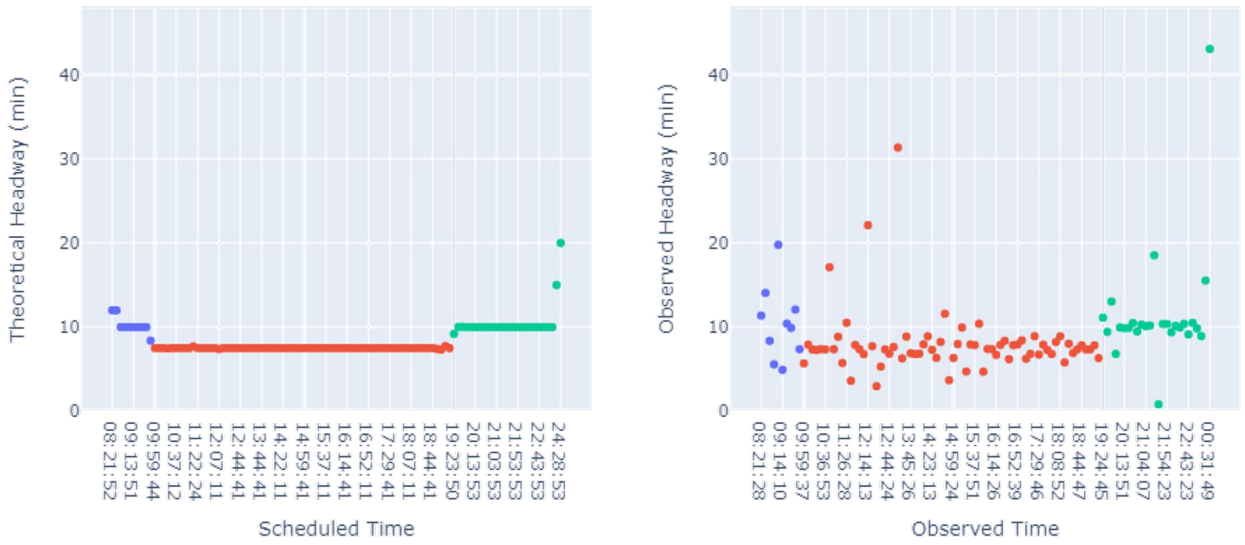


Figure – Theoretical and observed headways for Metro 1 at the stop de Brouckère in direction of Gare de l'Ouest on the 11th of September

Key Insight

EWT per cluster:

- Cluster 1: 1min25sec
- Cluster 2: 1min55sec
- Cluster 3: 4min11sec

Delays do propagate through the line

Problematic Segment Identification – FP Growth

1. Identify trends in the delays for each line, direction and day

- Fit the data into a polynomial to better perceive the growth and declines in the delays
- A stop is defined as
  - *delayer* when delays are growing
  - *catcher* when delays are decreasing

2. FP Growth

- Transaction
  - Delayers : sequence of stop growing the delay aggregated on the median per day per line
  - Catcher : sequence of stop decreasing the delay aggregated on the median per day per line
- Transactions : [[$stop_1$, $stop_3$, $stop_4$, $stop_6$], [$stop_3$, $stop_4$]]
- Analysed per transportation type

Median delay through the line

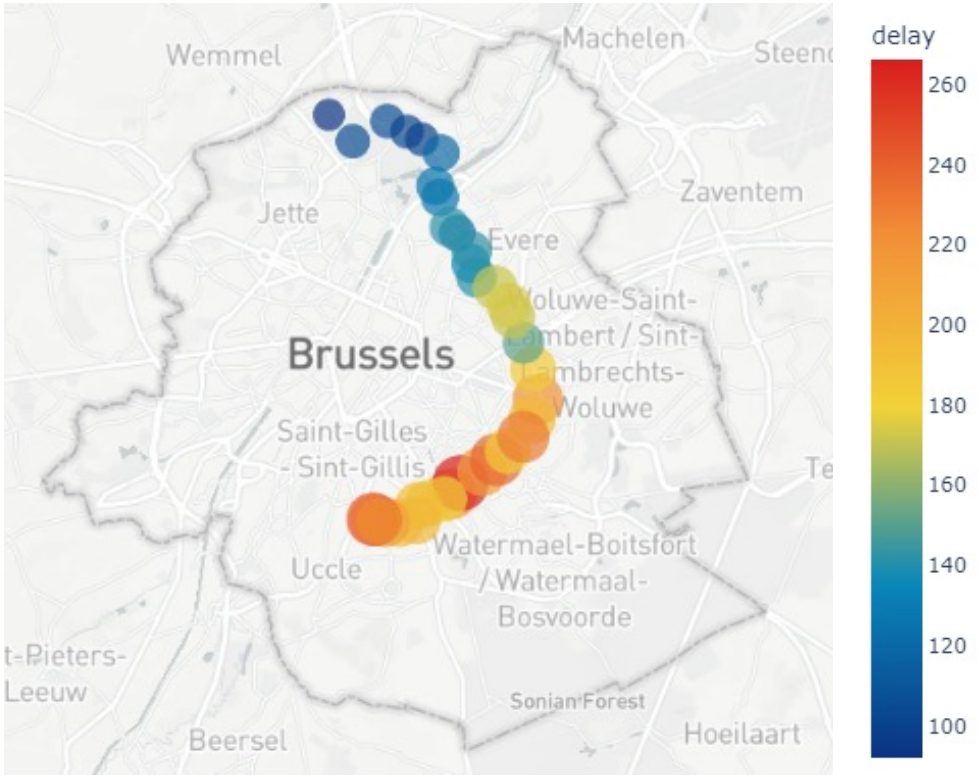


Figure – Median delay at each stop for Tram 7 in direction of Vanderkindere on workdays

Similar pattern can be inferred from trips

Daily median delay for a line

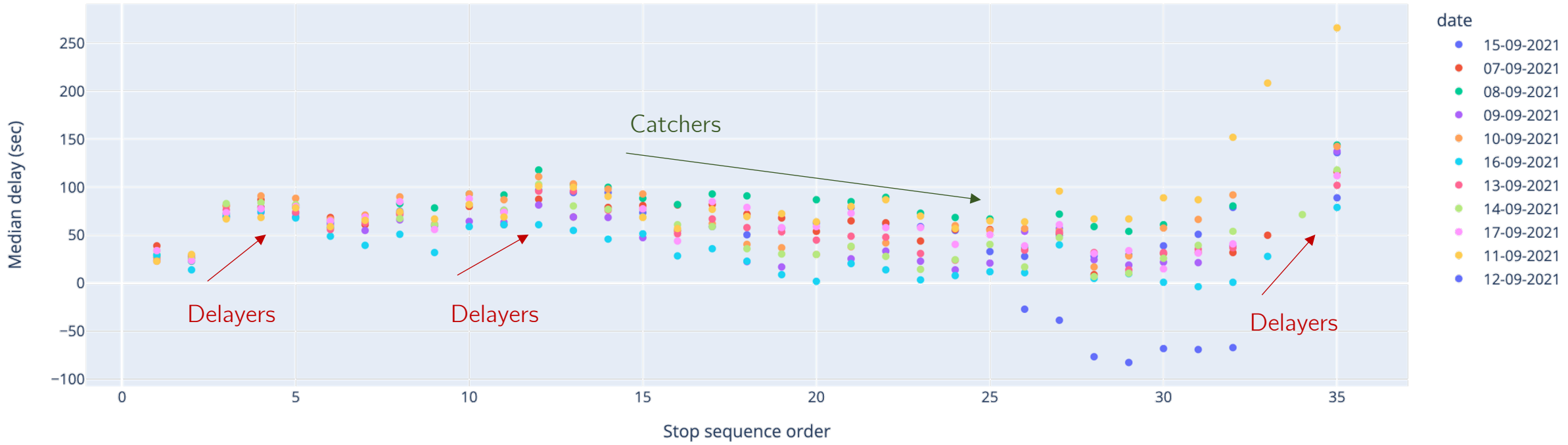


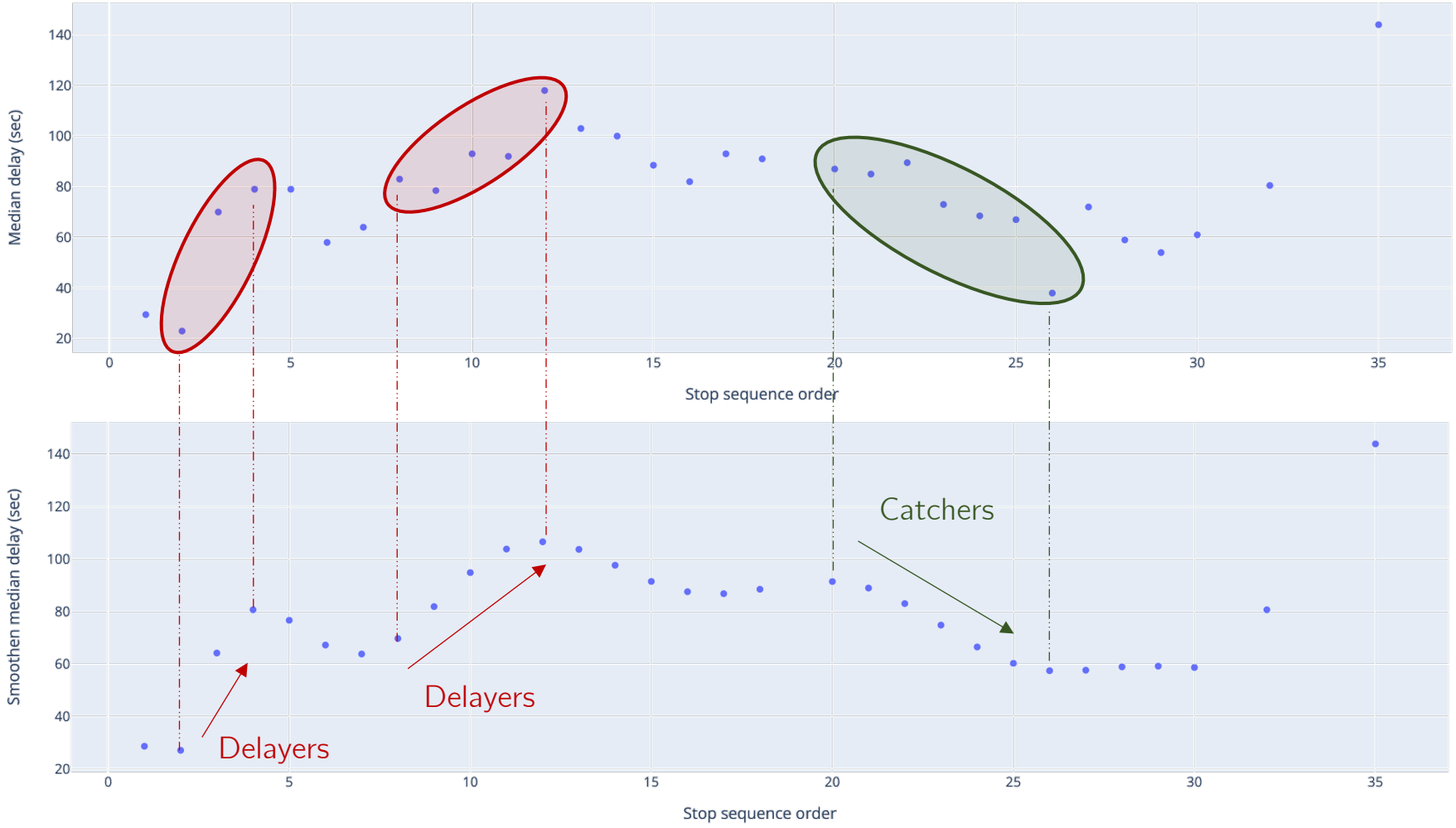
Figure – Median delays at each succession for Tram 7 in direction of Vanderkindere

Polynomial fitting helps at detecting trends

Polynomial Fitting

Figure – Median delays at each succession for Tram 7 in direction of Vanderkindere on the 8th of September

Figure – Median delays after polynomial fitting



ECOLE POLYTECHNIQUE DE BRUXELLES

Frequent patterns highlights problematic stops across the STIB network

Top delay offender and rescuer for trams

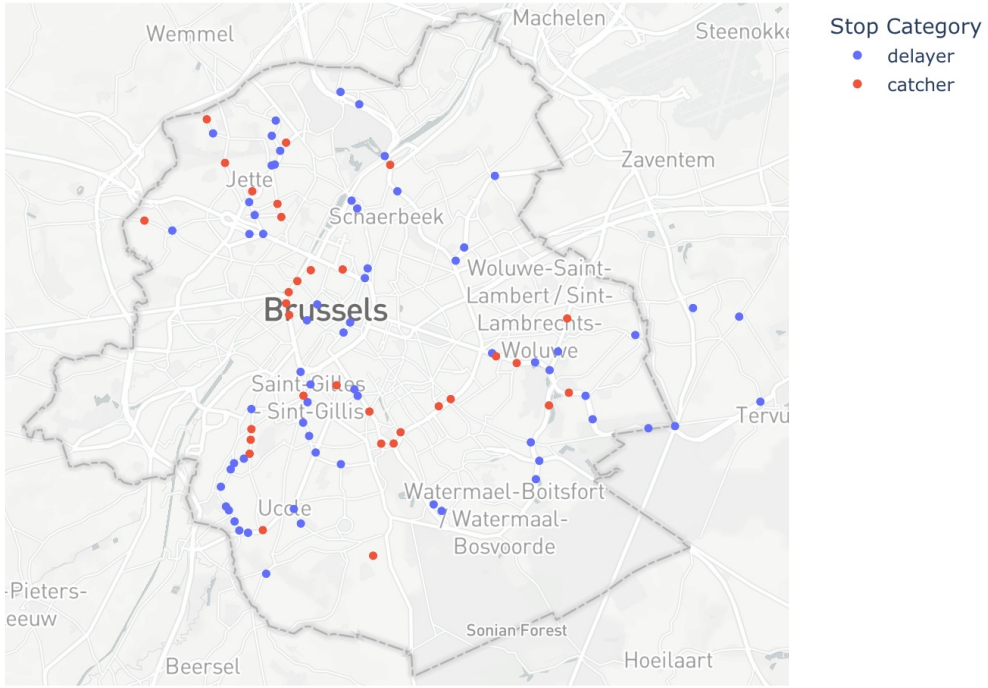


Figure – Frequent patterns in delays (tram)

Top regularity offender and rescuer for trams

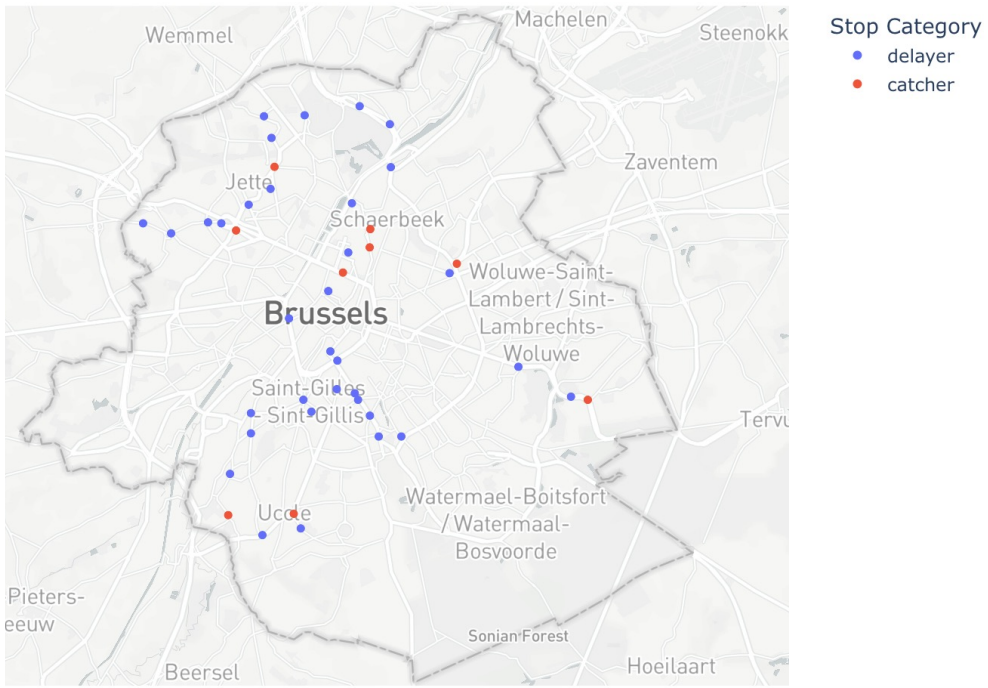


Figure – Frequent patterns in irregularity (tram)

ECOLE POLYTECHNIQUE DE BRUXELLES

— Conclusion

ECOLE POLYTECHNIQUE DE BRUXELLES

Specific road sections must be investigated

Punctuality

- With a loose threshold, the STIB network has an OTP of 54% on the punctuality clusters
- 35% of trips experience at least 2 min of delays at each stops passed by.

Regularity

- Key rule : EWT >= 1 min → cluster is irregular
- The network is experiencing only 30% of regular clusters.

Propagation

Delay

- Delays do propagate on stops but vehicles are also able to catch their delays at some other stops.
- Metros are extremely good at not propagating too much. It is often due to their low delays.
- Trams can catch delays more effectively probably due to specific lane allocation.
- Buses are experiencing delays in the suburbs rather than in the center. It seems that, in the city center, there is enough time between stops to catch delays.

EWT

- The irregularity can be propagated but is rarely caught up

— Dashboard time !

ECOLE POLYTECHNIQUE DE BRUXELLES

— Open to Questions

Thank you for your listening !

ECOLE POLYTECHNIQUE DE BRUXELLES

Sources used in the presentation

Bibliography

Bibliography

STIB Annual Report, Statistics 3, 2021.

Killick, R., Fearnhead, P., and Eckley, I.A. (2012). "Optimal Detection of Changepoints With a Linear Computational Cost." Journal of the American Statistical Association. 107. 1590-1598. https://dx.doi.org/10.1080/01621459.2012.737745

ECOLE POLYTECHNIQUE DE BRUXELLES

— Appendix

Matching theoretical to observed data

Matching challenges

Issue 1: Linking dates with service_id

- Create a label for the day in both tables: "workday, Saturday, Sunday"
- Define a date window that takes the days in the service_id with the right label
- Replicate in the schedule the observations for each date in the date window

Issue 2: Matching the schedule with the observed data

- One trip is considered as a vehicle, therefore for each trip,
- And for each row, the observation t in the schedule, we look in the observed data the record that is the closest to t
- Adjusting the matched record

Table illustration

date_normalized	route_short_name	stop_name	departure_time	adjusted_arrival_time
09/09/2021	7	BOILEAU	05:53:02	05:55:15
09/09/2021	7	PETILLON	05:54:00	05:56:41
09/09/2021	7	ARSENAL	05:55:52	05:58:17
09/09/2021	7	VUB	05:57:09	05:59:03
09/09/2021	7	ETTERBEEK GARE	05:58:46	05:59:51
09/09/2021	7	ROFFIAEN	06:01:00	06:02:55
09/09/2021	7	BUYL	06:03:00	06:05:01
09/09/2021	7	CAMBRE-ETOILE	06:04:06	06:06:03
09/09/2021	7	LEGRAND	06:06:00	06:07:43
09/09/2021	7	BASCULE	06:07:05	06:08:07
09/09/2021	7	LONGCHAMP	06:09:00	06:10:16
09/09/2021	7	GOSSART	06:10:15	06:11:17
09/09/2021	7	CAVELL	06:11:19	06:12:43
09/09/2021	7	CHURCHILL	06:11:47	06:13:16
09/09/2021	7	CHURCHILL	06:12:00	06:13:48
09/09/2021	7	CHURCHILL	06:12:10	06:14:11
09/09/2021	7	VANDERKINDERE	06:13:00	06:15:24

Table – Matched schedule for Tram 7 for the first vehicle of the day on 9th of September