PondiônsTracker: A framework based on GTFS-RT to identify delays and estimate arrivals dynamically in Public Transportation Network

Pedro Pongelupe Lopes

Programa de Pós-Graduação em Informática

December 04 2023

PUC Minas

Contents

PUC Minas

Introduction

Motivation

- Public Transportation Network
- Huge centers
- GTFS and GTFS-RT specifications

PUC Minas

Motivation

GTFS-RT Matching Identifiers Issue

To work with GTFS-RT, it is **required** to track vehicles in real-time. But, in some cases, it is not easy to match the identifier between a real-time record and the GTFS static data. In Rome in 2016, this issue was reported by Raghothama et al. (2016). We still face this issue in Belo Horizonte in 2023.

Objectives

Main Objective

- Proposing and validating PondiônsTracker

Specific Objectives

- Collecting data from the real-time API and combining with the GTFS
- Understanding if Belo Horizonte's delays are spatial and temporal dependent
- Comparing the arrival times defined at the GTFS with the ones generated by *PondiônsTracker*.

PUC Minas

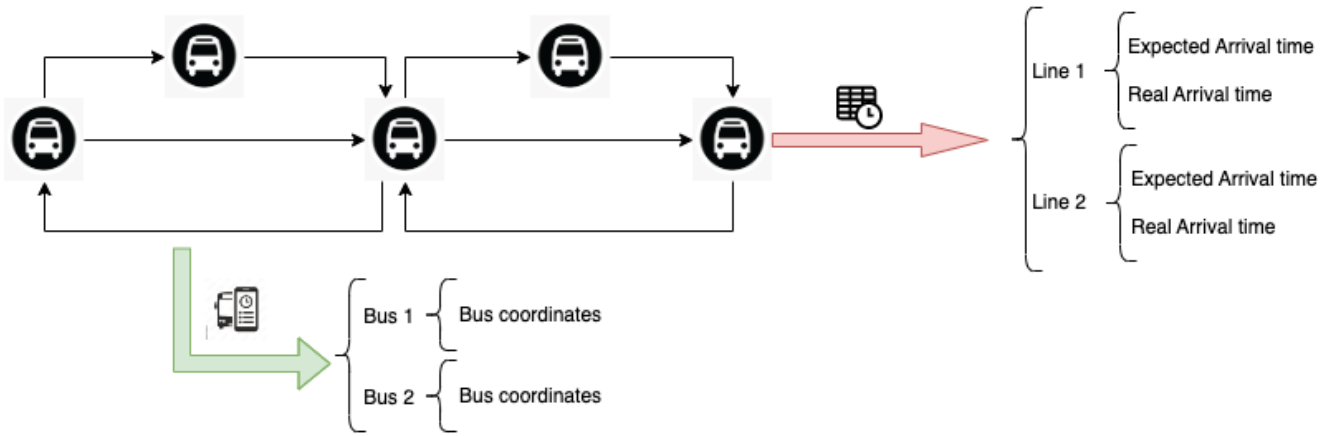
Theoretical Reference

Main Ideas

- Smart Cities
- Urban Computing
- Human Mobility
- Graphs and Complex Network

Public Transportation Network as a Complex Network

$$G^{l} = (V_g, E_g, X_g, A_g)$$



G^{l}: Graph G at a given time l

V_g: Bus stops

E_g: Routes connecting two bus stops

X_g: Additional information about bus stops (V_g)

A_g: Additional information about routes connecting two bus stops (E_g)

PUC Minas

PondiônsTracker

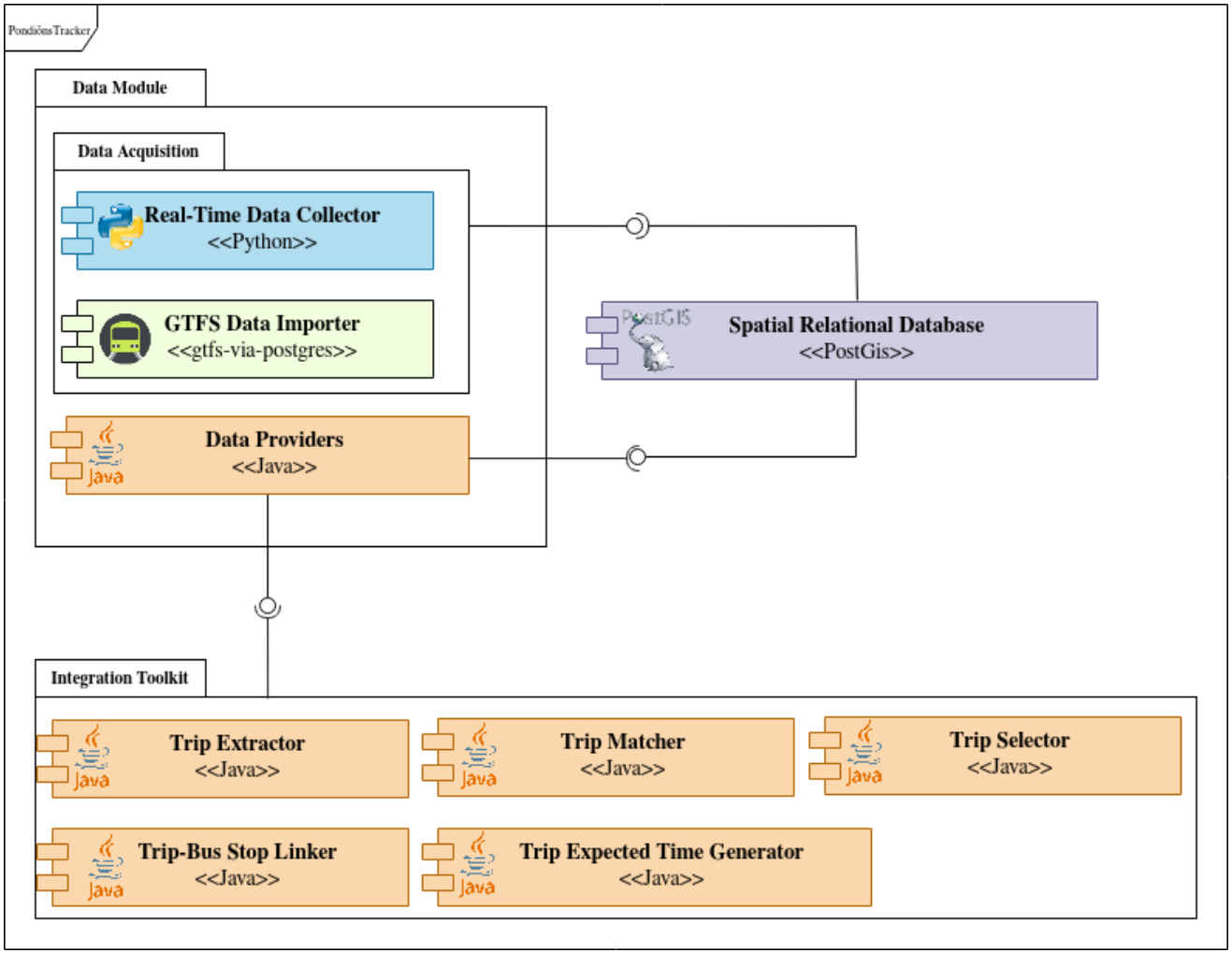
Overview

PondiônsTracker[a] is a framework to enrich GTFS data with real-time data. The name *PondiônsTracker* is a small gag from the sonority of the expression *bus stop* when pronounced in Portuguese with the accent from Minas Gerais.

[a] Available at `https://github.com/Pongelupe/PondionsTracker/`

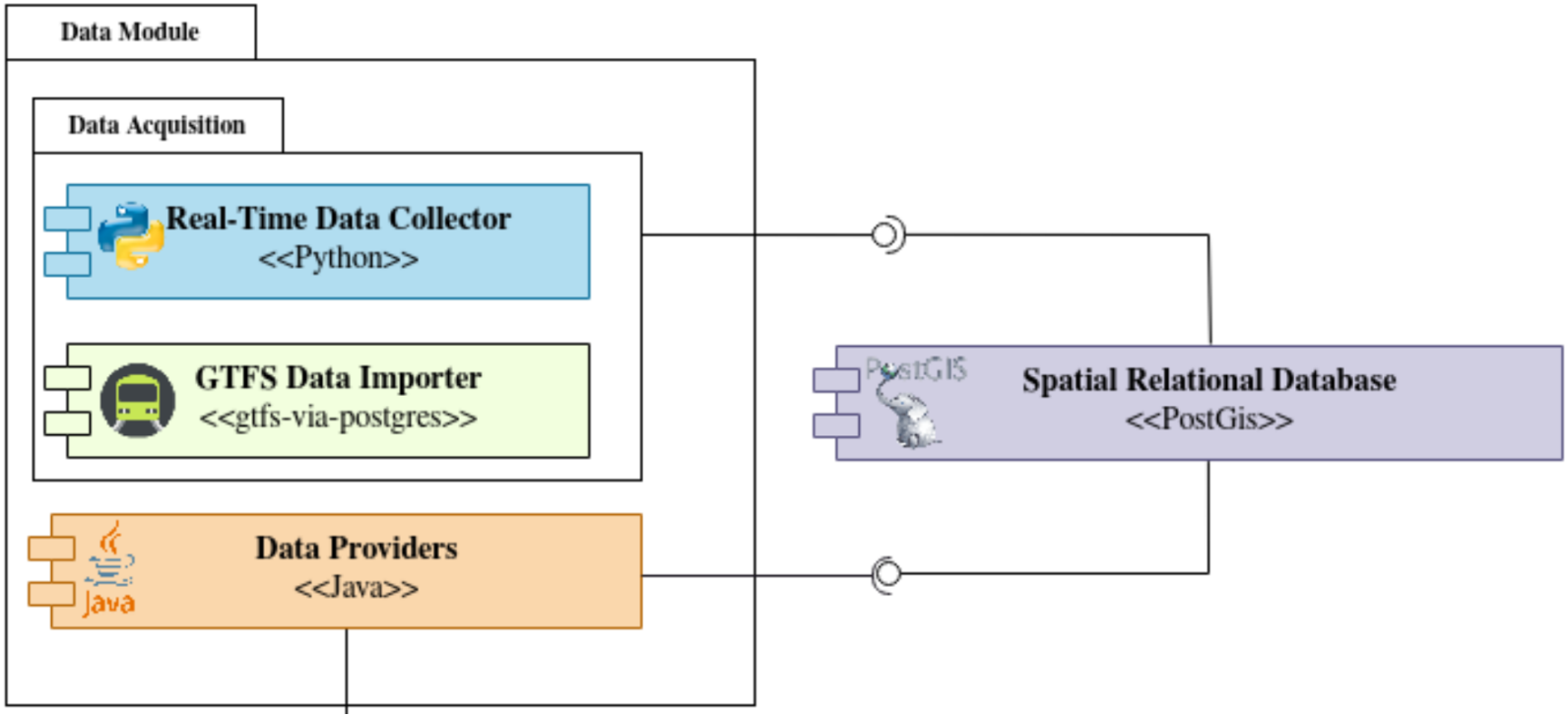
PUC Minas

PondiônsTracker's Architecture



PUC Minas

Data Module Overview



PUC Minas

Data Providers

```
1  <dependency>
2    <groupId>br.pondionstracker</groupId>
3    <artifactId>data-module</artifactId>
4    <version>1.0.0</version>
5  </dependency>
```

Figura: *DataModule*'s maven dependency

Integration Module

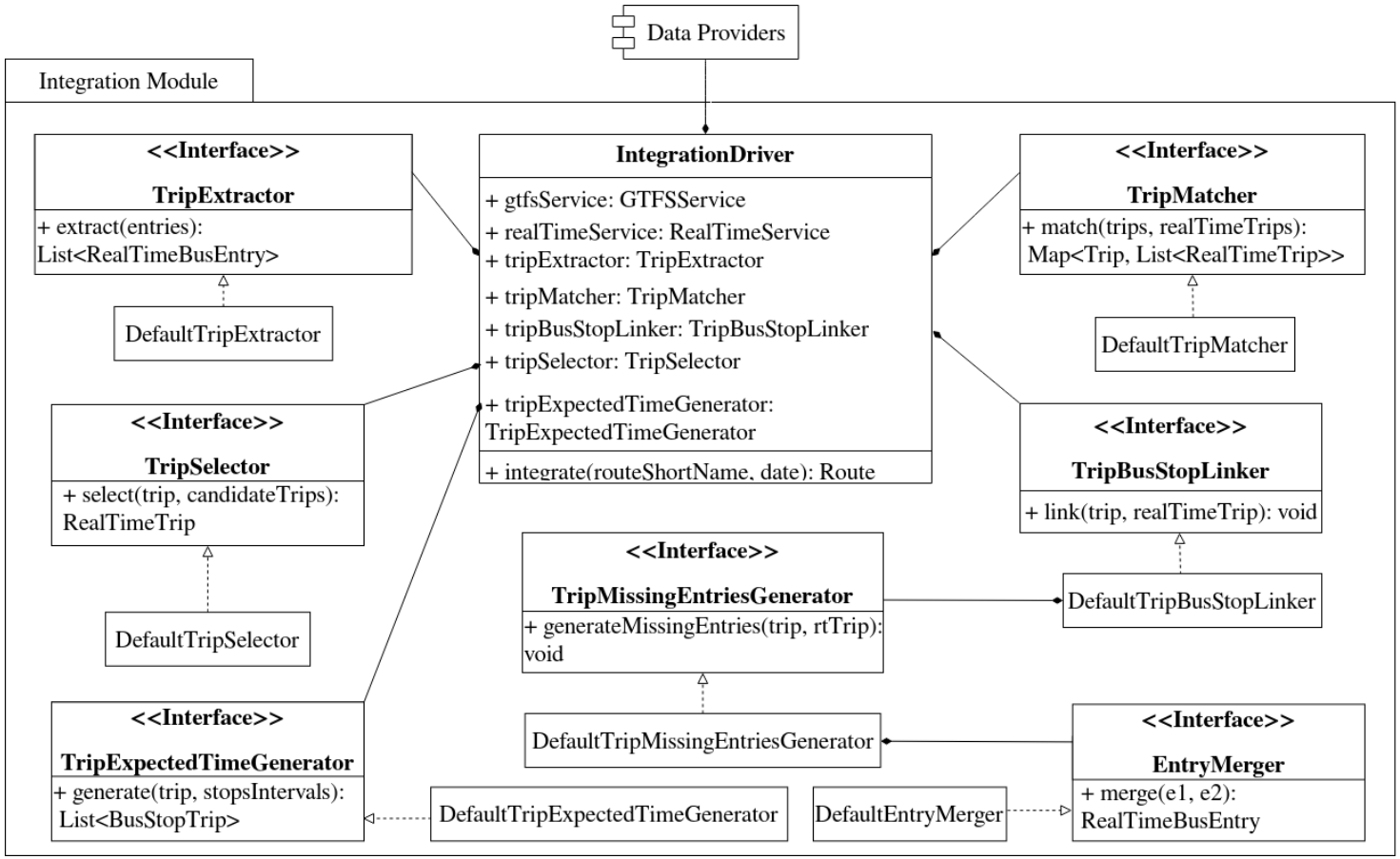


Figura: Integration Module Class Diagram

PUC Minas

Integration Driver

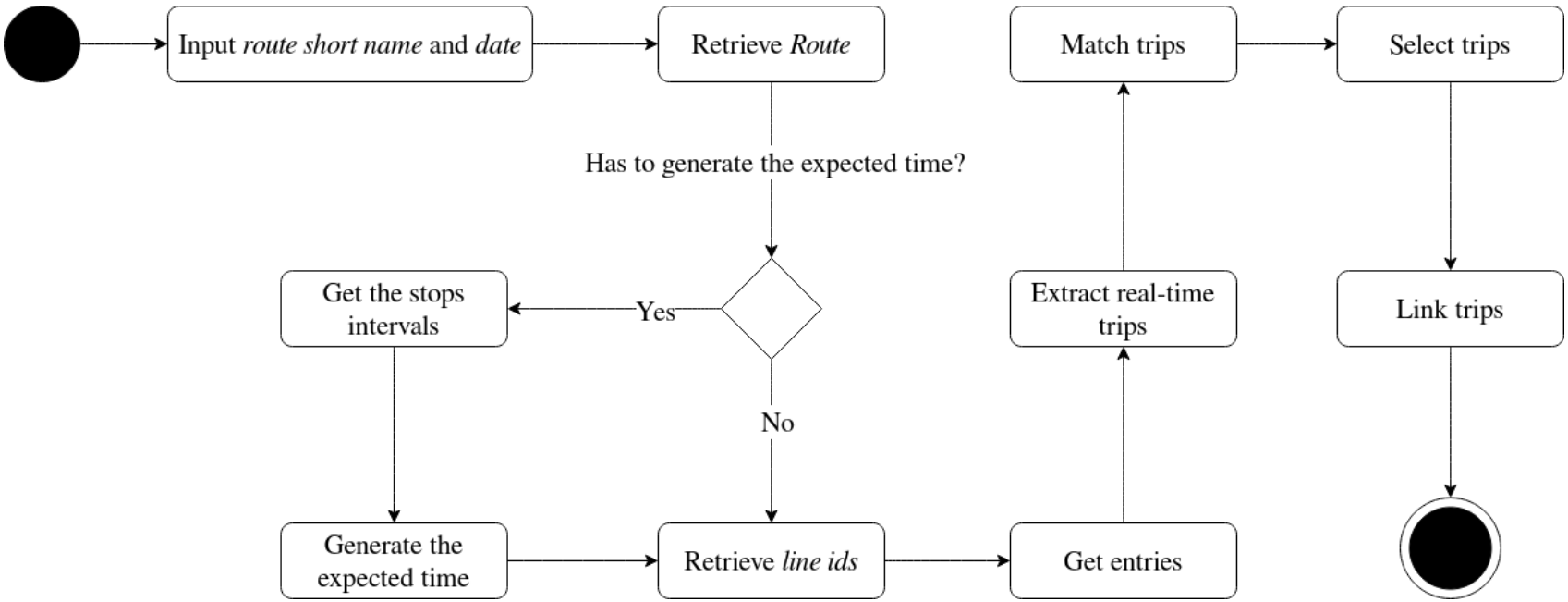


Figura: Integration Driver Activity Diagram

Integration Driver - 1st Step

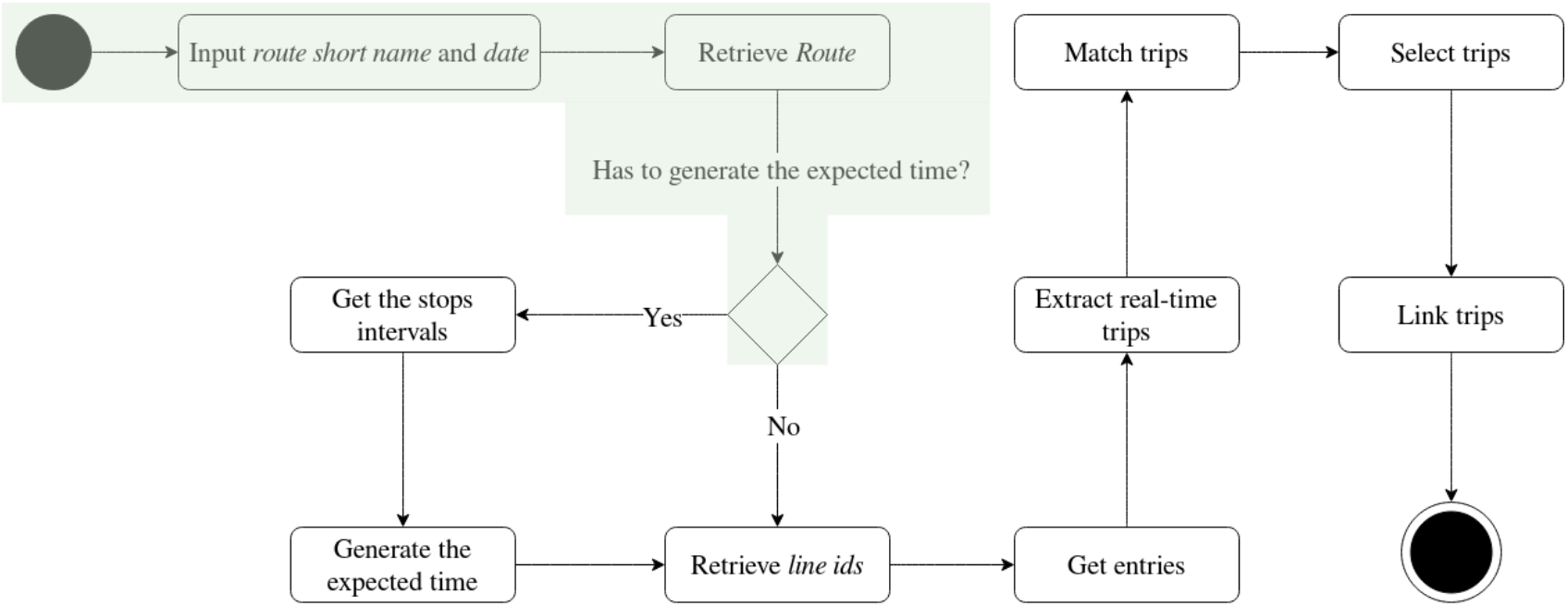


Figura: Integration Driver Activity Diagram

PUC Minas

Has to generate the expected time?

`arrival_time`	Time	Conditionally required	Arrival time at a specific stop for a specific trip on a route. If there are not separate times for arrival and departure at a stop, enter the same value for `arrival_time` and `departure_time`. For times occurring after midnight on the service day, enter the time as a value greater than 24:00:00 in HH:MM:SS local time for the day on which the trip schedule begins. Scheduled stops where the vehicle strictly adheres to the specified arrival and departure times are timepoints. If this stop is not a timepoint, it is recommended to provide an estimated or interpolated time. If this is not available, `arrival_time` can be left empty. Further, indicate that interpolated times are provided with `timepoint=0`. If interpolated times are indicated with `timepoint=0`, then time points must be indicated with `timepoint=1`. Provide arrival times for all stops that are time points. An arrival time must be specified for the first and the last stop in a trip.

Figura: *arrival_time* definition from *stop_times.txt*

Integration Driver - 2nd Step*

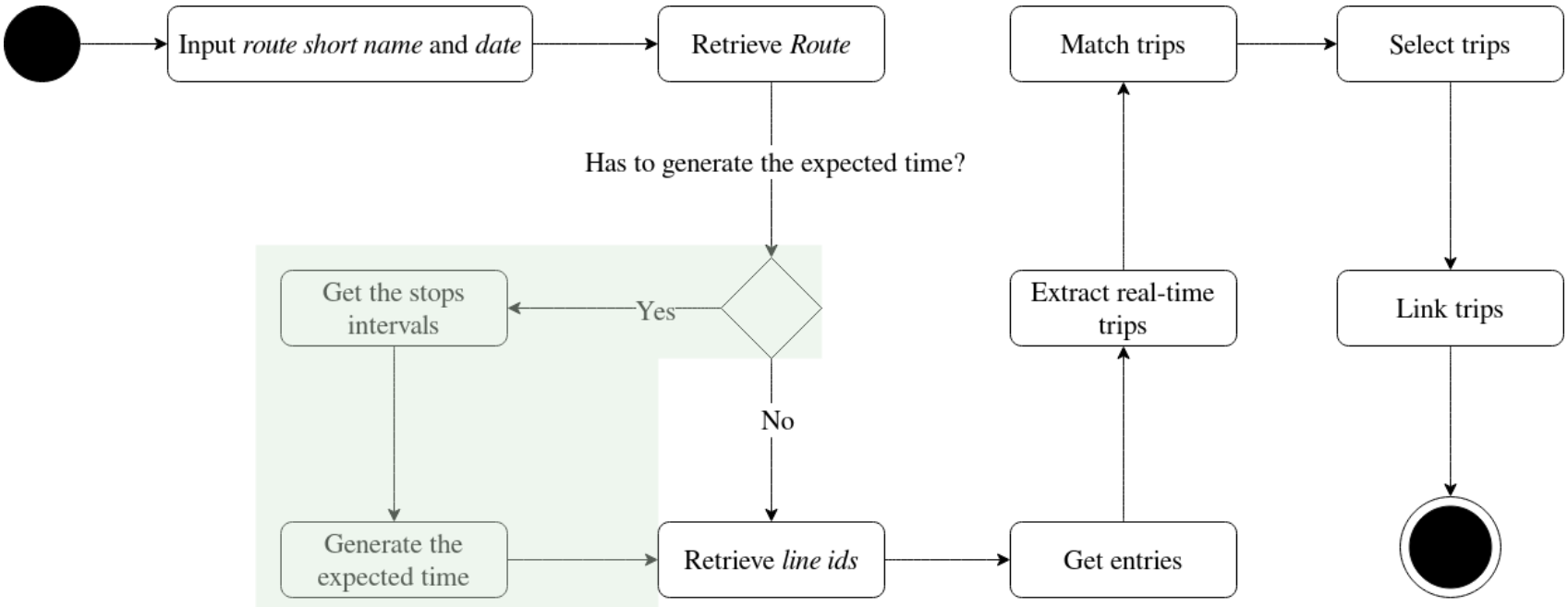


Figura: Integration Driver Activity Diagram

PUC Minas

Integration Driver - 3rd Step

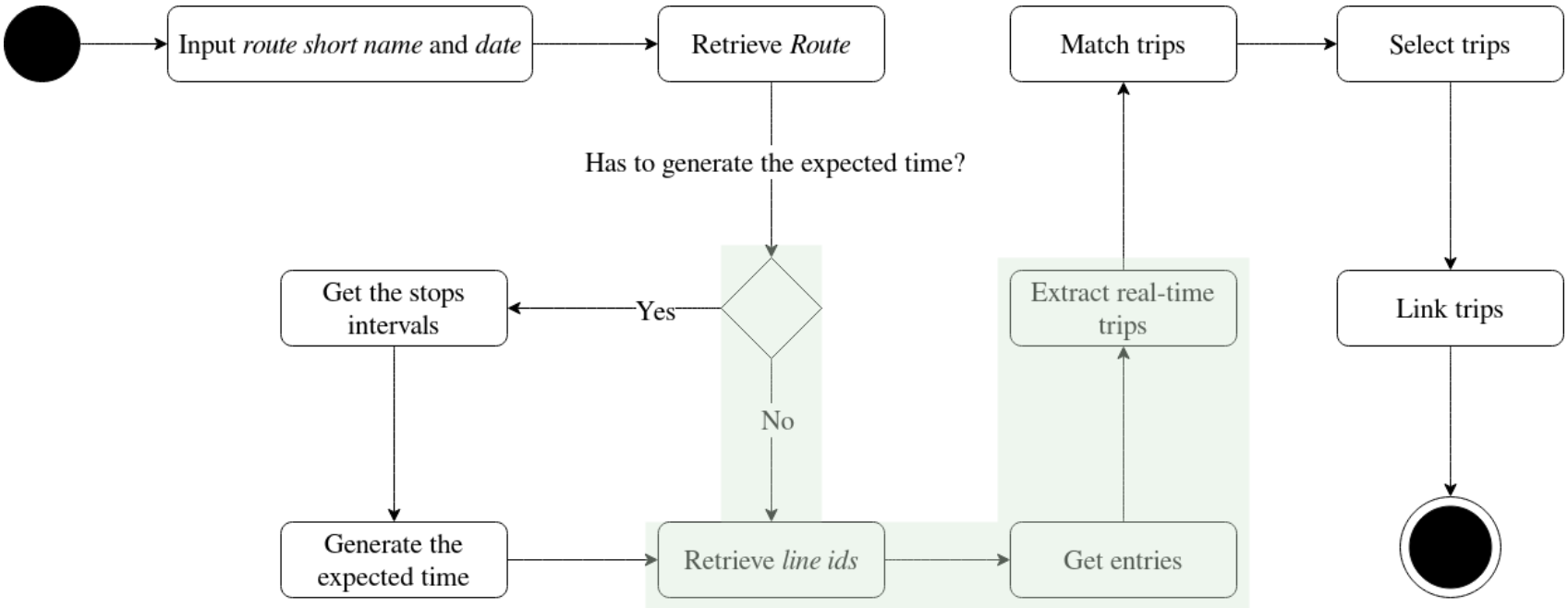


Figura: Integration Driver Activity Diagram

PUC Minas

Integration Driver - 4th Step

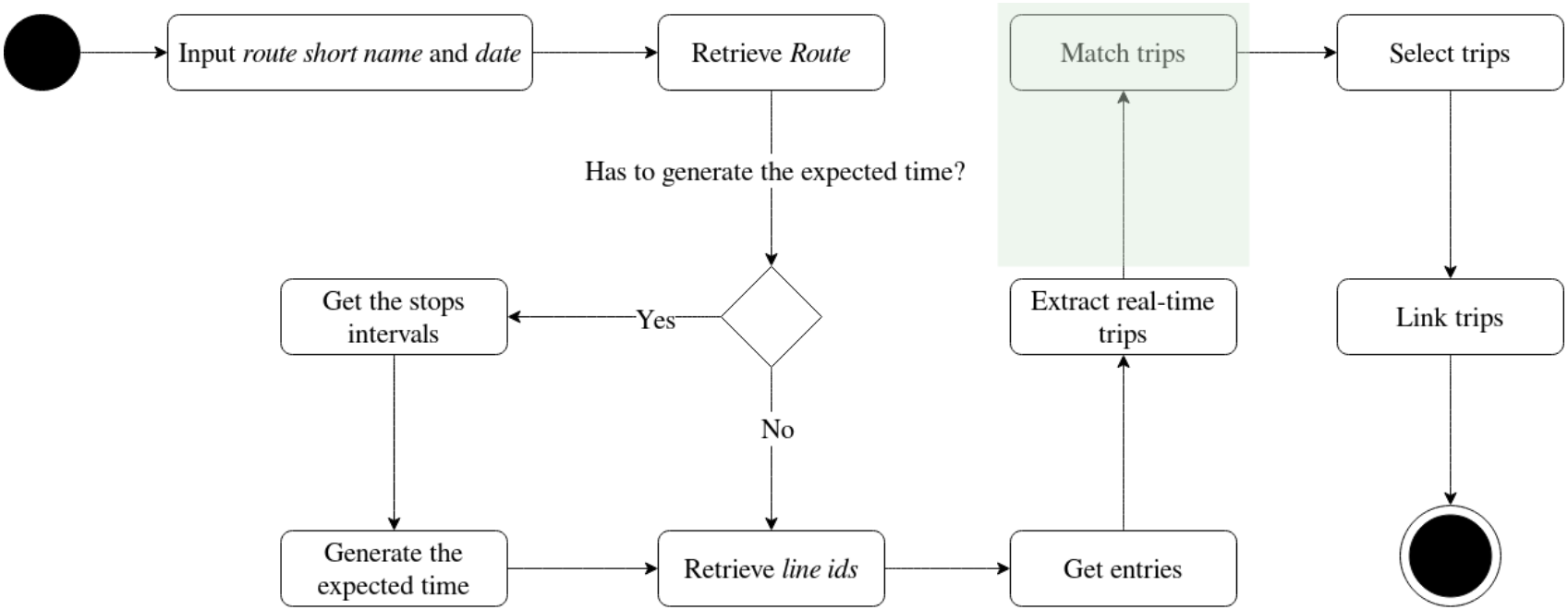


Figura: Integration Driver Activity Diagram

PUC Minas

Integration Driver - 5th Step

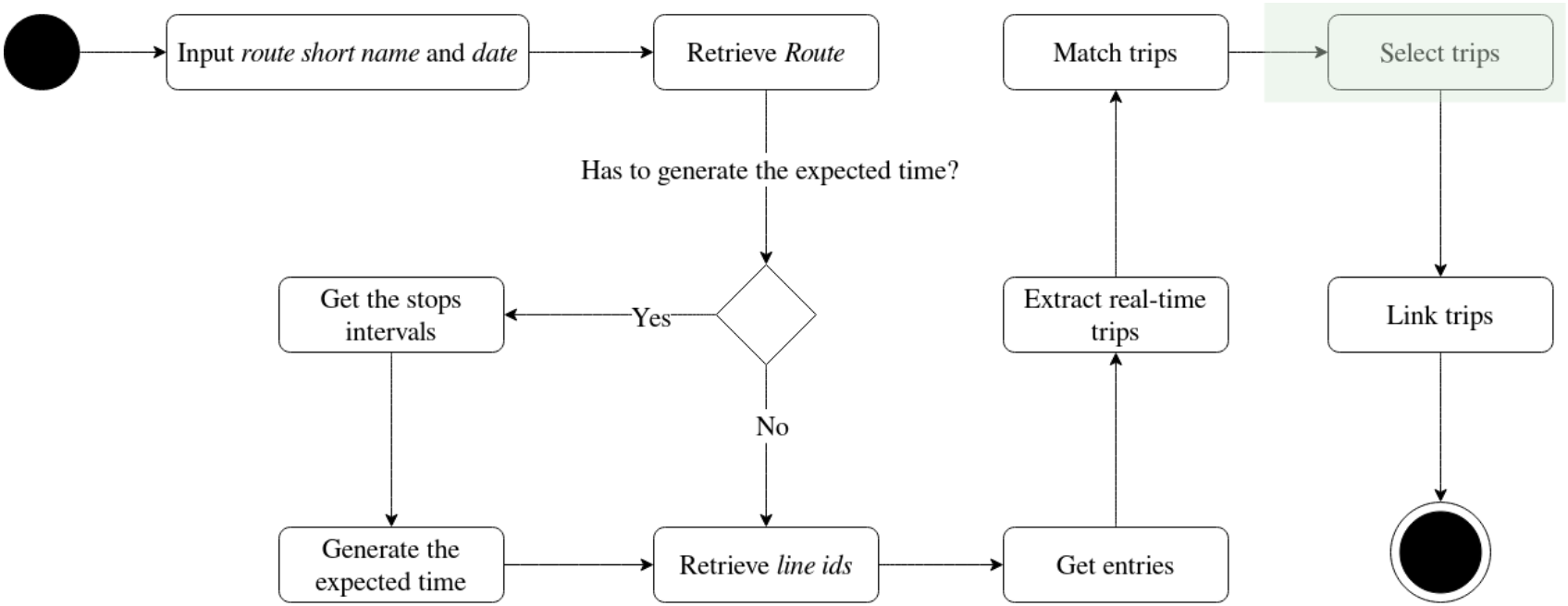


Figura: Integration Driver Activity Diagram

PUC Minas

Integration Driver - 6th Step

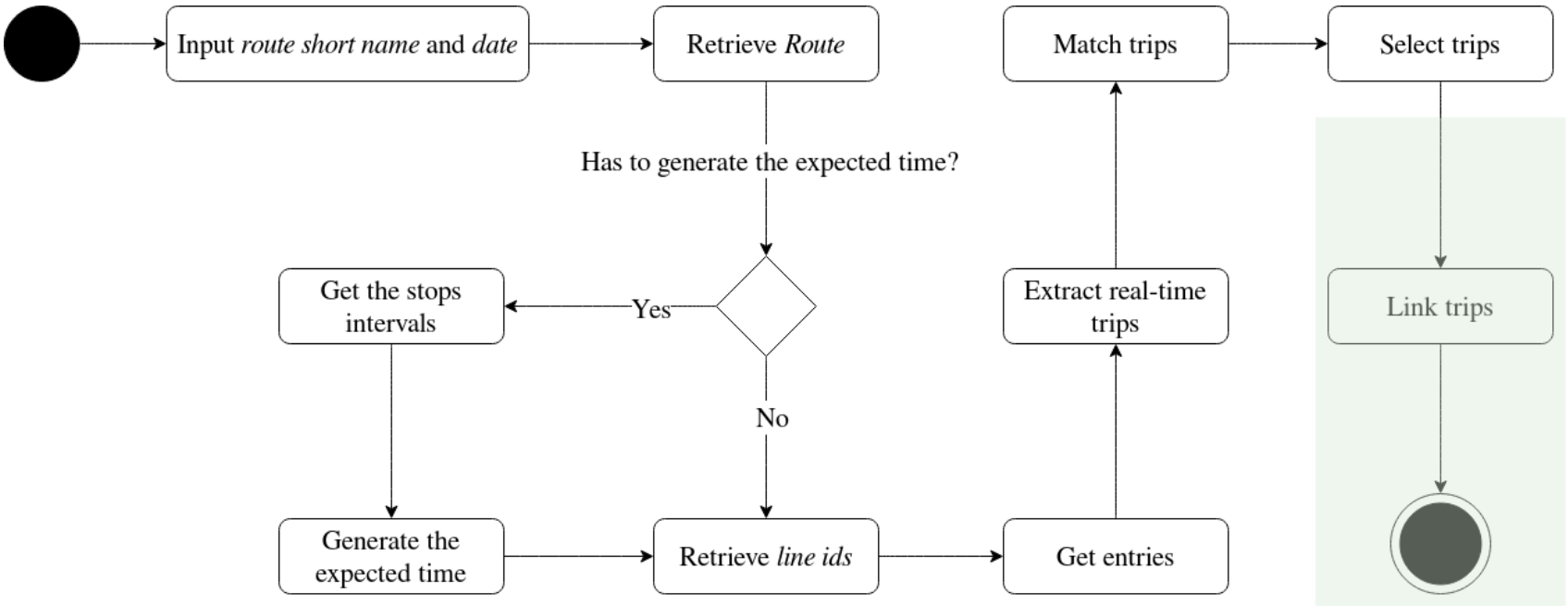


Figura: Integration Driver Activity Diagram

PUC Minas

Integration Module

```
1  <dependency>
2    <groupId>br.pondionstracker</groupId>
3    <artifactId>integration-module</artifactId>
4    <version>1.0.0</version>
5  </dependency>
```

Figura: *IntegrationModule*'s maven dependency

PondiônsTracker-BH

Overview

PondiônsTracker-BH[a] is our *PondiônsTracker*'s specialization to deal with Belo Horizonte's Network. So, we have implemented our own *Real-Time Data collector*, and we have overwritten a method from the *RealTimeService* from the *DataProviders*.

[a] Available at `https://github.com/Pongelupe/PondionsTracker-BH`

PUC Minas

Belo Horizonte's RealTimeService

BHRealTimeService - getIdsLineByRouteId

Due to a **one-to-many** relationship between the GTFS and the real-time data.

- *Transfacil* $\rightarrow$ *Traffic API*
- *BHTrans* $\rightarrow$ *GTFS*

PUC Minas

Workload Overview

Workload

- Data collected for 11 days straight in August 2023
- 30 Gigabytes

Date	Day-of-Week	Entries
29-07-23	Saturday	22,319,765
30-07-23	Sunday	22,635,117
31-07-23	Monday	22,583,380
01-08-23	Tuesday	22,432,739
02-08-23	Wednesday	21,970,073
03-08-23	Thursday	22,050,579
04-08-23	Friday	22,402,865
05-08-23	Saturday	22,642,955
06-08-23	Sunday	22,786,254
07-08-23	Monday	22,109,606
08-08-23	Tuesday	22,405,222
Total	-	**246,338,555**

Schedule Analysis

How much could our framework link the datasets?

Schedule-Filled Percentage = **Matched Trips** / **Scheduled Trips**

- **Total**: 156,628 / 205,884 = 76.08%
- **Weekdays**: 118,559 / 159,418 = 74.37%
- **Saturdays**: 22,796 / 28,200 = 80.84%
- **Sundays**: 15,273 / 18,266 = 83.61%

Delay Analysis

Delay Notation

When comparing a real-time to a expected time:

- **Delay**: $\geqslant$ 1 minute after
- **Ahead-of-Schedule**: $\geqslant$ 1 minute before
- **On time**: $\leqslant$ 59 seconds after OR $\leqslant$ 59 seconds before

PUC Minas

Delay Analysis

Distribution of each status over the network

- **Delay**: 89.8%
- **Ahead-of-Schedule**: 6.9%
- **On time**: 3.3%

Attention!

The predominance of *DELAYED* in the Public Transportation Network **does not imply** that the network is not working nor completely stopped!

PUC Minas

Delay Analysis

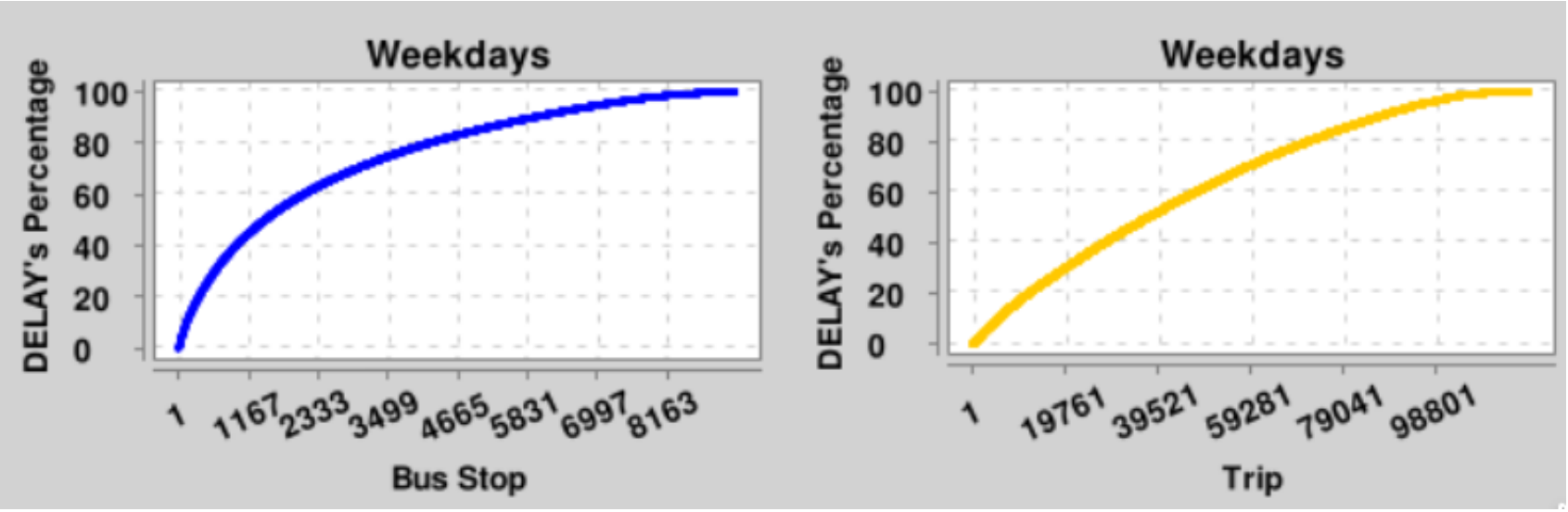


Figura: *DELAY*s Distribution: Bus Stop and Trip

PUC Minas

Delay Analysis

Figura: 300 Most Delayed Stops

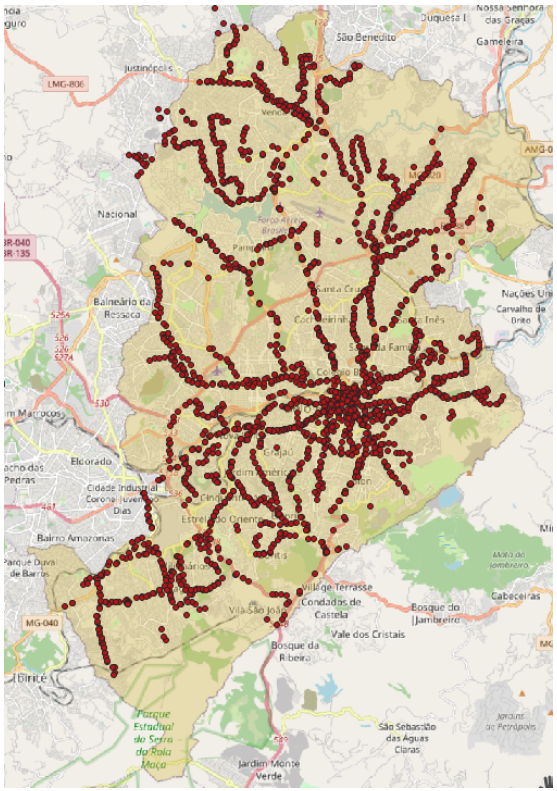


Figura: Fragment of the 50 Most Delayed Stops

PUC Minas

Delay Analysis

Network Constants

1. *Global Ahead Average*: 13.42 minutes
2. *Global Delay Average*: 20.49 minutes

Three Most Delayed Stops for Weekdays

1. #14793268 - *Avenida Dom Pedro II 1520* with 7,309 delays
2. #14791617 - *Avenida Amazonas 7309* with 7,009 delays
3. #14790997 - *Avenida Dom Pedro II 1980* with 6,692 delays

PUC Minas

Delay Analysis

Stops #14793268 and #14790997

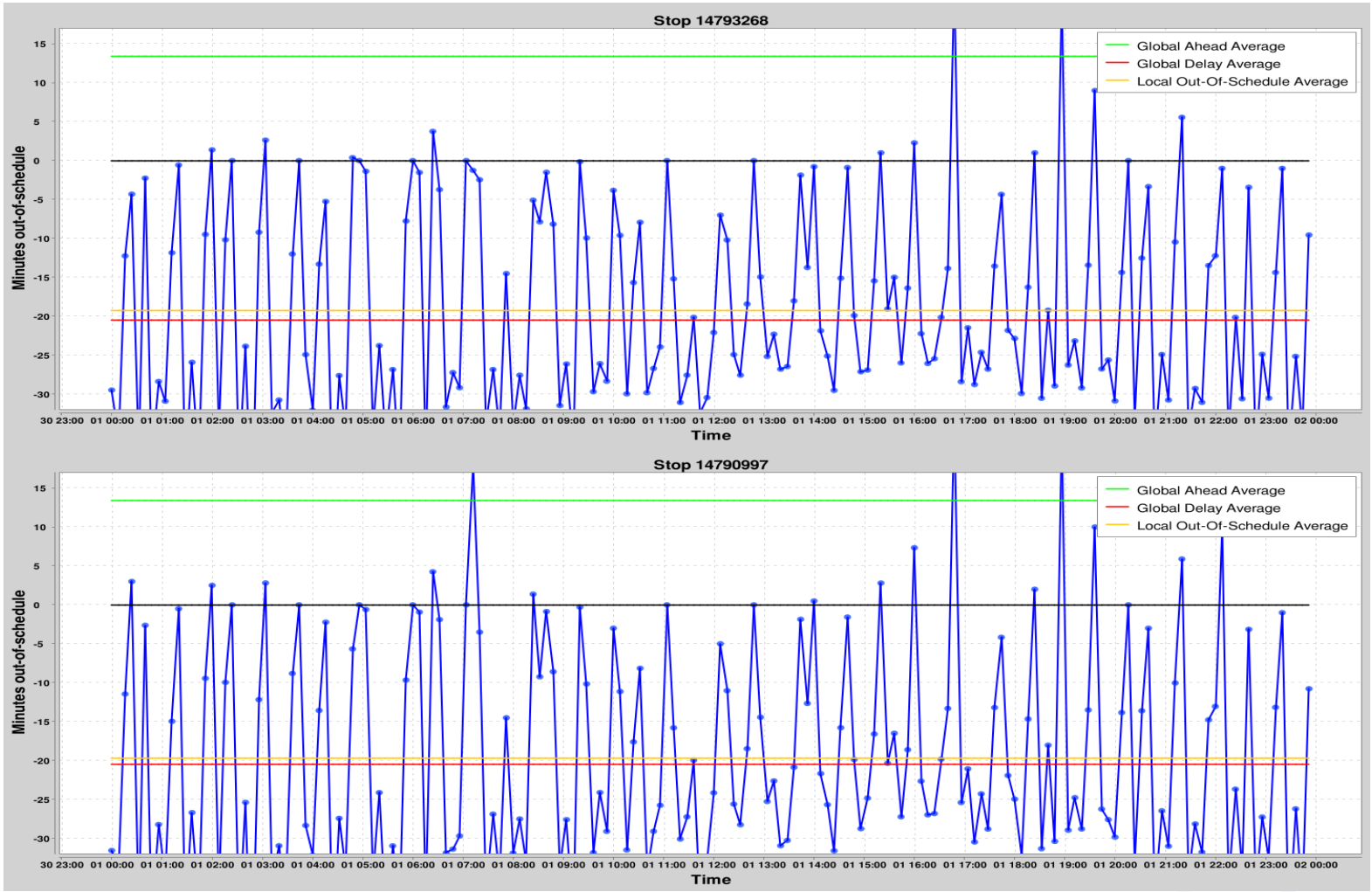
The stops #14793268 and #14790997 are the first and third most delayed in the Public Transportation Network, respectively. Also, these stops are **462** meters from each other on the same avenue, *Avenida Pedro II*, and share **2,590** common trips, so they are spatially related.

Local Out-Of-Schedule Average

- #14793268: 19.29 minutes
- #14790997: 19.68 minutes

PUC Minas

Delay Analysis



PUC Minas

Comparison Between Generated and Real Data

Overview

The previous analysis was only possible because Belo Horizonte's GTFS defines the expected time for all bus stops on every trip. The *Trip Expected Time Generator* generates the expected times when missing, so, we executed this component with Belo Horizonte's data and compared the expected times generated with those defined at the GTFS.

Comparison Between Generated and Real Data

		GTFS	Generated
Weekday	*ON_TIME*	3.3%	3.2%
	AHEAD_OF_SCHEDULE	6.9%	17.8%
	DELAYED	89.8%	79.0%
Saturday	*ON_TIME*	3.9%	3.5%
	AHEAD_OF_SCHEDULE	6.5%	18.4%
	DELAYED	89.6%	78.1%
Sunday	*ON_TIME*	4.4%	3.9%
	AHEAD_OF_SCHEDULE	5.4%	18.5%
	DELAYED	90.2%	77.6%

Comparison Between Generated and Real Data

Global Averages

- *Global Ahead Average*
  - GTFS: 13.42 minutes
  - Generated: 38.57 minutes
  - **Diff**: 25.15 minutes
- *Global Delay Average*
  - GTFS: 20.49 minutes
  - Generated: 24.75 minutes
  - **Diff**: 4.26 minutes

Local Out-Of-Schedule Average

- #14793268
  - GTFS: 19.29 minutes
  - Generated: 15.54 minutes
  - **Diff**: 3.75 minutes
- #14790997
  - GTFS: 19.68 minutes
  - Generated: 14.68 minutes
  - **Diff**: 5 minutes

PUC Minas

Conclusion

Concluding Remarks

- Delays in Belo Horizonte follow a *log-normal* distribution
- Analysis using data generated with the *Trip Expected Time Generator*
- *PondiônsTracker* as a viable option when GTFS-RT is unavailable

Future Work

- Futher explore Belo Horizonte Public Transportation Network using deep learning for graphs approaches
- Reproduce Belo Horizonte's results with other cities

Jayanth Raghothama, Vinutha Magal Shreenath, and Sebastiaan Meijer. 2016. Analytics on Public Transport Delays with Spatial Big Data. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data* (Burlingame, California) *(BigSpatial '16)*. Association for Computing Machinery, New York, NY, USA, 28–33. https://doi.org/10.1145/3006386.3006387

PUC Minas

Conclusion

Thanks!!