MEMO-F541 - Exploring Spatiotemporal Flanders Traffic Data

Pavlo, Brazhko
pavlo.brazhko@ulb.be

August 17, 2023

Internship report

Pavlo Brazhko

Supervised by:
Michiel Dhont, Ir.
data scientist
michiel.dhont@sirris.be

Curated by:
Sreeraj Rajendran, PhD
data scientist
sreeraj.rajendran@sirris.be

Abstract

The main objective of this work was to prepare Spatiotemporal Trajectory data using a methodology developed by Sirris. This methodology builds upon and extends methods outlined by my curator Michiel Dhont. It is described in the article [1], focusing on mobility data for the Small Ring Road of Brussels. However, the current task is even more significant: to study traffic data structures and prepare data for the entire Flemish region.

This work discusses the application of various techniques for data reduction and initial processing. Challenges encountered during the work are discussed, as well as an analysis of the pros and cons of the employed tools.

It is worth noting that the task of studying data preprocessing and primary processing to reduce input data was successfully completed. Throughout this experience, I gained valuable skills in working with real-world data. Collaborative efforts with my mentor to address ongoing challenges proved instrumental in enhancing the quality of our approach and the depth of our insights.

the code for this project can be found at the following link :

link to the gitlab of Sirris project repository

Keyword: Machine Learning, Traffic, vehicle, classification, prediction, Mobility Data, Time series, fingerprint recognition, trajectory

Contents

Chapter 1

Introduction

The EluciDATA Lab[1] is a dynamic and innovative research laboratory within Sirris, the esteemed Belgian scientific organization dedicated to advancing technological innovation. As a significant collaborative center for the technology industry, Sirris provides invaluable support to Belgian companies in realizing their innovation ambitions. The EluciDATA Lab, part of Sirris, is a research laboratory dedicated to data-driven solutions and digital transformation. With a team of experts and access to cutting-edge technologies, the lab supports Belgian companies in embracing innovation.

Operating as part of Sirris, the EluciDATA Lab focuses specifically on data-driven research and digital transformation. With a team of 150 experts and access to over 200 partners and ecosystems, the lab is at the forefront of cutting-edge technologies and data analytics.

During my internship at EluciDATA Lab, I worked on a project focused on studying traffic patterns in the Flemish region, specifically on a major highway bypassing Brussels. The project aimed to gain insights into the traffic flow using a unique data collection approach, which involved the utilization of magnetic sensors embedded directly in the road surface, instead of traditional cameras.

Gathering data from the Flemish government's official website[2] presented its own challenges. Despite these difficulties, the project successfully collected, analyzed, and organized traffic data from the region. The primary objective was to establish a foundation for future tasks related to improving road management, catering to the specific needs of various clients interested in traffic optimization.

[1] The Data and AI Competence Lab of Sirris, see https://elucidata.be/

[2] https://miv.opendata.belfla.be/miv/verkeersdata

Chapter 2

Technical Basis

A detection loop is a device utilized to detect the presence of vehicles and register their passage. These loops are commonly placed at entry lanes of traffic-light-controlled intersections or used for vehicle counting to determine traffic volumes. By appropriately placing several loops one behind the other at a specific distance, it becomes possible to measure the speed of passing vehicles. Detection loops are widely employed as traffic detectors for traffic control purposes.

Principle of Operation: When a vehicle passes over or stays on the loop, its inductance decreases, leading to an increase in the oscillator frequency. If this frequency change surpasses a predetermined threshold, it indicates the presence of a vehicle on the loop.

Figure 2.1: Induction loops[1]

Technical Parameters: Induction loops are embedded into the road at a depth of approximately 120 mm. Their size and orientation are selected based on their intended use. Loop dimensions typically range from 1 to 3 meters in length and 2 to 2.5 meters in width (matching the lane's width and preventing interference from vehicles in adjacent lanes). Special loops can be used for longer stretches, reaching up to 30 meters. The loop itself consists of a coil with a flexible face measuring between 1.5 to 2.5 mm². Additionally, the loop must be resistant to water and heat.[2]

[1] https://www.bikenorth.org.au/2023/01/04/traffic-light-changer/

Despite the abundance of various traffic detection systems such as radar sensors, ultrasonic sensors, infrared sensors, laser sensors, and mobile device sensors, induction detection loops remain one of the most widespread and well-established methods. The popularity of induction loops is attributed to their proven effectiveness and reliability.

Pros and Cons:

These loops provide essential traffic parameters, such as traffic intensity and occupancy, at a reasonable cost and are immune to adverse weather conditions.

However, a drawback of using induction loops is the lane closure and milling of the asphalt layer required during installation and maintenance. Moreover, malfunctions may occur due to poor road surface quality, necessitating re-installation during road repairs, which could potentially reduce the road's lifespan if not properly addressed.

Chapter 3

Data Collection

Meten-in-Vlaanderen[1] (MIV) provides minute-level traffic measurement data. These data are collected using double measurement loops, primarily located on highways in the Flemish region. The responsible entities for these data are the Agentschap Wegen en Verkeer (AWV)[2] and the Vlaams Verkeerscentrum (VVC)[3]. The dataset includes information on the number of vehicles and average speeds, categorized into 5 vehicle classes, aggregated per minute, along with corresponding locations of the measurement loops. The data is updated every minute.

It should be noted that at the moment there is no open source of data on car traffic in the Flemish region. In other words, you cannot openly obtain this data for any given period of time. At the moment, there is only one mechanism for requesting the official website of the Flemish government[1] and collecting data on its server. This was done by the efforts of the employees of the Sirris - EluciDATA Lab[4] division. With the help of the code they wrote, which allows you to request data from the official Flemish site every minute, the data used in this work was collected. Data has been collected since April 14, 2021.

[1]MIV data, see https://miv.opendata.belfla.be/miv/verkeersdata

[2]Agency for Roads and Traffic (AWV), see http://www.wegenenverkeer.be/

[3]Flemish Traffic Center (VVC), see http://www.verkeerscentrum.be/

[4]The Data and AI Competence Lab of Sirris, see https://elucidata.be/

Chapter 4

Data description

4.1 Metadata parameters

Studying the data structure and its encryption methods at the very beginning stage proved beneficial for the subsequent work and a deeper understanding of the examined system.

4.2 Placement of sensors in the Flemish region

At the moment, the number of sensors involved in the collection of traffic information in the Flemish region is 4477 positions.

Figure 4.1: Placement of sensors in the Flemish region

4.3 Sensor Table Parameters

The dataframe containing information about each sensor installed on the roads of the Flemish region provides key details about each unique sensor.

- `unieke_id` - unique identificator
- `beschrijvende_id`- descriptive ID
- `volledige_naam` - full road name
- `Ident_8` - road name and direction
- `lve_nr` - Number of the LVE[1]. An LVE processes the data of a group of measuring points.
- `Kmp_Rsys` - Reference points along the numbered roads used in the systems of the Flemish Traffic Center.
- `Rijstrook` - The letter indicates the type of lane.
- `X_coord_EPSG_31370` - X coordinate
- `Y_coord_EPSG_31370` - Y coordinate
- `lengtegraad_EPSG_4326` - longitude
- `breedtegraad_EPSG_4326` - latitude
- `geometry` - geometric point

	unieke_id	beschrijvende_id	volledige_naam	Ident_8	lve_nr	Kmp_Rsys	Rijstrook
1	3640	H291L10	Parking Kruibeke	A0140002	437.0	94.695	R10
2	3638	H292L20	Parking Kruibeke	A0140001	437.0	94.692	R11
3	3639	H292L30	Parking Kruibeke	A0140001	437.0	94.692	R12
4	2304	H202L10	Snede noordzijde sleuf Boom	N1770002	168.0	10.268	R10
5	2303	H208L10	Snede noordzijde sleuf Boom	A0121311	168.0	NaN	R10

[1]Lokale VerwerkingsEenheid - Local Processing Unit

X_coord_EPSG_31370	Y_coord_EPSG_31370	lengtegraad_EPSG_4326	breedtegraad_EPSG_4326	geometry
144474.5297	208293.5324	4.289731	51.184608	POINT (144474.530 208293.532)
144485.3962	208275.6893	4.289887	51.184447	POINT (144485.396 208275.689)
144483.6764	208278.8952	4.289862	51.184476	POINT (144483.676 208278.895)
150233.1199	200546.9427	4.372085	51.115002	POINT (150233.120 200546.943)
150226.4523	200546.5205	4.371990	51.114999	POINT (150226.452 200546.521)

Figure 4.2: Sensor table parameters

Road type	Mission
R	regional roads
N	national roads
B	special Crossing Bedding (BOB) or bus lane
TR	measurement of traffic in the opposite direction (usually at Tunnels) in the R lane
P	breakdown lane
W	parking or other road
S	rush hour lane
A	measuring point on a shaded part of the road

Table 4.1: Road numbering system in the Flemish Region

As illustrated in Table 4.2 numbering starts at R10 for the first regular lane on the main roadway. Numbering increases from right/trailer to left/fast lane. Lanes with 09, 08, 07, 06 are then to the right of this and usually represent entrance/exit lanes, merge lanes, newly added lanes, rush-hour lanes or BOB[2]. Lanes 11, 12, 13 are then located to the left of lane R10. The number 00 is used for a measuring point on the breakdown lane (P00). The TR lane is identical to the corresponding R lane (TR10 = R10, TR11 = R11, TR12 = R12), only the measuring point transmits only the minute data of the "ghost traffic". The measurement data for TR10 come from the same measurement loops as those for R10.

[2]Bijzonder Overrijdbare Bedding - waffle grid road marking

4th	3rd	2nd	1st					
		A11						P00
R13	**R12**	**R11**	**R10**	**R09**	**R08**	**R07**	**R06**	
				B09	B08	B07		
				S09				
				W09				
	TR12	TR11	TR10	TR09				

Table 4.2: Order and designation of lines on the main roadway

It is worth noting that the parameters `Ident_8` and `Rijstrook` use road designations slightly differently. `Ident_8` uses only the codes A, B, R, and N to denote road names, with the last one (N) not being used for `Rijstrook` coding. On the other hand, `Rijstrook` uses the following codes for lane designation: A, B, R, P, S, TR, W. It may seem counterintuitive, but the alphabetical codes for road names and lanes may not always match (e.g., `unieke_id` = 3640, `Ident_8` = A0140002, `Rijstrook` = R10). From this, we can conclude that the `Ident_8` parameter is more general than `Rijstrook`, as the latter provides more specific descriptions for each lane that makes up the road identified by `Ident_8`.

4.4 Time Table Parameters

The data structure of Time Table utilizes Hierarchical Indexing (MultiIndex), where the top level represents a unique sensor identifier, and the lower level consists of timestamps for that particular sensor, see Figure 4.3.

```
]: print(test_data.columns)

MultiIndex([('3162',                            'bezettingsgraad'),
            ('3162',                'class_2_verkeersintensiteit'),
            ('3162', 'class_2_voertuigsnelheid_harmonisch'),
            ('3162',                'class_3_verkeersintensiteit'),
            ('3162', 'class_3_voertuigsnelheid_harmonisch'),
            ('3162',                'class_4_verkeersintensiteit'),
            ('3162', 'class_4_voertuigsnelheid_harmonisch'),
            ('3162',                'class_5_verkeersintensiteit'),
            ('3162', 'class_5_voertuigsnelheid_harmonisch'),
            ('3162',                               'onrustigheid'),
            ('3163',                            'bezettingsgraad'),
            ('3163',                'class_2_verkeersintensiteit'),
            ('3163', 'class_2_voertuigsnelheid_harmonisch'),
            ('3163',                'class_3_verkeersintensiteit'),
            ('3163', 'class_3_voertuigsnelheid_harmonisch'),
            ('3163',                'class_4_verkeersintensiteit'),
            ('3163', 'class_4_voertuigsnelheid_harmonisch'),
            ('3163',                'class_5_verkeersintensiteit'),
            ('3163', 'class_5_voertuigsnelheid_harmonisch'),
            ('3163',                               'onrustigheid'),
            ('3164',                            'bezettingsgraad'),
            ('3164',                'class_2_verkeersintensiteit'),
            ('3164', 'class_2_voertuigsnelheid_harmonisch'),
            ('3164',                'class_3_verkeersintensiteit'),
            ('3164', 'class_3_voertuigsnelheid_harmonisch'),
            ('3164',                'class_4_verkeersintensiteit'),
            ('3164', 'class_4_voertuigsnelheid_harmonisch'),
            ('3164',                'class_5_verkeersintensiteit'),
            ('3164', 'class_5_voertuigsnelheid_harmonisch'),
            ('3164',                               'onrustigheid')],
           )
```

Figure 4.3: Hierarchical Indexing (MultiIndex)

The data represents temporal information with a one-minute time interval.

The parameters of the time table data contain the 23 following features:

- `actueel_publicatie` - current publication
- `beschikbaar` - available item`beschikbaarheidsgraad` - availability rate
- `bezettingsgraad` - occupancy rate has a range between 0 and 1, however, in practice values above 0.5 are rare
- `class_1_verkeersintensiteit` - class 1 traffic intensity

- `class_1_voertuigsnelheid_harmonisch` - class 1 vehicle speed harmonic
- `class_1_voertuigsnelheid_rekenkundig` - class 1 vehicle speed arithmetic
- `class_2_verkeersintensiteit` - class 2 traffic intensity
- `class_2_voertuigsnelheid_harmonisch` - class 2 vehicle speed harmonic
- `class_2_voertuigsnelheid_rekenkundig` - class 2 vehicle speed arithmetic
- `class_3_verkeersintensiteit` - class 3 traffic intensity
- `class_3_voertuigsnelheid_harmonisch` - class 3 vehicle speed harmonic
- `class_3_voertuigsnelheid_rekenkundig` - class 3 vehicle speed arithmetic
- `class_4_verkeersintensiteit` - class 4 traffic intensity
- `class_4_voertuigsnelheid_harmonisch` - class 4 vehicle speed harmonic
- `class_4_voertuigsnelheid_rekenkundig` - class 4 vehicle speed arithmetic
- `class_5_verkeersintensiteit` - class 5 traffic intensity
- `class_5_voertuigsnelheid_harmonisch` - class 5 vehicle speed harmonic
- `class_5_voertuigsnelheid_rekenkundig` - class 5 vehicle speed arithmetic
- `defect` - defective
- `geldig` - available
- `lve_nr` - unic number of the LVE[3]. An LVE processes the data of a group of measuring points.
- `onrustigheid` - restlessness
- `tijd_laatst_gewijzigd` - time last modified

Each of the five classes of vehicles has the following characteristics:

- `Voertuigklasse_1` - Motorcycles. This vehicle class was provided vehicles with approximate lengths between 0m and 1.00m. This data is no longer used by AWV[4] and the traffic center. The sporadic measurements in this vehicle class are little or not reliable.
- `Voertuigklasse_2` - Passenger cars. Vehicles with approximate length between 1.00m and 4.90m
- `Voertuigklasse_3` - Vans. Vehicles with approximate length between 4.90m and 6.90m

[3]Lokale VerwerkingsEenheid - Local Processing Unit
[4]Agentschap Wegen en Verkeer - Agency for Roads and Traffic

- `Voertuigklasse 4` - Unarticulated trucks. Vehicles with estimated length between 6.90m and 12.00m
- `Voertuigklasse 5` - Articulated trucks or buses. Vehicles with approximate length longer than 12.00m.
- `voertuigsnelheid rekenkundig` - Vehicle speed arithmetic

Vehicle speed arithmetic is equal:

$$V_{arithmetic} = \frac{\sum v_i}{n} \tag{4.1}$$

where v_i = individual speed of a vehicle within this vehicle class

Vehicle speed harmonic is equal:

$$V_{harmonic} = \frac{n}{\sum \frac{1}{v_i}} \tag{4.2}$$

where v_i = individual speed of a vehicle within this vehicle class

Value range from 0 to 200 km/h

Special values:

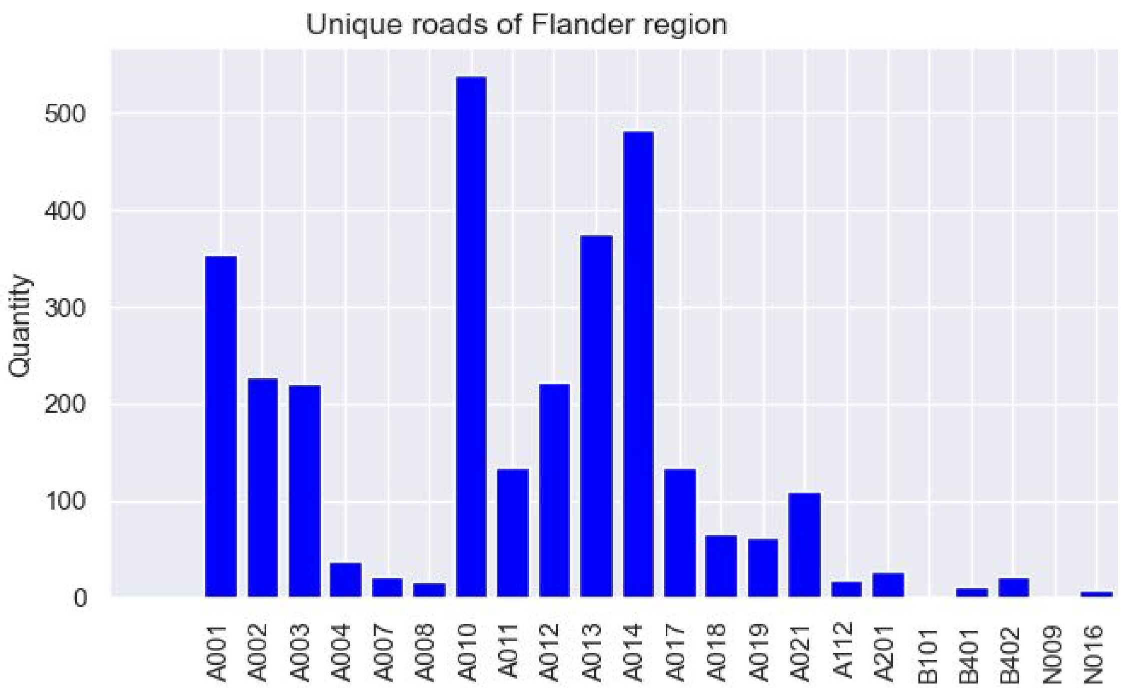
- 251: Initial value
- 254: Calculation not possible
- 252: no vehicles passed within this vehicle class.

Chapter 5

Exploratory Data Analysis

5.1 The Types of roads in Flanders region

The Flanders road transportation network stands as one of the densest in Europe, reflecting its pivotal role in facilitating connectivity and mobility. This intricate web of roads serves as a critical artery for the movement of goods and people, driving economic development and enhancing accessibility. In this context, the presented Figures 5.1 illustrates a comprehensive overview of road compositions within this dynamic region, capturing the diversity of road types and their significance in catering to the intricate transportation demands of densely interconnected community.



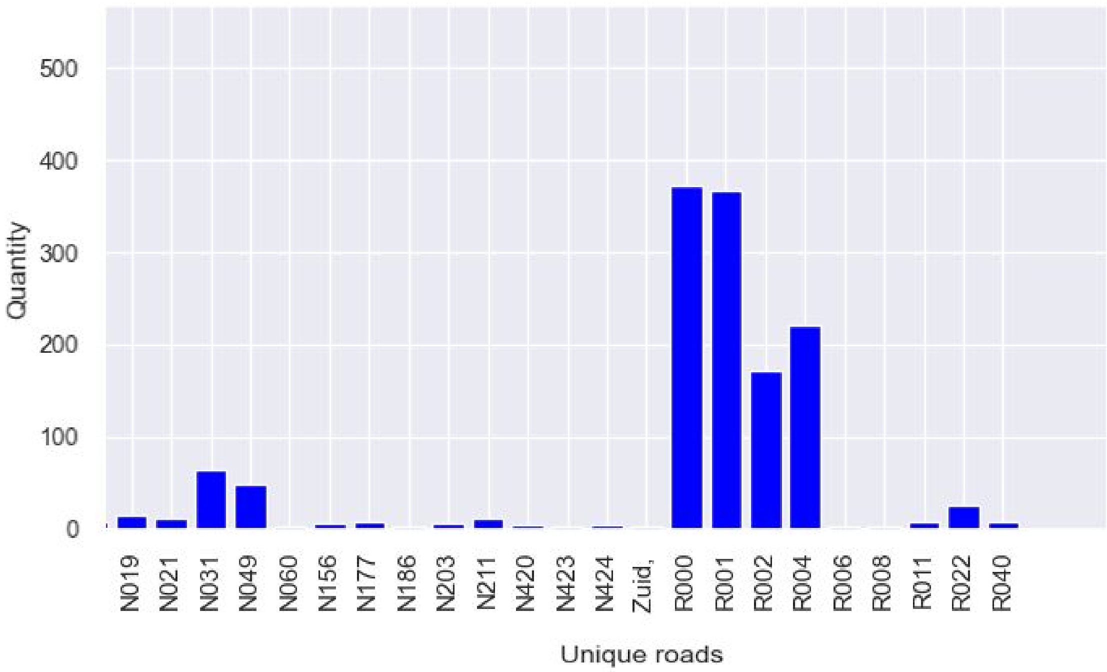


Figure 5.1: The Types of roads in Flanders region

5.2 The Sensor Locations by Local Processing Unit

Since one of the main tasks in the study of the data was to find the parameters by which it is possible to significantly simplify the spatiotemoral data, the analysis was carried out as grouped sensors according to the `lve_nr`[1] feature .

`lve_nr` is a group of sensors physically connected to a hub - a control panel that receives input data and from which this data is transmitted to the general system.

[1]Lokale VerwerkingsEenheid - Local Processing Unit

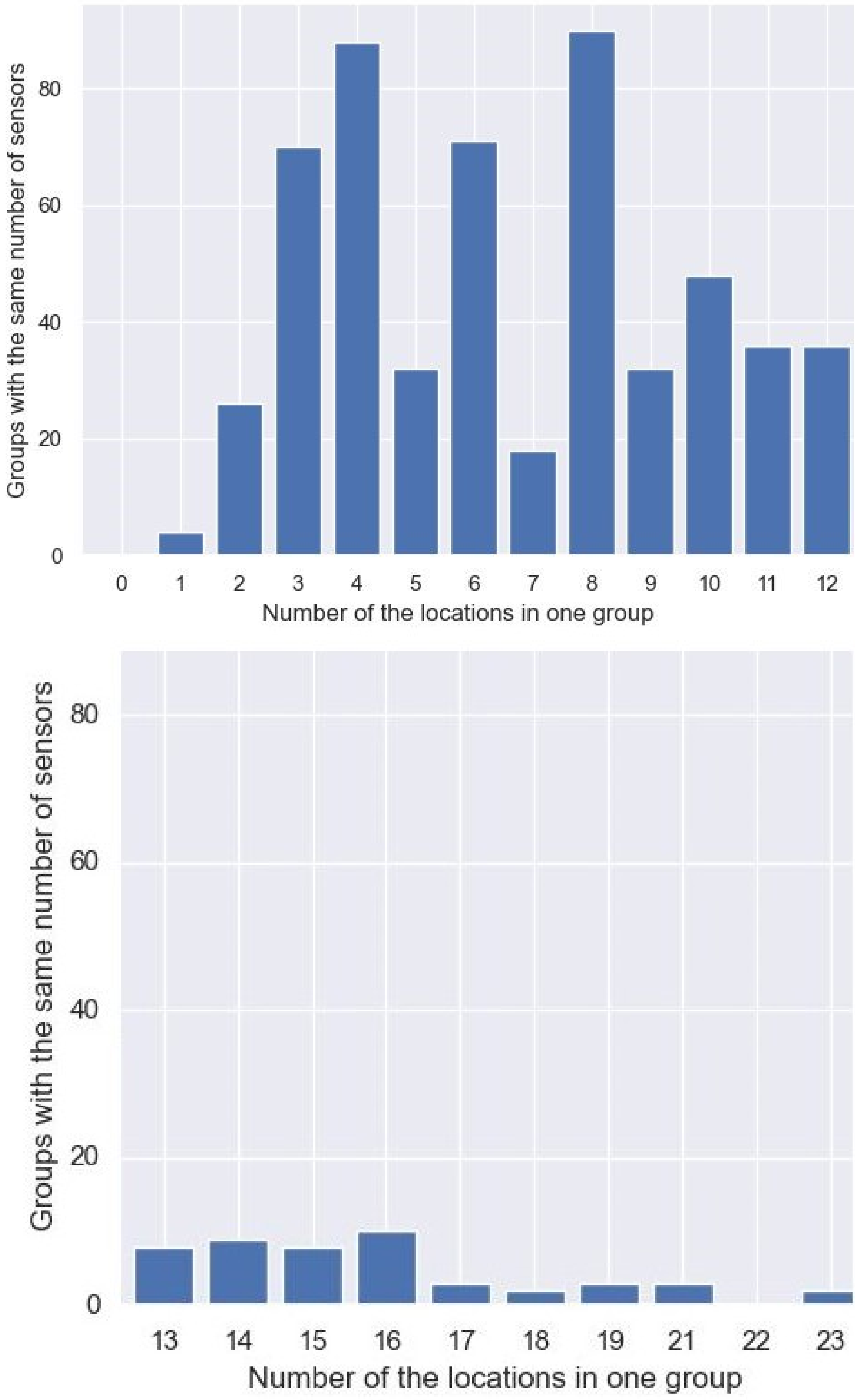


Figure 5.2: The Sensor Locations by Local Processing Unit

As the analysis showed, the number of sensors included in the group can be quite diverse. This is due to the technical parameters of connecting sensors. Figure 5.2 shows that the most common LVE groups are three, four, six and seven sensors.

Subsequently, the use of the `lve_nr` feature as a data reduction tool had to be abandoned. Instead, a tandem of `Ident_8` and `Kmp_Rsys` features were chosen, which will be discussed in Section 6.6.

5.3 The Sensor Locations by Traffic Lines

The table 4.2 above shows the order of the lines relative to each other on the roadway. Below are explanations of the designations of these lines.

The encoding of a specific lane of the road is coded in the `'Rijstrook'` column and has the following values:

- (B07) - special Crossing Bedding (BOB) or bus lane
- (B08) - special Crossing Bedding (BOB) or bus lane
- (B09) - special Crossing Bedding (BOB) or bus lane
- (P00) - breakdown lane
- (R06) - 4th additional lane on the right
- (R07) - 3rd additional lane on the right
- (R08) - 2nd additional lane on the right
- (R09) - 1st additional lane on the right
- (R10) - 1st lane
- (R11) - 2nd lane
- (R12) - 3rd lane
- (R13) - 4th lane
- (S09) - rush hour lane
- (TR09) - 1st additional lane on the right
- (TR10) - 1st ghost line
- (TR11) - 2nd ghost line
- (TR12) - 3rd ghost line
- (W09) - parking or other way

From the very beginning, an analysis of the sensor position statistics on various lanes of the Flemish roads was conducted. Table 4.2 depicting lane markings was provided earlier.

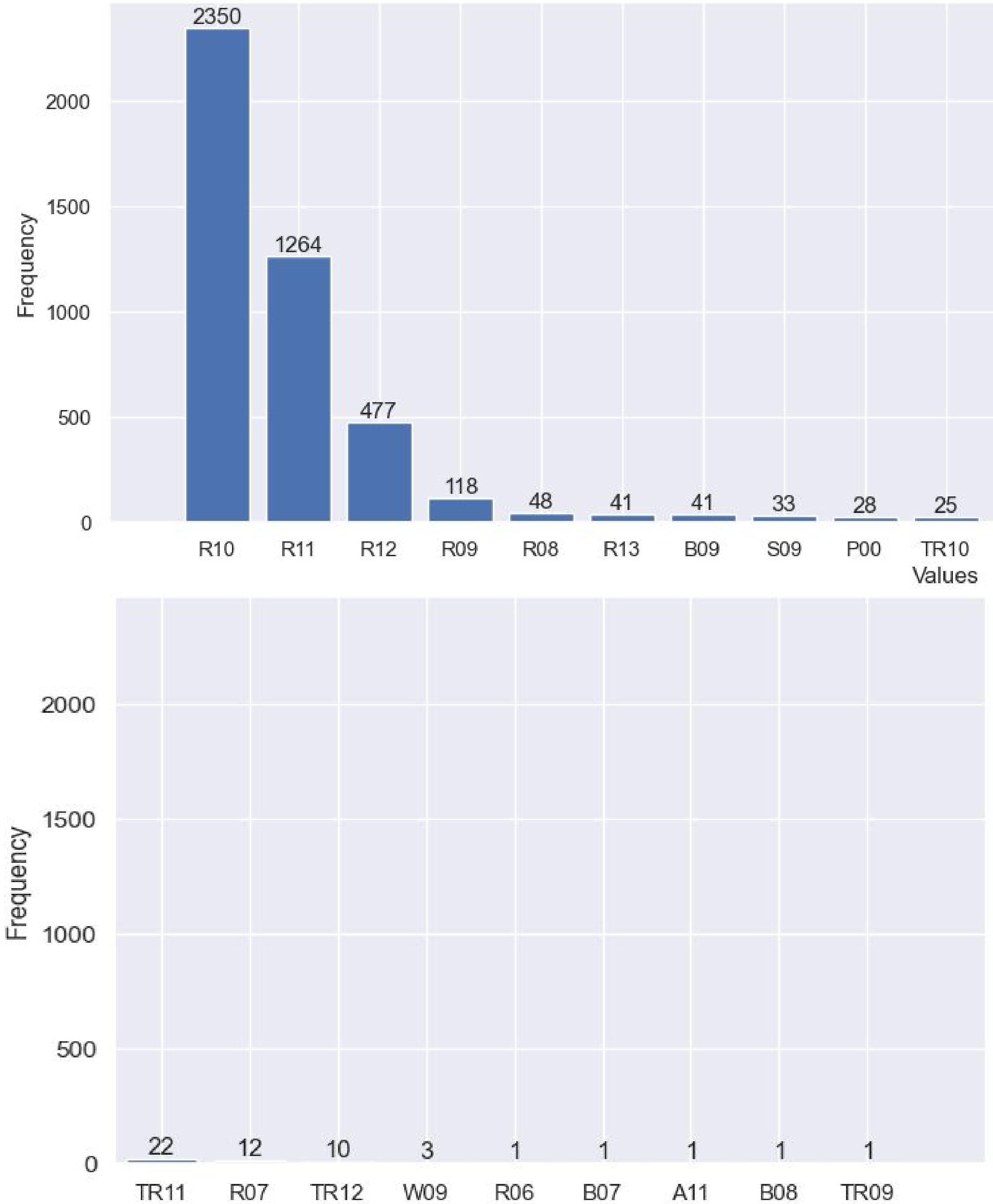


Figure 5.3: Types of roads included in the system

The majority of sensors are installed on the first (R10), second (R11), and third lanes (R12), as well as the first additional lane on the right (R09). This might justify disregarding the minor loss of information from the remaining lanes, which collectively constitute less than 6% of the total sensor volume.

The distribution of sensor data across traffic lanes is presented in Figure 5.3.

5.4 Characteristics of the direction of movement

During the study of the `Ident_8` characteristics, it was discovered that two different direction of "main road" having the last four digits `0001` and `0001` are quite balanced (1447 verse 1521 respectively). Additionally, it is worth noting two cases where the direction of movement is denoted as "Ottergemsesteenweg Zuid, Gent," representing a new road that does not fit into the general direction identification system.

5.5 Brussels Big Ring Data

In our work, we will mainly focus on examining the data related to the zone of the Brussels Ring Road, which is represented by the road R0. In this section surrounding Brussels, there are 193 sensors installed, making it the largest cluster among all the ones presented in the region.

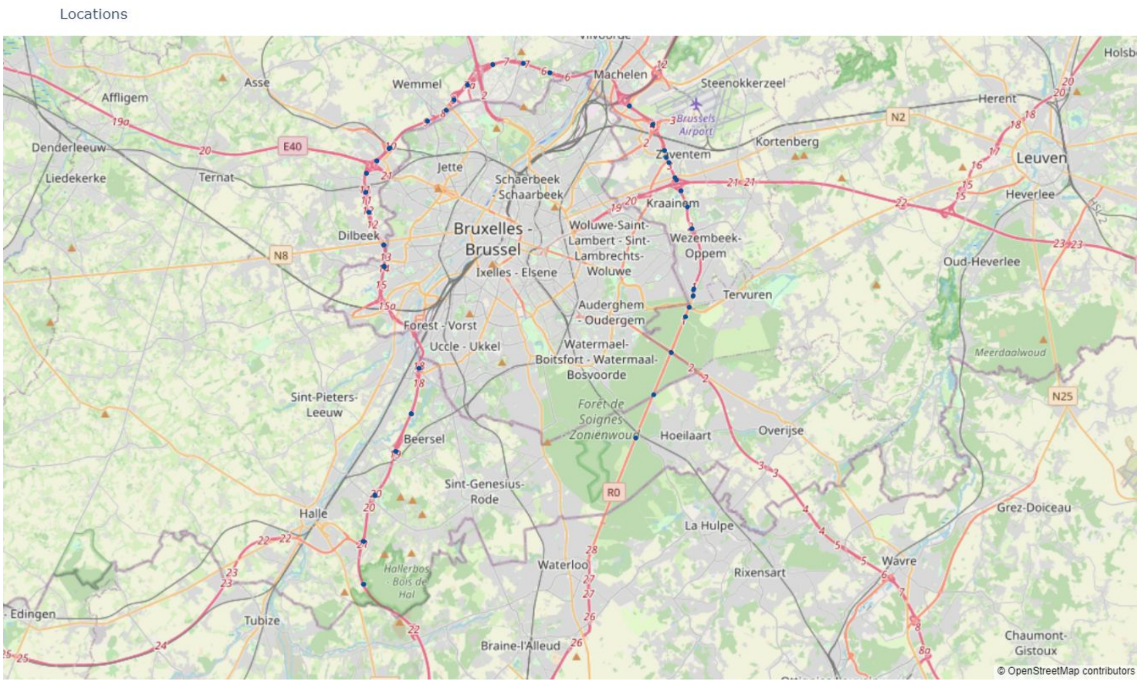


Figure 5.4: Brussels Big Ring - R0

At the initial stage of data exploration for the Brussels Ring Road, three sensors ['3162', '3163', '3164'] were selected, which belong to the same group (`lve_nr`) of sensors located at the largest of the five multi-level interchanges, Knooppunt Groot-Bijgaarden west (Figure 5.5).

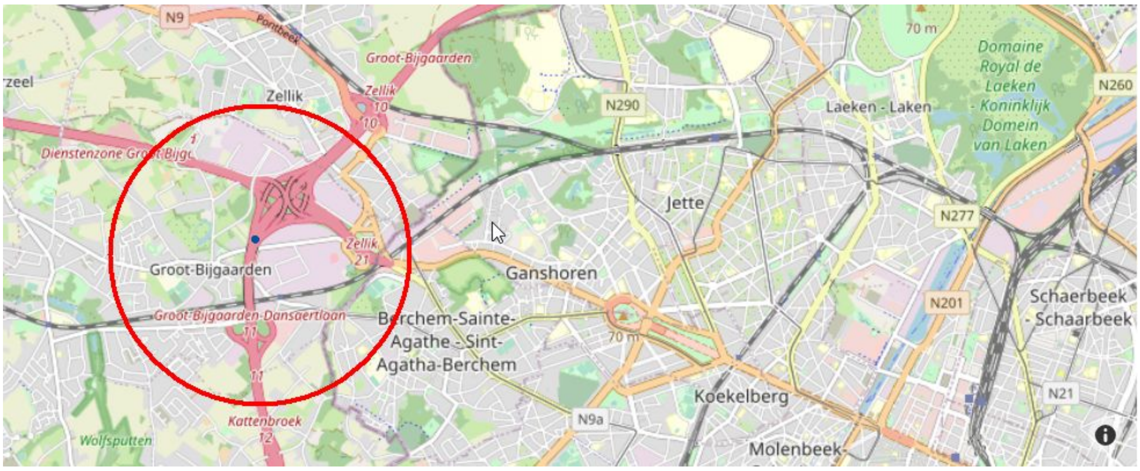


Figure 5.5: Roundabout of Groot-Bijgaarden

At this interchange, the central traffic lanes were selected as they were considered to best represent the average traffic characteristics along the Brussels Ring Road (Figure 5.6).

Figure 5.6: Locations 3162, 3163, 3164

The choice of multiple sensors was motivated by the goal of reducing the amount of data while retaining the ability to analyze the system as a whole.

5.6 Road Types of R0

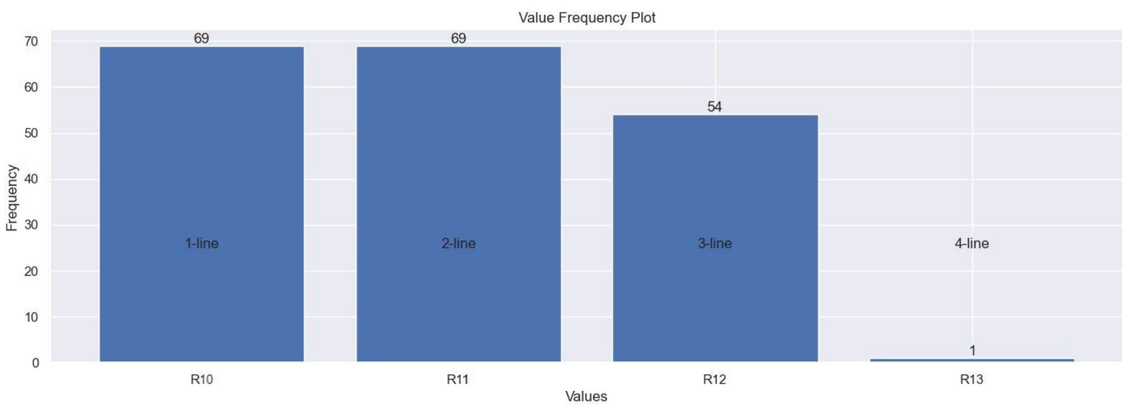


Figure 5.7: Road Types in the Big Brussels Ring

Let's examine the types of roads represented in the R0 cluster for the Brussels Big Ring. As seen in Figure 5.7, there are only four types of roads: R10, R11, R12, and R13, with one, two, three, and four lanes, respectively. Additionally, the four-lane road (R13) is represented by only one instance in the data set.

5.7 Missing Timestamps

During the study of missing data from sensors, for example, we analyzed the number of monthly gaps for the sensor ['3162']. The total time when the sensor did not send a signal to the system was also calculated. As can be seen from Table 5.1, this time can be quite significant. These gaps cannot be easily recovered using common date mining methods.

Time	Count	Duration
2021-04-30 00:00:00+00:00	2185	1 days 12:25:00
2021-05-31 00:00:00+00:00	639	0 days 10:39:00
2021-06-30 00:00:00+00:00	1142	0 days 19:02:00
2021-07-31 00:00:00+00:00	1581	1 days 02:21:00
2021-08-31 00:00:00+00:00	2892	2 days 00:12:00
2021-09-30 00:00:00+00:00	35826	24 days 21:06:00
2021-10-31 00:00:00+00:00	1635	1 days 03:15:00
2021-11-30 00:00:00+00:00	0	0 days 00:00:00
2021-12-31 00:00:00+00:00	5517	3 days 19:57:00
2022-01-31 00:00:00+00:00	224	0 days 03:44:00
2022-02-28 00:00:00+00:00	0	0 days 00:00:00
2022-03-31 00:00:00+00:00	554	0 days 09:14:00
2022-04-30 00:00:00+00:00	1537	1 days 01:37:00
2022-05-31 00:00:00+00:00	0	0 days 00:00:00
2022-06-30 00:00:00+00:00	745	0 days 12:25:00
2022-07-31 00:00:00+00:00	2775	1 days 22:15:00
2022-08-31 00:00:00+00:00	0	0 days 00:00:00
2022-09-30 00:00:00+00:00	40614	28 days 04:54:00
2022-10-31 00:00:00+00:00	0	0 days 00:00:00
2022-11-30 00:00:00+00:00	29955	20 days 19:15:00
2022-12-31 00:00:00+00:00	17417	12 days 02:17:00
2023-01-31 00:00:00+00:00	5500	3 days 19:40:00
2023-02-28 00:00:00+00:00	0	0 days 00:00:00
2023-03-31 00:00:00+00:00	2447	1 days 16:47:00
2023-04-30 00:00:00+00:00	3497	2 days 10:17:00
2023-05-31 00:00:00+00:00	9094	6 days 07:34:00
2023-06-30 00:00:00+00:00	40731	28 days 06:51:00
2023-07-31 00:00:00+00:00	0	0 days 00:00:00

Table 5.1: locations=['3162'] interruption statistics and total downtime on a monthly basis

5.8 Extracting information about missing data

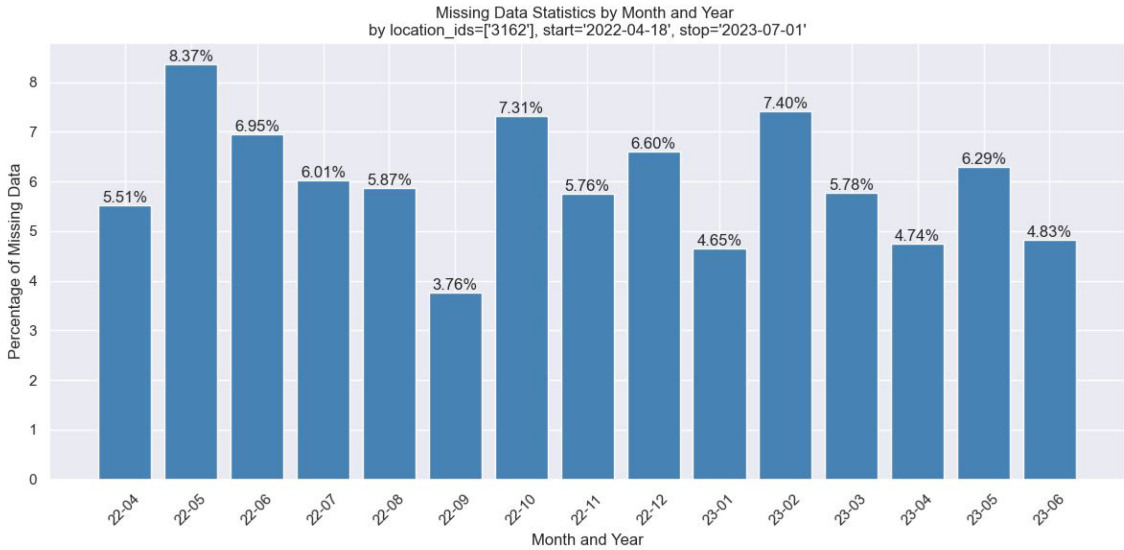


Figure 5.8: Percentage of Missing Data

5.9 Interpolation of missing data

As will be shown in section 6.4, we use the Rolling Window function to interpolate short periods of sensor failure. It allows not only to smooth data but also to interpolate missing data up to 4 minutes in size.

Rolling window

```
test_data_rolling = test_data.rolling(window=100, min_periods=1, center=True, win_type=None, on=None).mean()

test_data_rolling
```

_time	bezettingsgraad	class_2_verkeersintensiteit	class_2_voertuigsnelheid_harmonisch	class_3_verkeersintensiteit	clas
2021-04-18 13:48:00+00:00	4.000000	12.145833	94.416667	0.895833	
2021-04-18 13:49:00+00:00	4.040816	12.081633	94.326531	0.877551	

Figure 5.9: Applying rolling window technique in the code.

Chapter 6

Data Mining Techniques

6.1 Data Reduction Approach

Ultimately, the main strategy was to reduce the amount of data for subsequent work with them and identify insights. The idea is to merge information about different types of vehicles together. Since the traffic database in the Flemish region contains information about four different types of vehicles, we therefore have data on the number of vehicles passing and the harmonic speed for all four types (8 features). When we combine these data, the output we have is the combined traffic flow - the number of passing cars and their harmonic speed. In other words, we are preparing data for consideration in a more general way, which is logical for the initial stage of work on this large project. Over time, it will be easy to extract and analyze data for a specific type (for example, trucks). Now, when exactly for one sensor all 4 groups of vehicles are combined, we get one feature of the number of cars and another feature that contains harmonic speed (2 features). If normal summation is used to calculate the total number of passing vehicles, then the formula described in section 6.7 is used to calculate the average harmonic speed.

6.2 Enter-exit sensors

Since only major roads have 0001 and 0002 endings at the end of the `Ident_8` parameter, we can exclude all other sensors to start traffic analysis. Excluded sensors are installed on auxiliary roads intended for entry and exit from main roads But first, let's look at this types of sensors on a map.

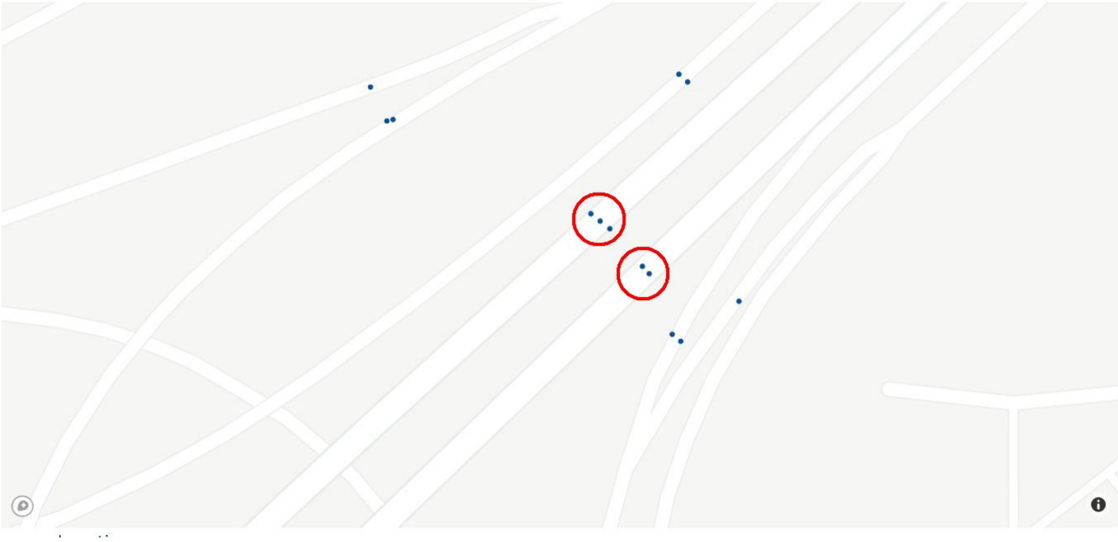


Figure 6.1: Sensors of main roads and enter-exit roads

Figure 6.1 shows all sensors in one of the randomly selected areas. Sensors are presented both on the main roads and on the roads intended for arrival and exit on the main roads.

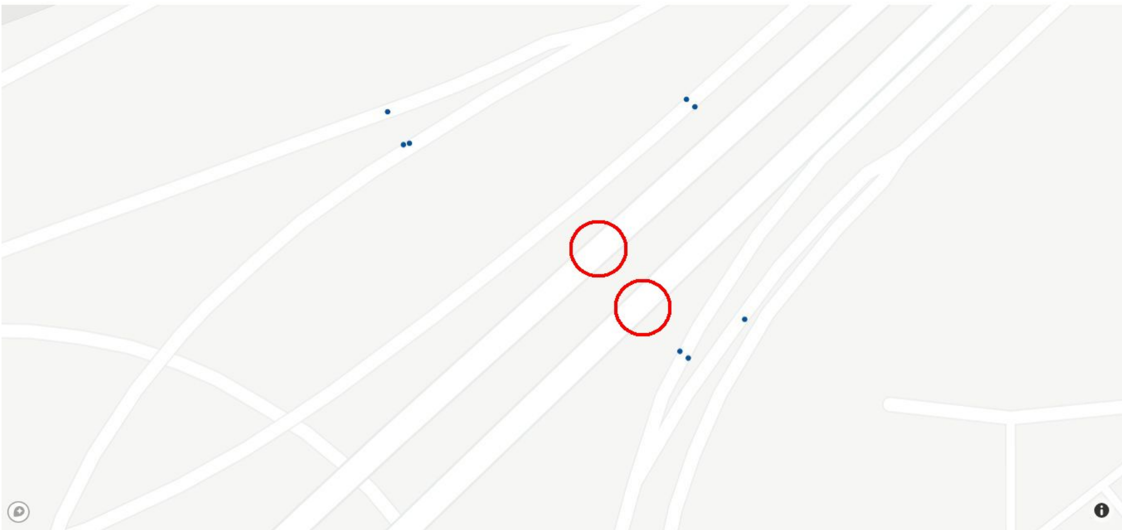


Figure 6.2: Enter-exit roads

While figure 6.2 is a visualization of only sensors installed on auxiliary sections (enter-exit) of roads.

The total number of sensors installed in the Flemish region is 4477. Among them, 1508 sensors are installed on roads designed for entry and exit to main roads, and 2969 sensors are installed on main roads.

In our study, we are examining sensors on main roads to identify characteristic patterns, as they contain the primary information about traffic behavior. While not disregarding the significance of data from sensors on entry and exit roads, it is important to note that they represent specific cases and may be of interest for particular focused tasks.

Each sensor belongs to a group of detectors (`lve_nr`) characterized by their close proximity in location.

6.3 'Not a Number' values - 252.0

The following features are used to present the average harmonic speed.

- `'class_2_voertuigsnelheid_harmonisch'`
- `'class_3_voertuigsnelheid_harmonisch'`
- `'class_4_voertuigsnelheid_harmonisch'`
- `'class_5_voertuigsnelheid_harmonisch'`

In cases where the sensor did not detect a passing car, the system introduces a conditional value of **252** for harmonic velocity to prevent division by zero. If the sensor generates an error, the system uses encoding **254**.

The harmonic mean velocity v is defined as:

$$v(x_i, t_i) = \frac{n_i}{\sum \frac{1}{v_i}} \tag{6.1}$$

6.4 Rolling Window

Time-series data often contain fluctuations and irregularities that can make analysis and interpretation challenging. In such cases, applying techniques like Rolling windows with a 5-minute interval can significantly contribute to stabilizing the data.

The application of Rolling windows also facilitates data interpolation, enabling the filling of missing values or gaps within the time-series. By allowing interpolation intervals of up to 4 minutes, the technique helps reduce the number of empty rows, effectively maximizing the data utilization.

The strategic application of rolling windows with a 5-minute interval presents a powerful solution for stabilizing time-series data, mitigating short-term fluctuations.

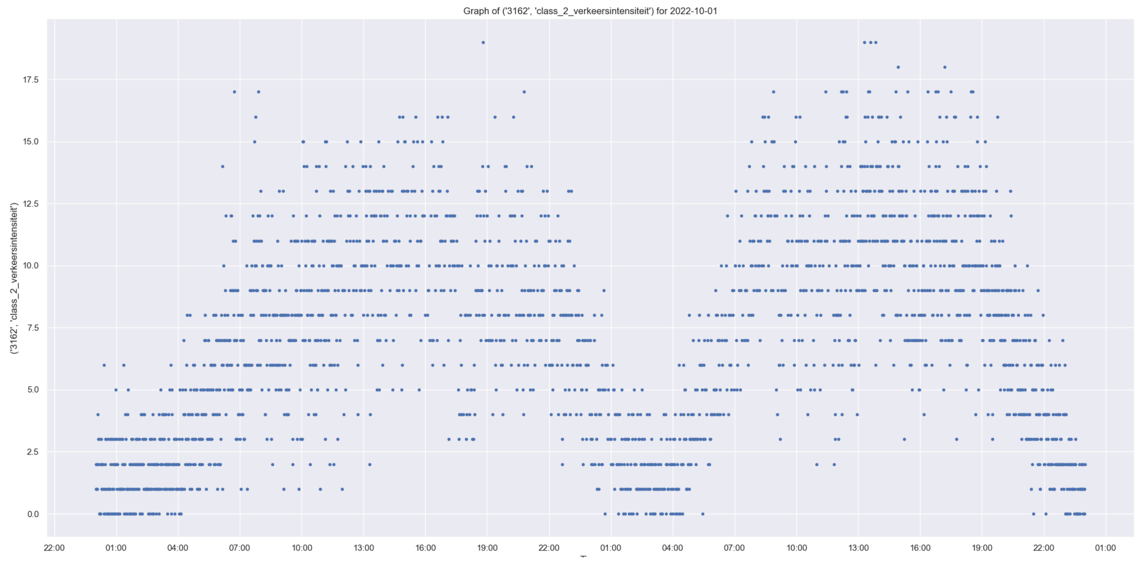


Figure 6.3: Sensor ['3162'] data before applying Rolling window

As can be seen from Figure 6.3, the data on traffic intensity changes during two days (from 2022-10-01 to 2022-10-03 for class `'class_2_verkeersintensiteit'`) have a discrete nature, and interpreting seasonality (in this case, daily) is possible but quite challenging.

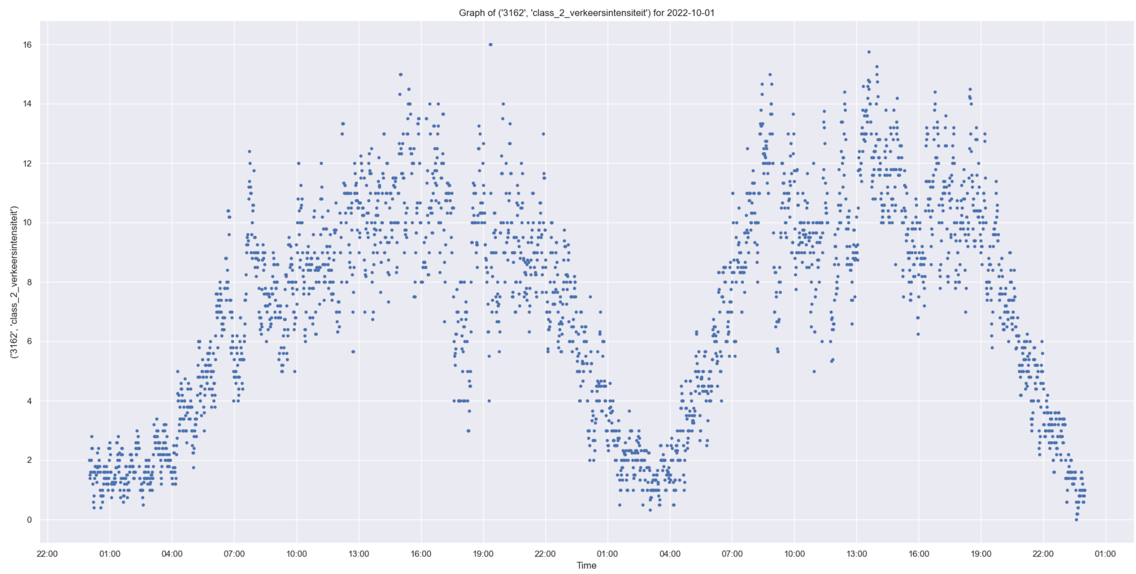


Figure 6.4: Sensor ['3162'] data after applying Rolling window size=5

As clearly evident from Figure 6.4, the data exhibit enhanced stability following the application of Rolling Windows. This is advantageous for uncovering spatiotemporal patterns.

Furthermore, Table 6.1 presents statistics regarding the reduction in the number of rows with missing data, based on the size of the Rolling Window. It is important, however, to strike a balance between data smoothing and the inevitable information loss it entails.

Rolling Window size	Number of NaN rows
row data	277 209
5	219 637
10	217 972
60	213 964
300	206 507

Table 6.1: Sensor['3162'] - NaN rows and window size dependency correlation

As much as we may desire, Rolling Windows is not a miraculous solution for handling missing data. Even when expanding the window size up to 300 minutes (Figure 6.5), a substantial amount of missing data still persists. These data gaps are characterized by prolonged time periods and can be attributed to technical sensor malfunctions requiring significant time for resolution, or road surface replacement, which also leads to significant data gaps.

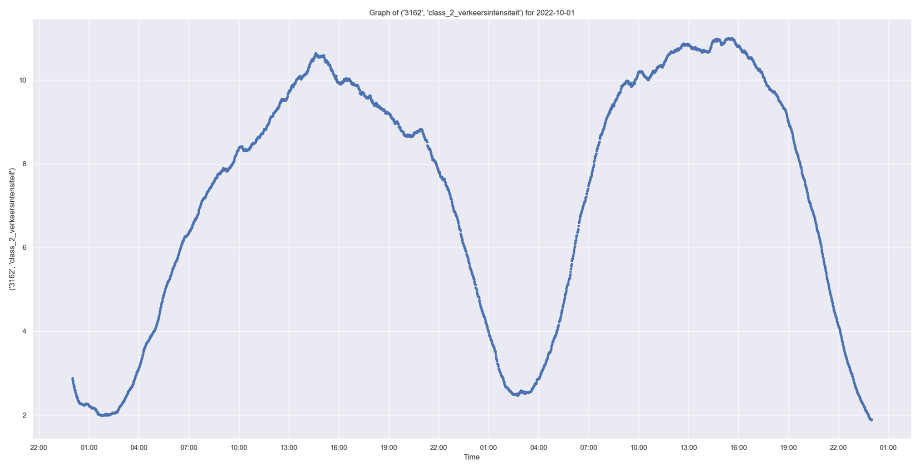


Figure 6.5: Sensor ['3162'] Rolling window size=300

6.5 Seasonality

One of the benefits of using a tool like Rolling Windows is that when you apply enough smoothing to the data, it becomes easier to detect seasonality. So Figure 6.6 illustrates the data received from the ['3162'] sensor for three weeks in October 2022. Seasonality by days of the week is easy to read on the graph.

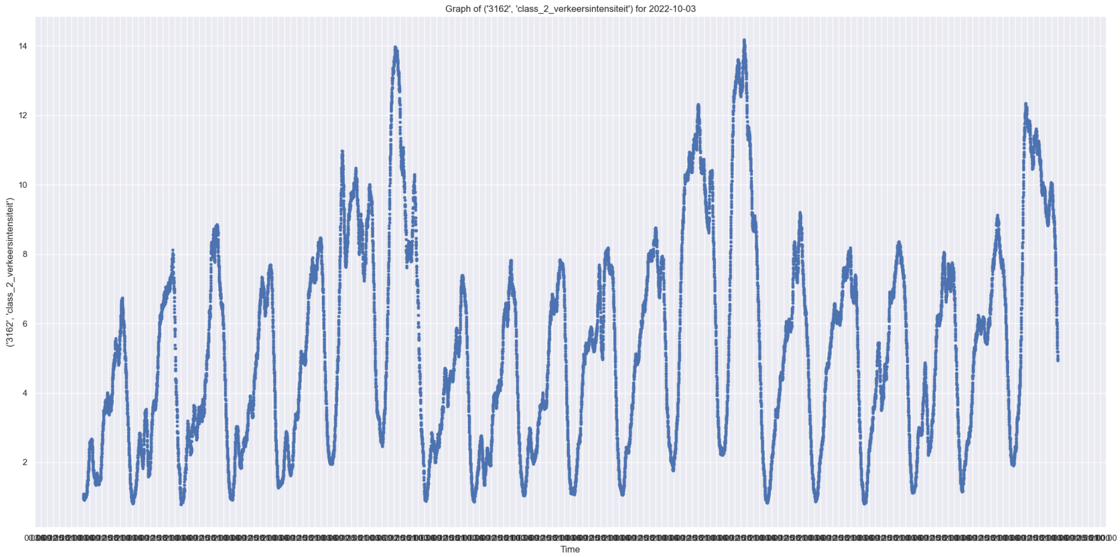


Figure 6.6: Seasonality. Sensor ['3162'] Rolling window size=100

6.6 Grouping by `Ident_8` and `Kmp_Rsys`

The concept of grouping sensors installed on the same road and in the same direction of travel was addressed using two distinct features, denoted as `Ident_8` and `Kmp_Rsys`. Feature `Ident_8` represents a unique road name, while feature `Kmp_Rsys` indicates the position from the beginning of the road. By grouping sensors based on features `Ident_8` and `Kmp_Rsys`, we form clusters of sensors situated on the same road (whether single or multi-lane) and in the same travel direction.

```
Group 1: ['2483' '2482' '2481']
Group 2: ['2485' '2487' '2486']
Group 3: ['800' '801' '802']
Group 4: ['1847' '1846' '1845']
Group 5: ['1891' '1890' '1892']
Group 6: ['2127' '2126' '2125']
Group 7: ['3815' '3816' '3817']
Group 8: ['3803' '3804' '3805']
Group 9: ['2647' '2646' '2645']
Group 10: ['1631' '1630' '165' '166']
...
Group 2325: ['2979' '2978']
Group 2326: ['3041' '3040']
Group 2327: ['2973']
Group 2328: ['2975' '2974']
Group 2329: ['3036']
Group 2330: ['2971']
Group 2331: ['1238' '1237']
Group 2332: ['1234' '1233']
Group 2333: ['1236' '1235']
Group 2334: ['1231' '1232']

Total Groups: 2329
```

Figure 6.7: Grouping by `Ident_8` and `Kmp_Rsys`

As can be seen in Figure 6.7, all sensors installed in the Flemish region and the number of which is 4477 are now grouped into 2329 unique groups.

It's important to highlight that this grouping process occurs along a single direction, as the `Kmp_Rsys` markers differ for different travel directions from the road's origin. This approach markedly outperforms the alternative method of grouping sensors based on feature `lve_nr` described in Section 5.2, which disregarded the direction of travel and primarily relied on physical proximity for clustering. Unfortunately, the latter approach often yielded unsatisfactory outcomes, particularly in complex junctions where sensors from different roads and travel directions could erroneously belong to the same group.

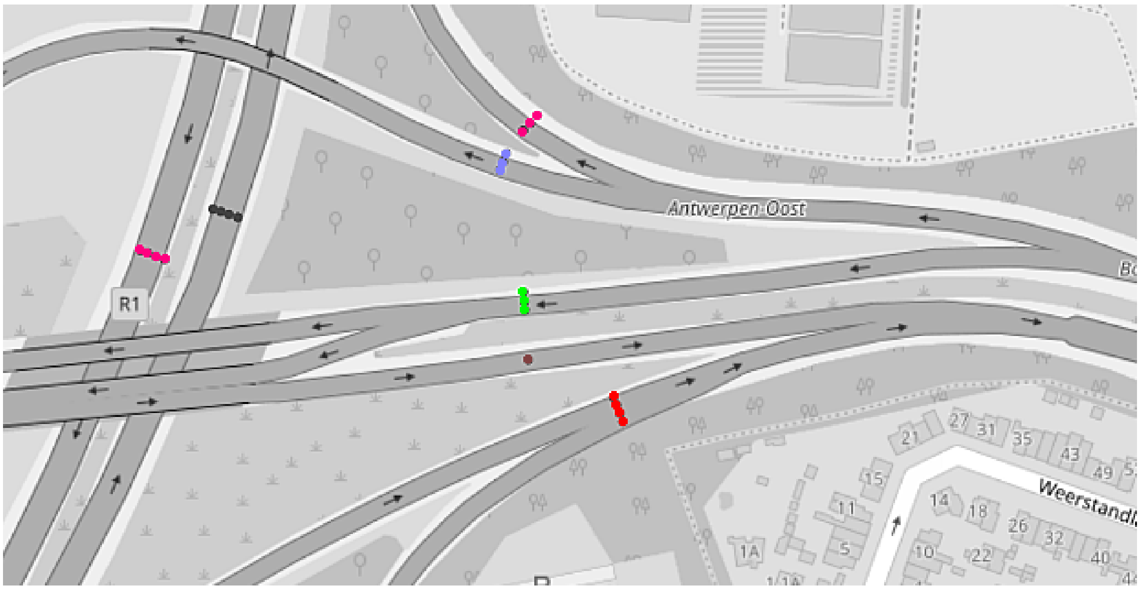


Figure 6.8: Visual representation of the grouping by `Ident_8` and `Kmp_Rsys`

Figure 6.8 clearly shows that now, despite the physical proximity of the sensors and even entering the same connecting hub (`lve_nr`), the sensors are grouped exclusively along different roads (`Ident_8`) and direction of movement (`Kmp_Rsys`). Different groups of sensors in figure 8 are indicated by different colors.

6.7 Mean harmonic speed

Since the goal at this stage is reducing of spatiotemporal data, it was decided to combine data from four classes of vehicle into one according to such parameters as intensity (`'verkeersintensiteit'`) and harmonic velocity (`'voertuigsnelheid_harmonisch'`).

It should be noted that the calculation of the average harmonic velocity has its own characteristics. As already mentioned in the Chapter 6.3, the harmonic velocity formula has the following form:

$$v(x_i, t_i) = \frac{n_i}{\sum_{k=1}^{n_i} \frac{1}{v_k}} \tag{6.1}$$

where x_i - location, t_i - given time period, n_i is number of vehicles pass by given sensor, v_k - is the velocity of vehicle k during the time interval.

The calculation of the average harmonic speed for all classes of vehicles cannot be simply achieved by computing a regular arithmetic mean. To address this challenge, we employ the following method. Despite lacking specific information about the speed of individual vehicles, we have access to the harmonic speed for each vehicle class every minute, along with the count of vehicles whose speeds were used by the system to compute their respective average harmonic speeds.

Below we give the formulas for the average harmonic velocity that was used in the data processing to calculate harmonic mean velocity for different vehicle classes (Class 2, 3, 4, 5) :

$$h = \frac{n}{\sum \frac{1}{v_i}} = \frac{n_1 + n_2 + \cdots + n_i}{\sum\limits_{v_i \in V_1} \frac{1}{v_i} + \sum\limits_{v_i \in V_2} \frac{1}{v_i} + \cdots + \sum\limits_{v_i \in V_3} \frac{1}{v_i}} \tag{6.2}$$

$$h_j = \frac{n_i}{\sum\limits_{v_i \in V_j} \frac{1}{v_i}} \rightarrow \sum\limits_{v_i \in V_j} \frac{1}{v_i} = \frac{n_i}{h_j} \tag{6.3}$$

therefore, the average harmonic velocity will be equal:

$$h = \frac{n_1 + n_2 + \cdots + n_i}{\frac{n_1}{h_1} + \frac{n_2}{h_2} + \cdots + \frac{n_i}{h_i}} \tag{6.4}$$

where n_i - number of vehicles per minute, and hi - harmonic mean velocity per minute, calculated for n_i number of vehicles, all these values are known to us.

6.8 Workflow

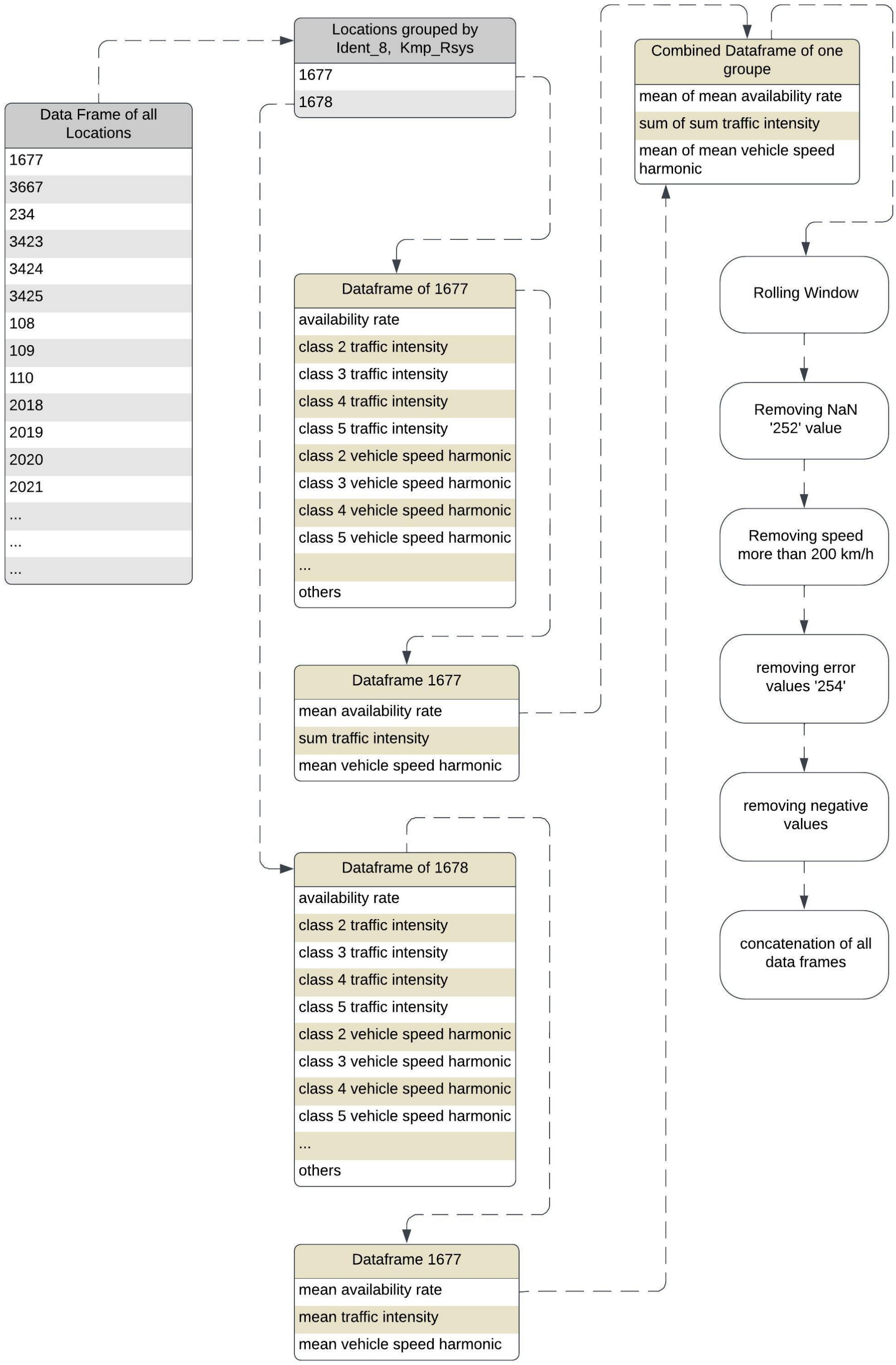


Figure 6.9: Data Processing Workflow Flowchart

In Figure 6.9, you can see the stages of preparation and processing of transportation data for the Flemish region.

Chapter 7

Results

The key outcome of my internship experience is the preparation of a transportation traffic database for the Flemish region. Through this work, a significant reduction in data volume was successfully achieved. For instance, raw data solely from the Big Brussels Ring (R0) since April 2021 amounted to 26.3 GB, which is just 4.5% of the total number of sensors installed across the entire Flemish region (193 out of 4477 items).

The reduction in data volume was accomplished by consolidating data from sensors placed on the same road segment and operating in the same direction (sensor grouping). Furthermore, the decision was made to abandon the categorization of vehicles based on their dimensions (`'verkeersintensiteit'`).

Even at this initial stage, efforts were made to eliminate outliers, smooth data (Rolling indow), and reformat the data structure. As a result, the processed data can now be employed to uncover hidden patterns, seasonal trends, and any other tasks that may be of interest to all stakeholders involved.

Chapter 8

Discussion

From the very beginning, it's worth noting other works related to the analysis of traffic in the Flemish region, such as the work[3] by Tom Van den Bulck , which utilizes the real-time distributed data processing system APACHE STORM.

Of particular interest, in my view, is the methodology proposed by my supervisor Michiel Dhont, where six different traffic states are introduced for further traffic data analysis, such as:

A) Traffic build-up B) Free-flow C) Stable non-saturated traffic D) Traffic intensity reduction E) Congested traffic F) Variable non-saturated traffic.

Utilizing data clustering based on these states, patterns of traffic movement are derived. Essentially, this results in time series data consisting of letters (traffic states), which can be treated as a sequence and processed using the TF*IDF approach.

This becomes particularly intriguing considering that the majority of techniques presently employed to cluster spatio-temporal data rely on density-based clustering methods. For instance, DBSCAN scrutinizes the entire dataset and classifies each data point as a core object (objects situated within a cluster), border object (objects located at the cluster's perimeter), or noise (objects not belonging to any cluster) by assessing the number of objects within a specific distance of each individual object.

I'm looking forward to witnessing the progression of this compelling project by the EluciDATA Lab.

Chapter 9

Conclusion

During my internship at Sirris' division EluciDATA Lab, I successfully tackled several assigned tasks within the Lab's laboratory environment:

- I engaged in a comprehensive study and analysis of data collected through induction loops deployed on roads across the Flemish region.
- I delved into the intricacies of the database's labeling system, which included road marking, lane delineation, traffic direction, and more.
- I devised a workflow capable of handling substantial data volumes in batches, streamlining the process of data retrieval and condensing the acquired information.

The outcomes of my work represent the initial phase of the broader project centered around spatiotemporal data analysis in the Flemish region. This phase will lay the foundation for the subsequent application of an algorithm focused on uncovering patterns, inspired by the research conducted by my mentor, Michiel Dhont, titled "Mining of Spatiotemporal Trajectory Profiles Derived from Mobility Data."[1]

The journey was not without its challenges:

- The traffic data collection system in the Flemish region lacks the robustness found in systems like Google's GTFS (General Transit Feed Specification), leading to extended periods devoted to data comprehension.
- The current system primarily serves real-time traffic status presentation, unlike prolonged data analysis. The Flemish Traffic Center website[2] vividly illustrates ongoing traffic congestion, roadworks, and accidents.
- Lack of open data services complicates access to data from various time periods. Laboratory Lab's efforts were instrumental in accumulating data starting from April 2021. However, having the possibility to obtain data from an earlier period would allow for a more in-depth analysis.
- Managing large data volumes necessitated batch-based workflow creation, a common practice when dealing with such extensive datasets.

[1] https://ieeexplore.ieee.org/document/10031212

[2] http://www.verkeerscentrum.be/

Despite these challenges, there were positive aspects:

- Leveraging prior experience from Lab's initiatives enabled me to apply techniques honed while studying similar datasets from other sources, such as traffic video surveillance in the Brussels Small ring (R20 and N0).
- The supportive EluciDATA Lab team, spanning different cities in Belgium, extended assistance and encouragement throughout my research, regardless of whether I was based in Brussels or Ghent.

I am gratified that my internship's results will contribute to the Lab's ongoing efforts. The data pertaining to Belgium's transportation system, embedded in an extensive road network, demands continuous research. This endeavor aims to optimize traffic flow, making the transportation system more efficient and environmentally friendly.

Chapter 10

Further research

The aim of this internship was to prepare the data in order to continue with understanding the characteristics of traffic on Flanders highways.

In a next phase, the goal is to replicate the traffic analysis methods applied on another dataset, covering the small ring of Brussels [1]. In this analysis, meaningfull traffic states were extracted, and every location got assigned one interpretable label every 15 minutes. The TF-IDF score allows for the identification of patterns in two distinct directions: over a time (along the columns) and over a space (along the rows). Furthermore, we aim to elucidate traffic state transition behaviour using cascades of Markov chains as has been done in [4].

Although the data format of the Flanders highways is very similar (after preprocessing), results are expected to be different because traffic on highways might contain different types of users and reasons for transportation.

The ultimate goal will be two-fold: I) validate the traffic analysis methods applied on another dataset, II) analyse difference in traffic behavior on highways vs normal roads.

All of these developments align with the plan set forth by the Flemish region under the "Slimme mobiliteit"[1] program, which includes the task of implementing Intelligent Transportation Systems (ITS)[5]. These research is being activities are being conducted under the XXX program and the

These research studies are being conducted as part of project MISTic by the Brussels-Capital Region – Innoviris[2] and and secured financial support from the Flemish Government's AI Research Program.[3]

[1]Smart mobility

[2]https://innoviris.brussels/

[3]https://www.flandersairesearch.be/en

Bibliography

[1] M. Dhont, E. Tsiporkova, and N. Gonz'alez-Deleito, “Mining of Spatiotemporal Trajectory Profiles Derived from Mobility Data. Publisher: IEEE,” Orlando, FL, USA, Nov. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/10031212.

[2] A. G. F. MIET, “Traffic signals - part 6 - vehicle detection,” 2014. [Online]. Available: https://www.linkedin.com/pulse/20141005200312-83983163-traffic-signals-part-6-vehicle-detection/.

[3] T. van Den Bulck, “Streaming traffic data with Spring Kafka & Apache Storm,” 2019. [Online]. Available: https://ordina-jworks.github.io/streaming/2019/03/25/streaming-traffic-data.html.

[4] M. Dhont and E. Tsiporkova, “Elucidating Transition State Behaviour from Mobility Data by Cascades of Markov Chains”. Publisher: Canadian Artificial Intelligence Association,” 2023. [Online]. Available: https://caiac.pubpub.org/pub/jv5jhck0/release/1.

[5] R. Venken, B. JANSSENS, B. WOLPUT, and K. LYEN, “VLAAMS STRATEGISCH ACTIEPLAN 2030,” p. 61, 2018. [Online]. Available: https://assets.vlaanderen.be/image/upload/v1590770849/ITS-actieplan-Visiedocument_brchhy.pdf.