Atieh Talebiahooie

Data-Driven Analysis of Sudden Braking Events in Helsinki’s Public Transport

Metropolia University of Applied Sciences

Master of Engineering

Information Technology

Master’s Thesis

20 May 2025

PREFACE

My background in civil engineering and water resources management has led me to explore how urban infrastructure, sustainability, and data integration can enhance the resilience of cities. A long time, MY experience in academic and practical engineering, particularly in environmental planning, public infrastructure, and intelligent systems, has shaped my interest in how modern technology can help cities to achieve greater safety and efficiency.

What led me to this topic was a curiosity about how real-time data and Data-Driven analysis could be applied to solve challenges in urban mobility. Living in Helsinki made me more curious about how cities use open data and technology to improve public services. I found it interesting how real-time transport data can explore safety aspects. This project allowed me to explore how these values can be turned into practical tools. It focused on detecting sudden braking in public transport. These moments often mean a near-miss with a cyclist or pedestrian. I have used a real-time public transport system to study this.

Before starting this thesis, I was unsure if I could manage everything quickly, studying and dealing with new challenges in daily life. It has not always been easy, but it has been possible. I would like to sincerely thank my family and loved ones, whose quiet strength and steady support carried me through each step of this work. Their encouragement, even during the most demanding moments, made this journey possible and meaningful.

I warmly thank Ville Jääskeläinen, our program director and supervisor, for his valuable guidance, support, clear feedback, and encouragement throughout my studies; special thanks to Amir Dirin for his kind help and for introducing me to Forum Virium. I would also like to thank the Forum Virium Helsinki team for providing access to HSL’s real-time datasets and their progressive work in innovative city development.

Lastly, I dedicate this thesis to all those working towards smarter, safer, and more livable cities.

Helsinki, Finland. 20 May 2025.
Atieh Talebiahooie

Abstract

Author:	Atieh Talebiahooie
Title:	Data-Driven analysis of Sudden Braking Events in Helsinki’s Public Transport
Number of Pages:	50 pages + 2 appendices
Date:	20 May 2025
Degree:	Master of Engineering
Degree Programme:	Information Technology
Professional Major:	Networking and Services
Supervisor:	Ville Jääskeläinen, Principal Lecturer

This thesis explores how real-time data from public transport services can help to identify and predict sudden braking events in Helsinki’s bus network. Sudden braking is often a sign of risky traffic situations, especially for pedestrians and cyclists. By studying these moments, we can better understand where and when safety issues are likely to happen, and what might be causing them.

The research uses open data from Helsinki Region Transport (HSL), focusing on vehicle movement and braking patterns over multiple days. By analysing the time, location, and vehicle behaviour during these events, this research aims to detect patterns that show risk areas. Bus location and movement data were combined with weather conditions to detect patterns behind sudden braking. The outcomes are supported with visualizations, such as heatmaps and cluster-based spatial analysis, to highlight risk zones for sudden braking. These insights could help city planners and traffic safety experts make streets safer, especially for those walking or biking. The goal was to predict when and where a sudden braking event might occur.

The finding emphasizes the importance of using open urban mobility data to improve traffic safety and protect vulnerable road users.

Keywords: Open Data, Real time data, Public Transport System, Forum Virium

The originality of this thesis has been checked using Turnitin Originality Check service.

Contents

List of Figures and Table

List of Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
GPS	Global Positioning System
HFP	High Frequency Positioning
HSL	Helsinki Region Transport (Fi: Helsingin seudun liikenne)
HTML	HyperText Markup Language
ITS	Intelligent Transportation Systems
MQTT	Message Queuing Telemetry Transport
NaN	Not a Number
STRADA	Swedish Traffic Accident Data Acquisition
UTC	Coordinated Universal Time
VRUs	Vulnerable Road Users

1 Introduction

Cities are growing fast, and the number of people using public transportation is also growing. Keeping public transport systems safe and efficient has become more important than ever. In recent years, the availability of real-time transport and environmental data has created new opportunities to better understand how traffic systems operate and how they can be improved. Data-driven analysis and developing pattern recognition can help us make sense of these large datasets, especially when finding patterns that may indicate safety risks.

In urban areas, one sign of possible safety issues is a sudden braking event which can reflect a near-miss scenario. When a bus driver suddenly brakes hard, it often reflects an immediate attempt to avoid collisions with vulnerable road users (VRUs), such as pedestrians or cyclists. Even though these event incidents usually do not appear in official reports, they can still point to places where close calls happen frequently and safety improvements might be needed. By looking closely at when and where these sudden stops occur, and how the weather was at the time, we may be able to predict them in advance and help city planners take action to improve the safety of those areas.

Sudden braking events in public buses may represent near-miss incidents involving VRUs. Such events may pose critical hazards and cause potential injuries for VRUs. In other words, even though they do not always result in collision, they can be a potential clash point. Traditional methods for identifying the most dangerous location depend on historical clash records, which may not be accurate enough. Therefore, leveraging real-time public transport data from Helsinki Regional Transport (HSL) [1] along with other factors such as weekdays, weekends, rush times, other lines of public transport, and environmental data from the Finnish Meteorological Institute (FMI) could improve the prediction of safety risks. This thesis explores how real-time data can be used to predict potential zones for a likelihood of sudden braking events, which are likely near-miss events with VRUs.

This research addresses the following key questions:

- How can real-time data be used to identify high-risk areas for VRUs?
- What speed, location, and vehicle behavior patterns are related to the sudden braking events in buses?
- Where and when do sudden braking events occur?
- What patterns or factors are linked to sudden braking events?
- How can real-time public transport help cities plan more safely?

So, this thesis covers the following topics:

- Analyzing HSL real-time data to detect sudden braking events.
- Exploring spatial and temporal trends in braking events.
- Applying clustering methods to identify meaningful groups in sudden braking events.
- Visualizing the high-risk locations in the Helsinki urban area.
- Offer valuable insights to urban planners and stakeholders.

The study focuses on public buses within the Helsinki Region. Throughout this study, open datasets from HSL have been used. The study looks only at sudden braking events and does not include any other types of traffic events. The analysis depends on the quality and coverage of data and how data-driven methods perform.

This thesis is structured in six sections, each one addressing a key aspect of the study. It begins with an introduction to the problem and motivation behind the research in Section 1. In Section 2, the state of the art of the field is presented. Data sources, cleaning, and data preparation are discussed in Section 3. The primary results are presented in Section 4, including the detection of sudden braking and clustering in detail. The results are visualized through interactive maps and graphs in Section 5. The final Section reflects on the findings and suggests how the results can be used to improve safety in public transport.

2 Method and Material

This Chapter describes the research design, data sources, and methods to analyze sudden braking incidents within Helsinki’s public transport network. A data-driven quantitative approach was focused on real-time vehicle movement records combined within geospatial analysis. The objective is to better understand when and where sudden braking happens most often and find areas where these events are more common. Knowing this can help make public transport safer and support urban planners’ efforts to reduce the chances of accidents.

2.1 Research Approach

Research can be categorized into several categories: basic versus applied research, quantitative versus qualitative methods, and field versus desk studies. Basic research is primarily aimed at expanding theoretical knowledge and understanding, often without a direct connection to practical applications or short-term outcomes [2:9]. Furthermore, Research activities can be conducted through field or desk studies. In Field studies, researchers go into real environments to collect their data and see what is happening firsthand [2:53]. In contrast, desk studies use information already collected by others, so the researcher does not need to be physically present at the scene.

Quantitative methods and exploratory approaches are used in this thesis. The process of quantitative research includes gathering and analyzing data, such as real-time transport movement, by using statistical methods to test hypotheses [2:546]. Quantitative research involves gathering and analyzing numerical data to explore patterns, relationships, or predictions. Therefore, this study adopts a quantitative approach, relying on numerical, real-time datasets collected from official sources. Moreover, exploratory study investigates previous patterns related to bus braking events under varying operational and environmental conditions.

Since the datasets used in this study were collected through official Application Programming Interfaces (APIs) and repositories rather than through direct observation or experimentation, the research is categorized as desk research rather than field research.

Research methodology also involves selecting an appropriate research strategy [2:57]. The choice of research strategy usually depends on the problem being studied, the environment around it, and the type of data needed to find practical solutions. In many business settings, surveys, case studies, and action research are common approaches [3]. Case studies are widespread in business, law, public policy, and healthcare. They help uncover practical lessons by examining how similar organizations or industries have dealt with comparable challenges [3].

According to the Frascati Manual 2015, applied research refers to original investigations carried out to acquire new knowledge, primarily focusing on addressing a specific, practical aim or objective [4]. This study follows an applied research approach, aiming to produce practical analysis to improve urban safety and support urban planning decisions.

While the structure of this work does not fully align with traditional action research, it still reflects the principle of applied research to solve real-world problems. Specifically, the study focuses on enhancing the safety of Vulnerable Road Users (VRUs) by leveraging real-time and contextual data to inform proactive decision-making [5].

This research adopts data-driven techniques to predict when and where sudden braking events will likely to occur within Helsinki’s public bus network. It applies the sudden braking location as a factor for traffic safety risks. By integrating transportation analysis, geographical data sciences, and urban safety planning, the study aligns with broader goals related to innovative city development and proactive road safety planning.

2.2 Research Design

This study adopts a data-driven, exploratory research design that utilizes real-time datasets from the urban transit system to investigate the characteristics, distribution, and patterns of sudden braking events in Uusimaa’s urban bus network. The primary objective is to detect spatiotemporal patterns associated with high-risk areas. These identified patterns are intended to inform road safety measures, particularly about near-miss events involving Vulnerable Road Users (VRUs).

This research integrates principles from traffic safety analytics, unsupervised learning, and geospatial visualization to identify sudden braking events in public transportation. It is grounded in the theoretical concepts of surrogate safety indicators, where non-crash events, such as abrupt deceleration, are used to infer underlying traffic risk [6]. These indicators are particularly valuable in environments with limited or delayed crash data [7]. Previous studies have shown that frequent harsh braking events, reliably identified through vehicle telemetry or smartphone sensors, often correspond to areas with elevated crash risk [8].

To achieve this objective, the study applied a five-stage methodological pipeline:

1- Data Collection: Real-time data from the urban bus network were collected through HSL, and the raw datasets were structured by Forum Virium Helsinki [1,9,10].

2- Data Preprocessing: Location, speed, and acceleration data were cleaned and preprocessed to ensure data quality and spatial accuracy.

3- Event Detection: Sudden braking events were detected by applying a deceleration threshold.

4- Clustering Analysis: The DBSCAN and K-Means techniques were used to group events based on the location of repeated events and the speed and braking intensity.

5- Spatial-Temporal Visualization: The results mapped over area and time offer valuable visual insights that can help urban planners improve road safety.

This approach follows best practices from recent studies, where real-time or crowdsourced vehicle data have been utilized to evaluate transport network performance and predict safety-critical conditions [11].

2.3 Data Collection and Analysis

This section describes the process of data collection and the subsequent data cleaning and preprocessing activities. First, the procedures for collecting real-time transport data are outlined, followed by the methods used to prepare the data for analysis.

2.3.1 Data Collection

The raw dataset collected from vehicle GPS systems was initially in High Frequency Positioning (HFP) format. Forum Virium Helsinki downloaded the HFP data files from the server, and parsed the MQTT[1] Message format, and converted the data into a structured format. Due to the large volume of data, the structured data format selected was GeoParquet [12].

The datasets included timestamped positional updates for each vehicle, and attributes such as positions, speed, and acceleration.

Figure 1 illustrates the service architecture for Helsinki Region Transport's real-time data stream, based on a message broker system using the MQTT protocol [13]. It shows how real-time data were collected from three types of public transport: trams, trains, and buses. All real-time data from these three vehicle modes were transmitted to a centralized MQTT broker hosted by HSL at mqtt.hsl.fi. Once the data reached this broker, it was broadcast to various client

1-a lightweight messaging protocol ideal for IoT and real-time transport applications.

devices, such as mobile apps, monitoring systems, and real-time processing and visualization tools [14].

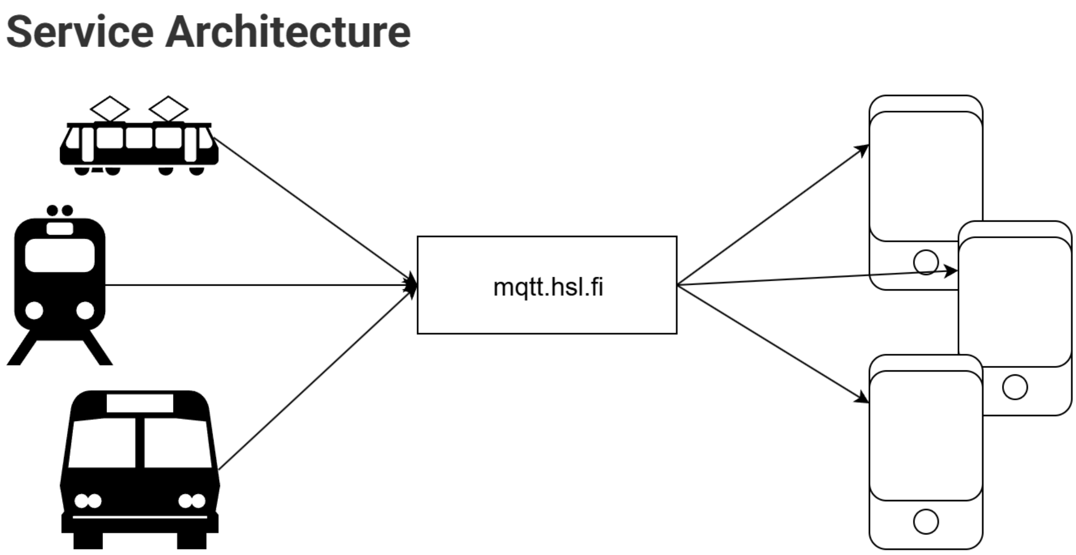


Figure 1. Service Architecture of HSL [14]

For this study, data were collected over 28 days between March and April 2025, covering both weekdays and weekends across 24-hour periods for Route 2200 in the Uusimaa area. Route 2200 was selected due to its length, the number of stations, the crossing of different zones, and the availability of accessible GeoParquet data [15].

2.3.2 Data Cleaning and Pre-Processing

This section describes the procedures used to clean and pre-process the collected data to ensure the accuracy and reliability of the analysis. Several data cleaning steps were performed to enhance the dataset quality:

- GPS Noise Filtering: Sudden jumps in location were detected by measuring the distance between two consecutive points. If the geographic distance was too considerable over a short time, the data point was classified as noise and removed. Additionally, low-speed GPS points were also removed [7]. These patterns often signaled either stationary buses at depots or temporary signal loss.

- Bounding Box Filtering: Since Route 2200, located between Elienlinaukio and Espoonkeskus, is in the Uusimaa area. Spatial filtering was applied. A bounding box was used to retain only those GPS points within the longitude range of 24.4-25.4°E and latitude range of 59.8-60.5°N, ensuring that location errors outside the planned study area were excluded [16].
- Speed Validation: Data records with speed values less than 1.5 m/s or greater than 35 m/s (approximately 126 km/h) were removed. Extremely high speeds typically indicate GPS or sensor errors.
- Handling Missing or Invalid Data: Records containing missing values (NaN) in acceleration, direction, or GPS coordinates were removed. Additionally, the direction variable was validated to include only acceptable values: 1 (forward) and 2 (backward). Further filtering was applied to remove points where the vehicle remained at nearly the exact rounded coordinates across many timestamps while moving slowly.

These preprocessing steps align with approaches used in previous public transportation studies. For example, Digital tachograph (DTG) data has been utilized to investigate passenger fall incidents by detecting abrupt decelerations and stop events using acceleration thresholds. In those studies, a structured data cleaning pipeline was applied to classify risk events, followed by regression modeling to explore passenger safety risk [17].

Similarly, harsh braking detection studies filtered raw telematics data and applied deceleration thresholds to identify unsafe driving behaviors [18].

Overall, the reviewed literature highlights the critical importance of data preprocessing in transportation safety analytics, particularly when managing large-scale sensor datasets. Accordingly, this study applied a systematic cleaning strategy to improve the accuracy and reliability of sudden braking event identification.

3 Data-Driven Traffic Safety

This Chapter presents a comprehensive review of recent research relevant to detecting and analysing sudden braking events in urban transit networks, focusing on safety, vulnerable road users, real-time data analytics, and unsupervised clustering techniques.

3.1 Data-Driven Approaches and Risk Identification

Increasing population mobility and improving cities have required an increase in the number of vehicles on the roads, leading to many challenges for road traffic management. To overcome these problems, researchers in both industry and academia have been working on implementing sensors, communication, and innovative technology to have more efficient traffic management systems [19]. Traditional approaches rely on slow historical crash data and official records, which need years of data to detect dangerous locations. Therefore, data-driven methods, especially in cities, are becoming preferred proactive detection methods. The data from sensors, along with the location and movement of the vehicle, is geospatial real-time data. Real-time data provides professional insights into traffic patterns, driver behavior, and potential collision hotspots [20].

Intelligent transportation systems (ITS) technologies help to identify the changing conditions and mitigate potential risks. The importance of utilizing a multidisciplinary approach is cumbersome, but worth it. When used to detect dangerous locations and improve urban safety, collaboration between data scientists, urban planners, and transportation experts is critical for developing and mixing the traditional methods. Detecting dangerous areas that can cause sudden braking events plays a significant role in improving road safety. Proactive methods, which identify potential safety hazards before near-miss events and crashes occur, can reduce the number of injuries on the road [21]. The result of proactive analysis informs road safety management and project development decision-making.

Data-driven techniques, including big data analytics and AI, can analyze a wide range of inputs such as traffic flows, vehicle telemetry, and near-misses to identify dangerous locations, especially for VRUs such as pedestrians and cyclists. Using various IoT sensors, computer vision, and machine learning models to predict or detect high-risk zones with real-time data is an explosion of research and leads societies to sustainable road safety. These methods allow analysts to detect areas before collisions and crashes happen, which is crucial for VRUs, who are often at high risk. Detecting unsafe crossings or bike-vehicle interactions can inform stakeholders before incidents happen [22].

Various types of sensors have been used in different research, such as proximity, positional, internal, optical, auditory, and light-based sensors. Data from these sensors helps to monitor and analyze road conditions, traffic flow, driver behavior, public transit location, and even predict accidents. More detailed data exists with the increase of GPS-equipped cars and smartphones in the last decade. These mobile sensors, such as GPS in vehicles, can collect movement information on routes that fixed sensors might miss. These intelligent sensors have led to smarter road safety, which can detect obstacles, monitor traffic, track public transportation, or alert authorities of accidents in real-time [22]. The types of sensors used for data collection in research are illustrated in Figure 2.

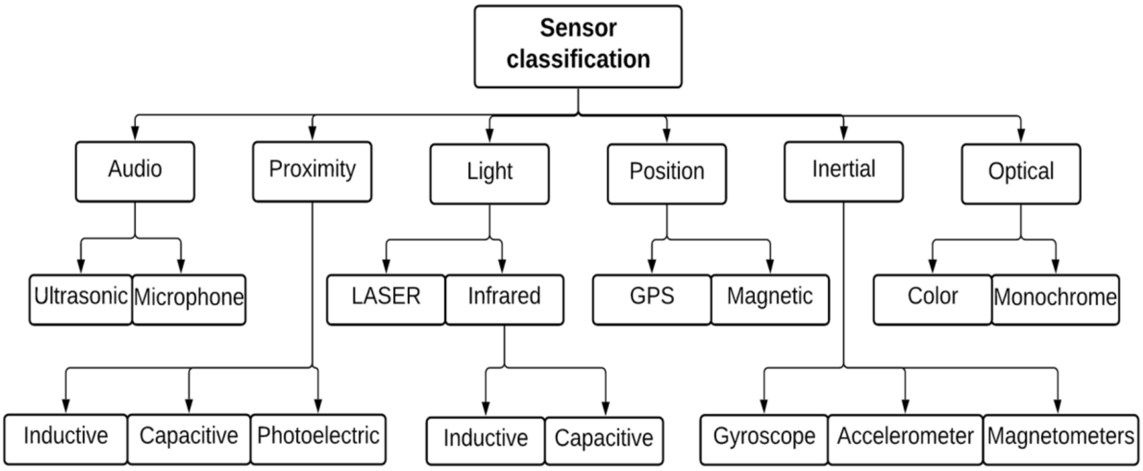


Figure 2. Data Acquisition Sensors [22]

A study focusing on New York City applied a grid-based spatial aggregation method, the Empirical Bayes (EB) approach, and spatial analysis tools, such as global and local Moran’s I. The findings show that near-misses have the highest correlation with crash frequency among all examined variables. Other variables, such as the number of intersections, the bus stops, the road length, the residential land use rate, and the open-space land use rate, significantly impact crash frequency. The estimated near-miss-to-crash ratio of 1957:1 may serve as a benchmark for other cities. The analysis of network variables shows that a higher number of interceptions, bus stops, and longer routes all affect increased crash frequency. On the other hand, open space land use and residential land use were negatively correlated with crash frequency [23].

3.2 Sudden Braking as a Surrogate Safety Measure

A harsh braking is typically a high deceleration by a driver to avoid a potential crash. In one study, 4.5 years (spanning from January 2016 to July 2019) of crash data were analysed with one month of hard-braking data (July 2019) at eight signalized intersections. Geospatial analysis of over a million records linked nearly 7,000 harsh braking events to specific locations. Sensitivity analysis showed that at least four weeks of hard-braking data are essential for reliable correlation with crash data. In this study, statistical modelling showed crashes to rise with increased hard-braking events and traffic volume. This method allows stakeholders to gather valid safety data within a few months. Additionally, histogram bars can identify where hard braking occurs [24].

According to the findings reviewed across multiple transport safety studies, passengers in public transport are at risk of injury even if there is no traffic crash. Harsh braking is one of the most common precursors to non-collision passenger injuries and near-miss events [25].

Sudden braking events in public transport, especially in urban bus systems, are a critical safety concern due to the potential to injure passengers in non-collision situations. These injuries often happen from sudden braking events that cause standing passengers to lose balance or sitting ones to be jolted. Studies across multiple countries confirm the frequency and severity of such events [25].

Silvano and Ohlin analyzed 3.5 years of public bus injury data from STRADA. They concluded that abrupt braking is a major cause of falls among passengers. The study also showed that middle-aged people are disproportionately impacted during the cruising period or transit [26]. The safety advantages of smoother driving practices were also highlighted by Jeong et al., who looked at the digital tachograph data from South Korean buses and showed a substantial statistical correlation between abrupt deceleration occurrences and passenger fall incidents [17].

In the depth study, Barnes et al. examined the causes and contexts of onboard passenger injuries in London buses using a mixed method approach that combined national road traffic incident statistics (STATS19), Transport for London's IRIS[1] datasets, and qualitative analysis of 500 operator-submitted incident reports and 70 CCTV[2]-reviewed events [27]. Their method included thematic coding of textual reports and video analysis to assess driver behaviors, deceleration dynamics, and passenger responses. A key finding from this research was that sudden or harsh braking was the most prevalent cause of non-collision passenger injuries, surpassing other operational maneuvers. Notably, these braking events often occurred in routine (non-emergency) situations, such as late responses to traffic or misjudged vehicle distances [26]. The study highlighted that passengers standing or seated near the wheelchair bay or rear-facing seat were especially vulnerable to being thrown forward during abrupt deceleration, frequently making contact with hard interior surfaces. The findings reinforced that monitoring and mitigating harsh braking is essential for proactive safety improvements in urban bus systems [27].

Driving a bus in cities is challenging as it needs to pay attention to the other road users who sometimes display unpredictable behaviors. Cyclists play an important

[1] -Incident Reporting Information System
[2] -Closed-Circuit Television

role in these interactions. In an online survey study among 639 bus drivers in Santiago, the findings show that younger and older drivers perceive cyclists better than middle-aged drivers [28].

Identifying high-risk locations in transport networks does not require years of crash data. Monitoring sudden braking events has proven to be a quicker and more cost-effective way. Instead of waiting for accidents, this approach helps detect emerging risks, early on, even in places where few or no crashes have been reported. Recent studies have shown a strong link between the frequency of sudden braking and accidents, especially in work zones. Using real-time braking data, transportation agencies can move from reactive to proactive safety management, focusing on improvements where they are needed most [29].

3.3 Deceleration Thresholds for Sudden Braking

Defining a clear deceleration threshold is crucial because it decides which braking events count as hard and sudden braking. In recent studies, sudden braking was determined based on a deceleration threshold of approximately 3 m/s^2 and 5 m/s^2. Although many publications simply report values such as 5 m/s^2 without indicating the sign, this value represents negative acceleration in the context of braking. For instance, the researcher has estimated deceleration events from raw connected vehicle data records around 629 interstate exits in Indiana for three months in 2023. The result showed that deceleration events between -0.5 g and -0.4 g (approximately -4.9 to -3.9 m/s^2) had the highest correlation with an R^2 of 0.69, which is shown in Figure 3. Among all tested deceleration ranges, events between -0.5 g and -0.4 g were the best predictors of where crashes occurred. The framework in this study enabled agencies and transportation professionals to perform safety evaluations on raw trajectory data [30].

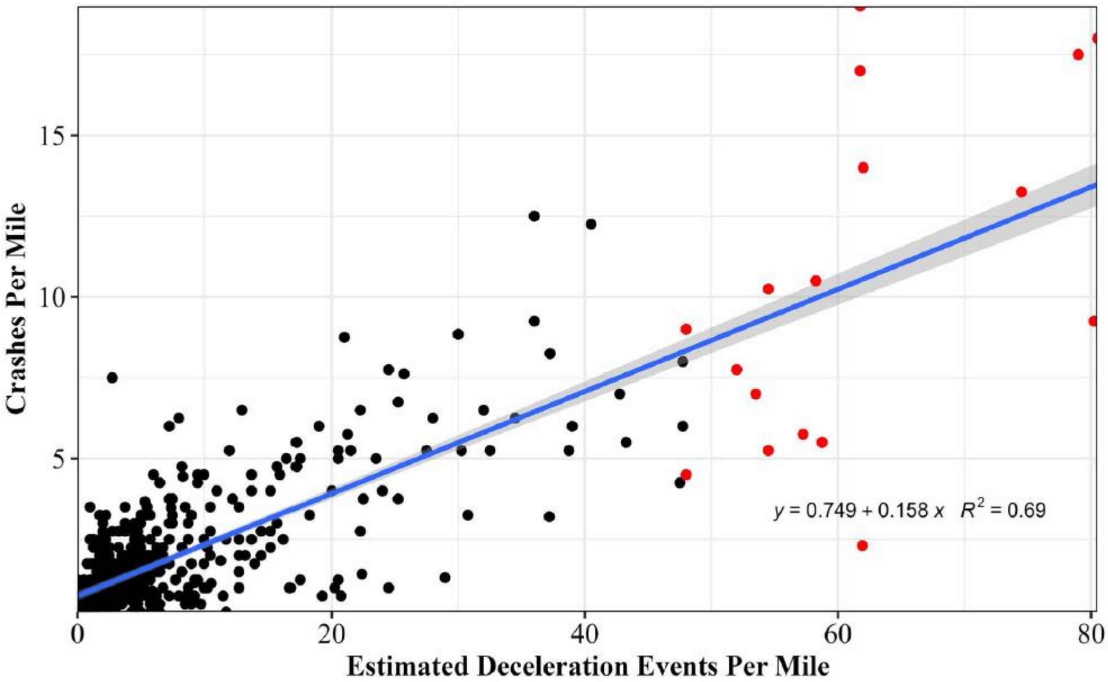


Figure 3. Scatter plot showing crashes per mile versus estimated deceleration events per mile for deceleration in the range of -0.5 g, -0.4 g [30].

In another unsupervised study on sudden braking behavior among heavy passenger vehicle drivers, 8,295 acceleration events and 7,151 braking events were extracted from 142 driving profiles collected using high-resolution GPS instrumentation. This analysis showed that the thresholds derived from clustering the data, the aggressive accelerations ranging between +0.3g and aggressive braking 0.42g and 0.27g (approximately -2.65 m/s^2 to -4.1 m/s^2) — exceeded the acceptable limits for passenger safety [31].

Similarly, a study in the Indiana metropolitan area utilized connected vehicle (CV) data to extract hard braking (HB) events. In this approach, an HB event was identified when the vehicle experienced a deceleration greater than 0.27 g, approximately 2.65 m/s^2, which typically corresponds to a speed reduction of about 18 mph within three seconds.

Vehicle acceleration was derived from the sampled speed and time values between consecutive points along the vehicle's route to detect HB events.

First, the speed difference between two consecutive path points is calculated:

$$\Delta s_i = s_i - s_{i-1} \tag{1}$$

Where:

S_i = speed at the current path point,

S_{i-1} = speed at the previous path point.

Next, the time difference between the same two points was computed as:

$$\Delta t_i = t_i - t_{i-1} \tag{2}$$

Where:

t_i = timestamp at the current path point,

ti-1 = timestamp at the previous path point.

Finally, the instantaneous deceleration at path point i, calculated as:

$$\propto_i = \frac{\Delta s_i}{\Delta t_i} \tag{3}$$

If the deceleration $\propto_i$ exceeds 2.65 m/s^2, the event is classified as a hard braking (HB) event [32].

Based on the review of recent studies, the deceleration thresholds used to define sudden braking events usually range between -2.65 m/s^2 and -5 m/s^2. Lower thresholds, like -2.65 m/s^2, can pick up many regular but quick slowdowns, while higher thresholds (around -5.0 m/s^2) focus more on serious and risky braking events [30,31,32]. This study chose a threshold of -5 m/s^2 to capture better braking situations that are more likely related to real safety risks for vulnerable road users (VRUs). This threshold helps avoid detecting normal driving behavior and concentrates on moments that could represent near-miss incidents. The selected value also matches other recent studies that aim to highlight major safety concerns rather than everyday small changes in speed.

This study applied a minimum speed threshold of 7.0 m/s based on visual inspection of vehicle behavior across the dataset. This value ensured that only deceleration events during active motion were included. It helped to remove speed deceleration events that often occur near stops, which do not represent a safety risk.

3.4 Clustering Approaches for Identifying High-Risk Locations

Clustering algorithms are widely applied in traffic safety research to identify areas where crashes or sudden braking events are concentrated. Rather than focusing on each incident, clustering allows researchers to detect patterns by identifying locations where many high-risk events happen close to one another. Among the many clustering methods, partitioning and density-based techniques are commonly employed in analyzing traffic and transport safety. As a density-based methodology, DBSCAN is well-suited for identifying groups of arbitrary shape and size without changing the number of groups. This flexibility is perfect for datasets where risky events are irregularly distributed. In contrast, K-Means clustering provides a more organized categorization by dividing the data into a collection of groups. This classification is beneficial when relatively clear boundaries between event types or locations are expected.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) has been widely implemented for hotspot detection in traffic safety studies. The main advantages of the DBSCAN lines are the ability to discover groups of different forms and to identify single points that do not belong to any cluster. This capability is particularly valuable in road safety applications, where crash or sudden braking event data is often spatially distributed and uncertain. The DBSCAN approach helps researchers to find significant patterns in safety-critical transportation data by grouping closely located events and treating scattered points as noise.

An interesting example of DBSCAN in traffic safety can be found in a study conducted in metropolitan Athens, Greece. This research collected smartphone-based connected vehicle data from more than 200 drivers with more than 100,000 trips to examine risky driving behaviors across the urban road network. Harsh acceleration and sudden braking events were extracted and spatially clustered with the DBSCAN algorithm to detect critical roadway segments where extreme driving behaviors were concentrated. The analysis generated hazard maps, as shown in Figure 4, which underlined high-risk areas, particularly near intersections and traffic signals, where aggressive driving responses were frequently observed. Consequently, the researchers could localize zones of repeated risky behavior, therefore providing a clear basis for planning targeted interventions and improving urban traffic safety [33].

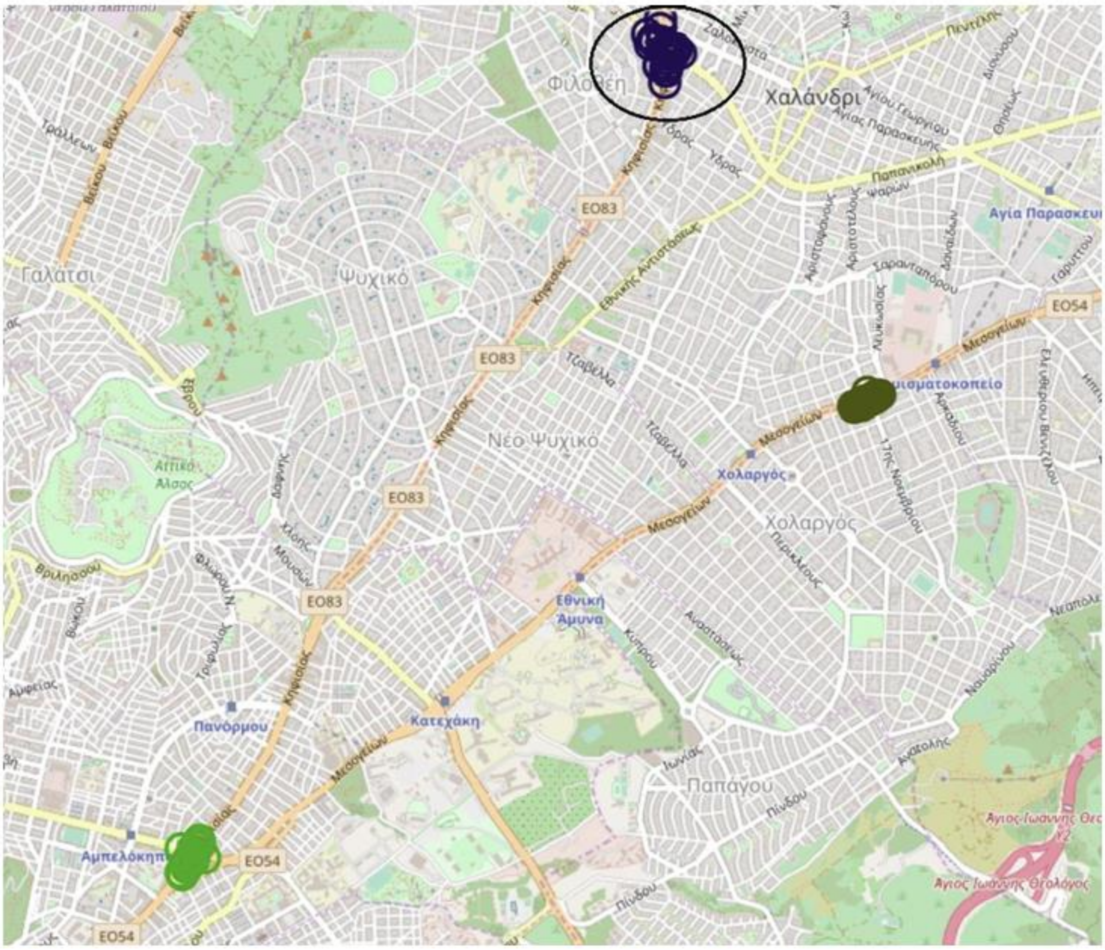


Figure 4. Harsh Braking map with a DBSCAN clustering method [33]

K-Means clustering is another widely used unsupervised learning technique for partitioning a dataset into a specified number of clusters depending on feature similarity. The method operates by iteratively assigning each data point to the nearest cluster center and updating the centers until convergence. The K-Means algorithm’s strength is the simplicity, efficiency, and ability to travel underlying group structures in complex datasets [34]. The K-Means could be used in sudden braking events, such as low-speed versus high-deceleration.

In transportation safety studies, K-Means clustering has been widely used to group crash incidents and sudden braking events. A study focused on identifying locations where cyclists felt unsafe within an urban traffic network. In this research, click-point data were collected from 78 cyclists riding in Lund, Sweden, marking spots where they experienced unsafe situations [35]. An iterative K-Means clustering approach was applied to organize these reports, and was supported by DBSCAN to manage outliers more effectively. The study progressively removed isolated points and adjusted groupings based on how close the unsafe locations were to each other to make the clusters more reliable. This combination helped address some common problems of traditional K-Means, such as differences in cluster size and sensitivity to noise. Local traffic safety experts in Lund then reviewed the resulting clusters, who confirmed that the identified area matched real safety concerns [35]. Through this methodology, Figure 5 illustrates the iterative clustering methods for three different `epsilons` with an outlier detection approach. The DBSCAN method proved more successful in isolating closely packed incidents, filtering out noise, and highlighting critical danger zones. Though it provided more extensive spatial coverage, K-Means included points that may not accurately reflect risk. The red dots are the DBSCAN, and the black ones are the K-Means method. [35].

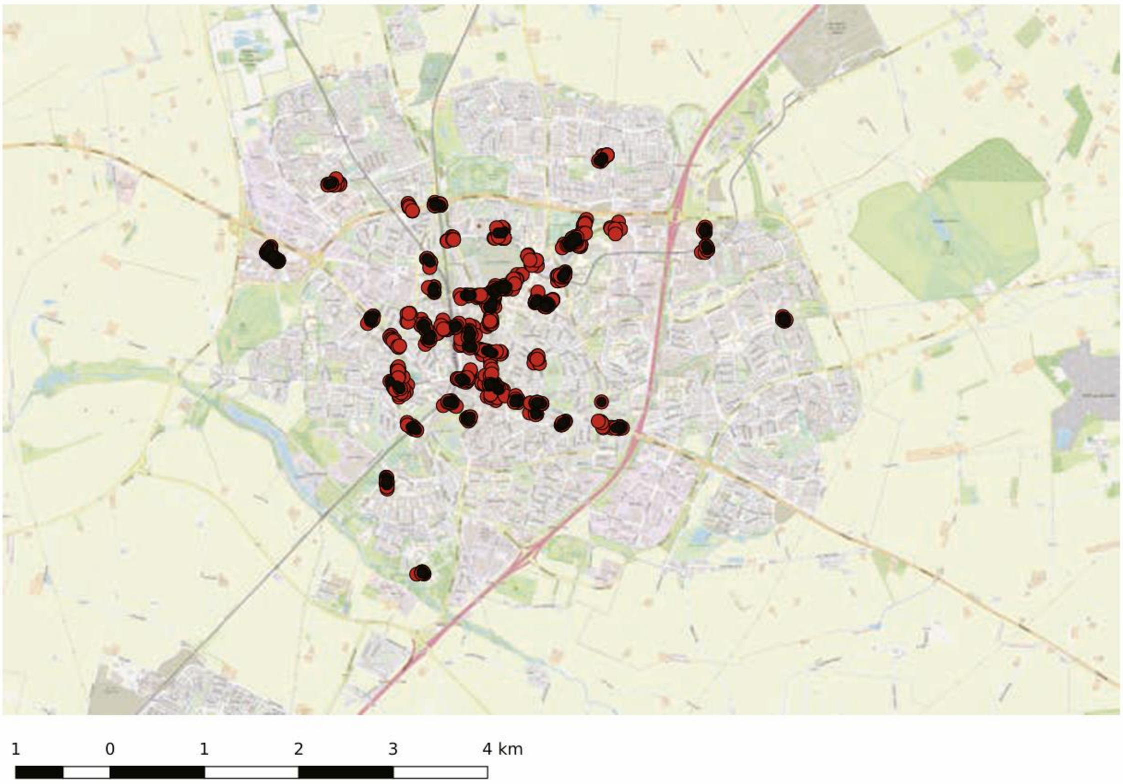


Figure 5. Visualization of K-Means clustering (black dots) with `min_samples=8` and DBSCAN with three different `epsilon` values (red dots) [35]

While K-Means is popular for its simple implementation and quick convergence, it is sensitive to the choice of the number of clusters (K). It assumes clusters to be spherical and identical in size. In contrast, density-based methods such as DBSCAN do not require specifying the number of clusters in advance and are capable of detecting clusters with irregular shapes and varying densities. Nevertheless, K-Means remains a highly effective and interpretable tool in structured tasks such as segmenting roadway sections based on crash risk or braking behavior.

This study, employed K-Means and DBSCAN to analyze sudden braking event data from Helsinki's public transport system. K-Means was used to explore grouping patterns among braking events and spatially distinguish between typical and abnormal braking behavior. Meanwhile, the DBSCAN was utilized to identify natural clusters of sudden braking without requiring prior assumptions about the number of groups and to differentiate significant patterns from random noise.

Combining these two approaches allowed for a more detailed understanding of how sudden braking events are distributed across the city. The analysis identified specific locations where these events cluster, offering valuable insights for improving road safety planning.

3.5 Open-Source Tools and Libraries in Transportation Safety Analysis

Modern transportation safety studies frequently leverage powerful open-source Python libraries for data analysis, modelling, and visualization. In academic research, tools such as GeoPandas, scikit-learn, Seaborn-Matplotlib, and Folium have been validated and are increasingly standard in the analyst's toolbox. These libraries not only accelerate analysis but have been explicitly mentioned in the literature for their roles in reproducing results:

- GeoPandas: As a geospatial extension of Pandas, GeoPandas allows researchers to easily manipulate geographic data such as accident coordinates, road networks, and zone polygons. Its use is documented in road safety studies that require spatial joins and map visualizations. Recent research on traffic accident hot-spot detection in Brazil demonstrated using Python and the GeoPandas library to generate accident concentration maps [36]. Further reviews of geospatial factors applied to road accidents have highlighted GeoPandas as a key Python-based tool for geospatial data analysis, alongside libraries such as Shapely and PySAL [37].

  Overall, GeoPandas serves as a core GIS library in Python for numerous peer-reviewed studies involving collision mapping and infrastructure analysis.
- Scikit-learn: Scikit-learn is a widely cited machine learning library used in numerous transportation safety research projects for crash prediction, severity classification, and clustering tasks. Its reliability and extensive algorithms (K-Means, DBSCAN, etc) have made it a go-to in academic experiments. In research, Scikit-learn was used to develop, train, and evaluate various classification models to forecast pedestrian injury

severity in road accidents and identify the optimal predictive algorithm [38].

- Seaborn and Matplotlib: These libraries are the standard tools for data visualization in Python, and researchers frequently use them to create publication-quality charts for traffic safety data. Although studies may not explicitly mention them in the text, their presence is often evident from the figures and supplementary code. In a study on crash severity analysis of vulnerable road users using machine learning, visualization tasks were conducted using the Matplotlib library to illustrate model performance and feature importance [39]. Seaborn, which is built on Matplotlib, is particularly popular for creating statistical visualizations, such as heatmaps (e.g., correlation matrices of crash factors) and distribution plots (e.g., pre-crash as distributions), which have been widely used across many studies. The consistency and clarity of visuals these libraries produce meet journal standards [40]. Thus, Matplotlib and Seaborn have earned academic validation through their sheer ubiquity, having been used in countless transportation studies to illustrate key findings.
- Folium: A Python library for interactive mapping. It allows researchers to create dynamic maps of accidents or simulation results that can be shared or included in online supplements. While Folium is often seen in industry or data journalism, it has also made its way into academic contexts, particularly for demonstrating results. An AfricaNLP workshop paper on geo-visualization of road traffic collisions in Nigeria explicitly states, was done using Folium, a Python library for geographic data visualization, and the output of geo-parsing was displayed on an interactive leaflet map, which was referred to plotting accident locations from news reports on an interactive Folium map [41]. Including Folium in studies indicates that journals and conferences value interactive visual aids for spatial data. While a printed paper may show a static image, authors often host or provide links to Folium maps so stakeholders can explore high-risk locations in detail. The ease of creating choropleth maps and markers with Folium has made it a valuable tool to complement

static analyses. After producing clustering results, researchers often create interactive city maps that help local authorities quickly recognize high-risk areas. Compared to static figures, these dynamic maps make it much easier to communicate and apply the findings.

In conclusion, open-source Python environments like GeoPandas for spatial data, Scikit-learn for machine learning, Seaborn and Matplotlib for visualization, and Folium for mapping have become a regular part of transportation safety research. They are not just easy to use, but the community has tested them over time. Algorithms in Scikit-learn produce results that align well with those from commercial software, and the maps and graphs made with these libraries are more than good enough for academic publications. So, when working on a data-driven traffic safety thesis, there is every reason to trust these open-source libraries; they have been used successfully to pull out important insights and share them.

4 Analysis of Sudden Braking on an Urban Transit Line

This chapter presents the analysis conducted on real-time vehicle trajectory[1] data from Helsinki's public transportation system. The study focused on identifying potential safety risks for VRUs by analyzing sudden braking events' spatial and temporal distribution. This part illustrates the detection of hard braking events based on vehicle deceleration thresholds, which were defined upon the theoretical foundations and methodological approaches in the previous chapter.

Python and associated geospatial libraries such as Pandas, GeoPandas, Scikit-learn, Matplotlib, and Folium were employed for the analysis. The raw datasets were cleaned to remove GPS anomalies and other noise before doing any meaningful analysis. The filtering included removing wrong GPS points, for instance when a bus seemed to jump several kilometers in just a few seconds. After filtering out such problems, the clustering techniques were explored to identify locations with frequent braking events. Then, sudden braking events were highlighted by visualizing and interpreting sudden braking patterns for illustrating the high-risk areas for more research if needed.

4.1 Data Analysis and Pre-Processing

Before analyzing the patterns of sudden braking incidents, it was first necessary to outline how the real-time geospatial data collected from Helsinki's public transport system was prepared for analysis. The datasets, which include vehicle movement records for multiple bus lines, were used to investigate patterns of sudden braking incidents.

[1] - A Trajectory describes the path a moving object follows, like a bus, over time. Based on GPS location, it shows where the vehicle was at different moments, and helps us understand how it moved along its route.

Data collection spanned a month during March to April 2025 for selected bus line 2200 (200), chosen for its relatively long route, numerous stops, and importance within Helsinki's bus network. These features made it particularly suitable for analysing braking behaviours.

The original format of data was based on the MQTT messaging protocol. The Script *hfp_data_processor.py* from the official GitHub repository was used to parse these MQTT messages and convert them into a structured GeoParquet format, supporting tabular and geospatial attributes [42].

Each record represents a timestamped geolocation and motion status of buses that informs variables such as [12]:

- Timestamp (UTC of record)
- Latitude and longitude (vehicle position)
- Speed (km/h)
- Acceleration (calculated from changes in speed over time, m/s^2)
- Vehicle-ID (unique identifier for each bus)
- Direction of travel (1 or 2)
- Route-id
- Delay
- Door status

Since data preprocessing is a critical phase to ensure analytical validity, several steps were undertaken to clean and prepare the data for analysis. These procedures were implemented sequentially to remove noise, correct anomalies, and ensure that only meaningful data points remained for future analysis. Each filtering step addressed a specific error commonly found in urban transit datasets. The following subsections detail these procedures and their rationale.

4.1.1 Filtering GPS Noise

To ensure spatial accuracy, the raw GPS trajectories were first examined for outliers and positional jumps. Sudden, unrealistic movements, such as a vehicle appearing kilometers away between two short intervals, were identified by calculating the geodesic distance between consecutive GPS points. If the displacement exceeded 1.5 km within less than 3 minutes, the entry was classified as GPS noise and removed. This step effectively eliminated data spikes caused by signal drift or temporary sensor disruptions.

```
# Calculate distance
    veh_df["dist_m"] = veh_df.apply(
        lambda row: geod.line_length(
            [row["geometry"].x, row["prev_geom"].x],
            [row["geometry"].y, row["prev_geom"].y]
        ) if isinstance(row["geometry"], Point) and
isinstance(row["prev_geom"], Point) else 0,
        axis=1
    )

# calculate the time difference in minutes
    veh_df["time_diff_min"] = (
        pd.to_datetime(veh_df["ts"]) - pd.to_datetime(veh_df["prev_ts"])
    ).dt.total_seconds() / 60

# Remove Jumps (>1500 m in <3 min)
    jump_mask = (veh_df["dist_m"] > 1500) & (veh_df["time_diff_min"] < 3)
    cleaned_df = veh_df[~jump_mask].copy()
    cleaned_chunks.append(cleaned_df)
```

Listing 1. GPS jumps filtering

Additionally, records with null values, empty or non-point geometries, and those located at (0,0) typically reflect severe positioning errors.

```
# Drop empty, invalid, or non-point geometries
gdf_all = gdf_all[
    gdf_all["geometry"].apply(lambda geom: geom is not None and not
geom.is_empty and geom.is_valid and geom.geom_type == "Point")
```

Listing 2. An invalid GPS filtering

4.1.2 Filtering with Bounding Box

The study concentrated on a specific public transport corridor in the Uusimaa region; a bounding box filter was applied to retain only the data points within a defined geographical boundary. This filter included only those coordinates with a longitude ranging from 24.4°E to 25.4°E and a latitude from 59.8°N to 60.5°N. This bounding box enabled the elimination of GPS traces caused by route diversions, service deviations, or inaccurate location reporting. The investigation has concentrated exclusively on the movements of operational vehicles inside the selected transit corridor.

```
# Spatial filter. Geographical boundary
gdf_all = gdf_all.cx[24.4:25.4, 59.8:60.5
```

Listing 3. A spatial filtering

4.1.3 Speed Validation

The speed filter was applied to improve the reliability of the data used for braking analysis. First, all records with a speed of 1.5 m/s or less were removed. These usually happen when the bus is parked, waiting at the stop, or is stuck in traffic and does not show functional movement patterns. Subsequently, extremely high speeds above 35 m/s (approximately 126 km/h) were excluded. These values are impossible for public buses and are most likely attributable to GPS or sensor inaccuracies. Removing the impossible values enhanced the accuracy of later calculations, particularly for acceleration.

4.1.4 Management of Incomplete or Invalid Data

Records with incomplete or invalid information were removed to maintain the reliability and integrity of the dataset. These records included data with missing values in critical fields such as acceleration, direction, and position. Only records with valid directional indicators were labeled as 1 and 2 and representing the two operating directions of the bus line were retained. A further filter targeted repetitive low-speed records at the exact location. When GPS coordinates remained the same across multiple successive records, the points were assumed to show either a stationary vehicle or repeated data. They were removed from the datasets since they do not represent real movement.

```
#filter only low speed points
low_speed_df = gdf_all[gdf_all["spd"] <= 0.5]

# Group by vehicle, direction, and approximate location
stuck_counts = (
    low_speed_df
    .groupby(["veh", "dir", "rounded_point"])
    .size()
    .reset_index(name="repeat_count")
)

# Identify "stuck" points repeated at least 15 times
stuck_points = stuck_counts[stuck_counts["repeat_count"] >= 15]
#print(f" Found stuck points (low speed & repeated location):
{len(stuck_points)}")

# Merge and flag rows to be removed
gdf_all = gdf_all.merge(
    stuck_points[["veh", "dir", "rounded_point"]],
    on=["veh", "dir", "rounded_point"],
    how="left",
    indicator="remove_flag")
```

Listing 4. A repeated and low speed filtering

4.2 Sudden Braking Detection

The first step in identifying high-risk braking incidents involved calculating deceleration values based on changes in vehicle speed over time. The acceleration threshold was used to achieve braking events by examining speed changes between consecutive GPS points, as explained in the previous chapter. Hence, it was possible to detect how suddenly the bus slowed down at different moments along its route. Based on findings from recent transport safety studies, several threshold deceleration values have been suggested to identify an abrupt braking event. These values varied from -3 m/s^2 to -7 m/s^2. In this work, a deceleration threshold of -5 m/s^2 was used to identify sudden braking events in the Helsinki urban area. This value offered a balanced compromise for urban planners and was high enough to capture abrupt stops. It is not so low as to include routine slowdowns. So, this threshold helps distinguish sudden braking events, such as those involving sudden reactions to pedestrians or cyclists, from the smoother deceleration patterns during routine traffic flow or bus stops. The acceleration values were calculated using Python libraries such as Pandas and NumPy and applied to the cleaned real-time data collected from bus movements. These acceleration values were stored in GeoParquet format and preserved spatial and temporal attributes. Trajectory variables such as timestamp, geographic coordinates, speed, acceleration, vehicle ID, and travel direction were recorded for each sudden braking value. These records provided input for detecting where and when these events occurred and how they were distributed across time and zone.

Additionally, to finalize the exact sudden braking in line 2200, the braking records obtained from the previous step were filtered. The points with very low or high speeds were removed. Also, the location within the Helsinki-Espoo bounding box and direction within 1 or 2 were applied for the final and exact events.

In addition to the automated filtering steps, a manual filtering step was applied to remove events that did not match typical movement patterns observed in the urban bus network. A set of records showed abnormal deceleration (exceeding -25.83 m/s^2) within a short time. There were sudden braking events repeated from the exact vehicle within a 10-minute window near the bus terminal. These could be due to sensor or GPS drift.

Manual filtering had a significant impact on the result. Some braking events did not match real-world driving conditions, so if these events had been included, they could have affected the accuracy of hotspot detection and clustering results and led to wrong decision-making by urban planners. Removing unreal events made the study sure that the dataset represented real driving behaviors.

4.3 Clustering Analysis

Clustering techniques were employed on the cleaned geospatial braking dataset to explain the trends in the likelihood of sudden braking events, which were identified in the previous step. Two clustering algorithms were applied to investigate different aspects of the data. The DBSCAN was utilized to identify coordinate hotspots, whereas K-Means was employed for speed and acceleration behavior patterns. This method aimed to detect locations within Helsinki’s public transport network where bus drivers must brake suddenly, which may indicate safety hazards for VRUs.

As explained in chapter 3, clustering is widely accepted for identifying hotspots in traffic safety studies. Two algorithms were selected: K-Means, for partitioning the study area into organized groups based on spatial proximity and event frequency, and DBSCAN (Density-Based Spatial Clustering of Applications with Noise), for its capacity to identify naturally shaped clusters and noise from patterns.

The first clustering approach utilized the DBSCN algorithm. This method does not require prior specification of the number of clusters and is well-suited for detecting irregularly shaped groups. The noise that does not belonging to any dense region is filtered out in the clustering algorithm.

The cleaned GeoParquet file was imported and converted into a list of coordinate pairs (x,y) extracted from the geometry column. The `epsilon` parameter, representing the maximum distance between two locations in each cluster, was selected at 0.0015 degrees, equivalent to 150 meters, and `min_samples` was chosen to 5, indicating that each cluster must contain at least five events.

```
# Convert geometry to (x, y) coordinates
coords = gdf.geometry.apply(lambda p: [p.x, p.y]).tolist()

# Apply DBSCAN clustering (eps ≈ 1500m in degrees)
db = DBSCAN(eps=0.0015, min_samples=5, metric='euclidean').fit(coords)
gdf["cluster"] = db.labels_
```

Listing 5. DBSCAN clustering algorithm

Sudden Braking events assigned a cluster value of -1 were classified as noise, but other values indicated specific clusters of braking patterns. The cluster was visualized with Matplotlib, and the plot showed several dense clusters throughout the Helsinki-Espoo line.

The final number of clusters excluding noise was printed as:

```
# Print number of clusters
print("Number of identified clusters (excluding noise):",
len(gdf["cluster"].unique()) - (1 if -1 in gdf["cluster"].unique() else 0))
```

Listing 6. Printing the number of clusters

This method revealed critical urban segments where braking was not random but frequently repeated. Hence, the DBSCAN method identified several dense clusters of sudden braking events along line 2200. These locations can be flagged as potential high-risk zones for future investigation.

While the DBSCAN clustered events according to geographic proximity, K-Means clustering was applied to classify braking events based on speed and acceleration as features. This classification led to the exploration of different braking behavior types, such as high-speed versus high-deceleration, low-speed versus sharp deceleration, or smooth deceleration pattern versus moderate speed. The number of clusters was chosen to be 4, according to visual investigation and interpretative analysis.

```
# ---- K-Means Clustering ----
kmeans = K-Means(n_clusters=4, random_state=42, n_init=10)
gdf["cluster_kmeans"] = kmeans.fit_predict(gdf[["spd", "acc"]])
```

Listing 7. A K-Means clustering

The clusters were visualized using a scatter plot with speed on the x-axis and acceleration on the y-axis.

```
# ---- Scatter Plot for KMeans ----
plt.figure(figsize=(8, 5))
for cluster in gdf["cluster_kmeans"].unique():
    cluster_data = gdf[gdf["cluster_kmeans"] == cluster]
    plt.scatter(cluster_data["spd"], cluster_data["acc"], label=f"Cluster
{cluster}", alpha=0.6)
plt.xlabel("Speed (m/s)")
plt.ylabel("Acceleration (m/s²)")
plt.title("KMeans Clusters of Braking Events")
plt.legend()
plt.grid(True)
plt.tight_layout()
plt.show()
```

Listing 8. A Scatter plot for K-Means clustering

K-Means clustering identified unique braking behavior patterns that may support operational evaluations, such as driver training or route optimization.

4.4 Temporal Analysis of Sudden Braking

A temporal analysis was applied to the cleaned dataset to identify when sudden braking occurs. The Python tools and GeoPandas library were used for parsing timestamps, grouping events by date and hour, and generating aggregated distributions. The sudden braking patterns across days of the week, hours of the day, and working days versus weekends were identified. Weekdays typically involve higher public transport demand, more frequent bus operations, and vulnerable road user flows that may increase the likelihood of abrupt deceleration.

Also, the timestamps were analyzed to extract the hours of each braking event. The temporal distribution illustrated the peak hours during which sudden braking events most frequently occurred.

4.5 Visualization of Sudden Braking Events

Several visualizations were created to identify and understand the trends better. This visualization provided more profound insight into the spatial and temporal distribution and density of braking patterns that were observed along Line 2200 in Helsinki.

All visual outputs were created using Python tools such as Matplotlib, Seaborn, and Folium that followed the event detection process.

4.5.1 Heatmap of Spatial Event Density

A heatmap was developed to investigate the spatial concentration of braking events. Each braking record, represented by a point with latitude and longitude, contributed to the overall density visualization. The heatmap was created with Folium’s Heatmap plugin, which enabled its direct display on an interactive web-based interface.

The parameters of the heatmap, including radius, blur, and opacity, were fine-tuned to improve the heatmap’s quality. These modifications highlighted zones with recurrent braking events. In addition, the heatmap helped detect local hotspots of braking incidents without relying on clustering algorithms.

4.5.2 Scatter Plot of Speed and Acceleration

A two-dimensional scatter plot was created to illustrate the distribution at the time of each event. The Matplotlib and Seaborn libraries were used to plot each event along the x-axis (speed in m/s) and the y-axis (acceleration in m/s^2). This plot illustrated a visual overview of the variation in braking intensity across different speed ranges.

4.5.3 Interactive Map with Folium

In addition to static visualizations, interactive web maps were created to facilitate dynamic spatial analysis of braking data. The Folium library was used to create interactive HTML maps, plotting each sudden braking with its exact GPS location.

Several features have been implemented to increase insight and usability. A MarketCluster plugin was used to group nearby braking events, which significantly reduced visual clutter when zoomed out. Each event could be analyzed through pop-up windows to display details such as vehicle ID, speed, and acceleration.

The maps were full browsers, which allowed the user to zoom, pan, and investigate the spatial patterns in detail. Also, the maps were exported as HTML files and used to analyze events without relying on numerical outputs.

While clustering maps, such as those using DBSCAN and K-Means, were addressed earlier, the interactive maps focused on unclustered braking incidents. The main idea was to simplify the detection of spatial braking events.

5 Result of Sudden Braking Analysis

This part presents the main findings based on the processed data on sudden braking events, building on the methodology introduced in Chapter 3 and the analysis steps outlined in Chapter 4. The main objective of this chapter is to determine the principal results of the investigation and highlight the patterns of braking events. These findings describe the temporal and spatial distribution of sudden braking incidents and their potential impact on safety in the urban transport environment.

The results are based on the final filtered dataset, which contains valid and confirmed braking incidents. The analysis focused on identifying trends, such as daily and hourly variations, spatial concentrations, high-risk zones, and driving behavior, such as speed and deceleration. The charts and visual outputs are used to analyze the results.

5.1 Overview of Sudden Braking Events

This section presents an overview of the final datasets used to analyse sudden braking events in Helsinki's public transport network. The records included here represent harsh deceleration events that remained after cleaning, filtering, and manual validation processes. These datasets were variable for the spatial, temporal, and clustering analyses.

The sudden braking events were detected based on a deceleration threshold below -5.0 m/s^2 and a minimum speed threshold of 7.0 m/s, approximately 25.6 km/h, to ensure the vehicle was in motion and the event reflected an abrupt stop. Direction validity and spatial location within the Helsinki-Espoo bounding box were also applied. This step resulted in 188 braking events. Another validation step removed events within speeds outside the realistic speed range, which was below 1.5 m/s or above 35 m/s. After applying these conditions, the number of events was reduced to 104.

Additionally, another filter was added to remove extreme outlier deceleration values. Events with deceleration values below -15 m/s^2 were removed as such values are impossible for urban buses and are probably related to GPS anomalies.

A manual filtering process was performed to verify the accuracy of the final dataset. This phase involved the visual inspection and identification of inaccurate positive braking incidents from vehicles with IDs 943, 914, and 915. certain records represented trends inconsistent with standard bus operations. For example, a few data points included anomalous deceleration levels exceeding -20.0 m/s^2 or recurrent braking points within a few seconds, even though the bus was in motion or completely stopped. On March 14, 2025, several zones' records from vehicle 943 were generated within a tiny area and a brief time frame. Similar anomalies were observed in vehicles 914 and 915 data on March 20, where abrupt acceleration-deceleration occurred without any movement and were subsequently removed.

After completing this manual filtering, the final dataset comprised 86 validated sudden braking events. These events were stored in the GeoParquet format file for subsequent analyses. In Table 1, the details of the sudden braking records were summarized.

Table 1. The details of the Sudden Braking incidents

Metric	Value
Total Braking events	86
Data collection period	March-April 2025
Maximum deceleration recorded	-10.68 m/s^2
Maximum speed recorded	20.89 m/s
Minimum speed recorded in the event	7.05 m/s

These final records represent the most significant and valuable data on sudden braking for identifying patterns and high-risk locations and understanding key traffic safety concerns in urban environments.

The maximum deceleration observed in the final dataset was -10.68 m/s^2, the value in the upper range of what is acceptable for emergency braking in the urban transit system. These kinds of deceleration values are unusual. However, they can happen when the driver stops quickly to avoid a possible crash. The highest speed recorded during braking events was 20.89 m/s, approximately 75 km/h, which is reasonable for some parts of the Helsinki-Espoo corridor. The lowest speed among the sudden braking was 7.05 m/s, aligning with the lower threshold applied during filtering to ensure that in-motion records were used in the analysis.

Figure 6 shows the distribution of deceleration values within 86 braking events in the Line 2200, Helsinki urban transit system. Most of the events are concentrated between -5.0 and -7.0 m/s^2, which indicates that the sudden braking events belong to the range for urban buses.

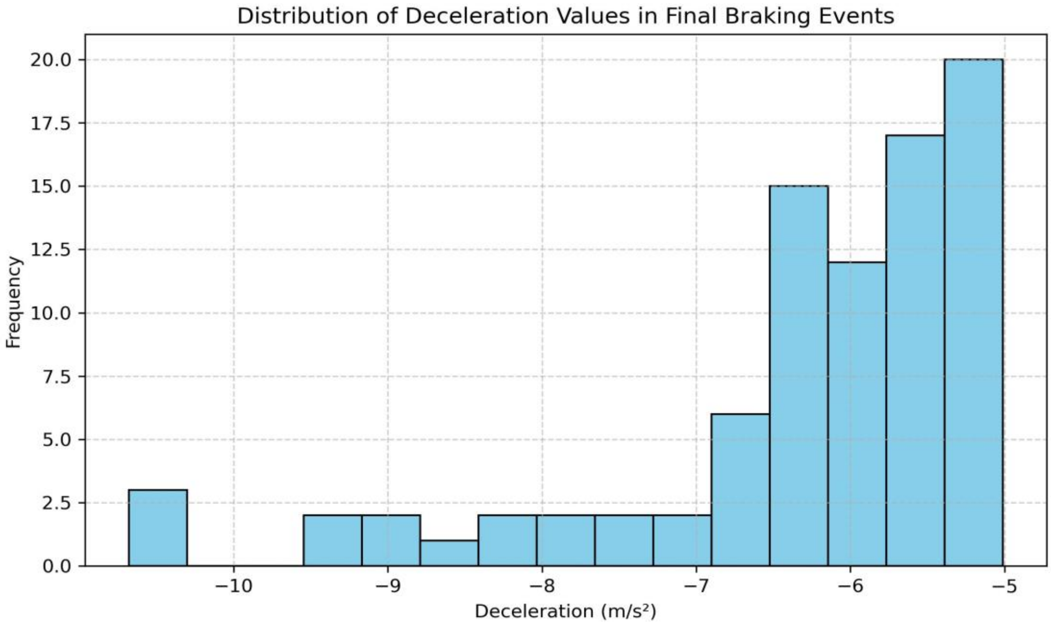


Figure 6. Distribution of deceleration values in braking events

This histogram supported the decision to use -5.0 m/s^2 as the detection threshold. A stricter threshold, such as -7.0 m/s^2, would have excluded many relevant events that reflected safety patterns. Only a few cases exceeded -9.0 m/s^2, confirming that extremely sharp braking was rare and defined emergencies. The distribution confirmed the effectiveness of the filtering approach.

5.2 Temporal Distribution of Braking Events

To identify the temporal pattern of sudden braking events, some temporal attributes were extracted from the final dataset. The braking events were grouped by hour, day of week, and weekday versus weekend classification. The distributions were visualized by histograms, bar charts, and pie charts to recognize daily and weekly patterns in braking events.

5.2.1 Hourly Distribution

The histogram of hourly braking events in Figure 7 indicates that most events took place between 9:00 and 13:00, with a strong peak around noon. This timeframe reflects when buses operate at higher frequencies and encounter pedestrian and cyclist traffic. The decrease in events after 16:00 also corresponds with the natural operation of urban bus activity in the evening hours.

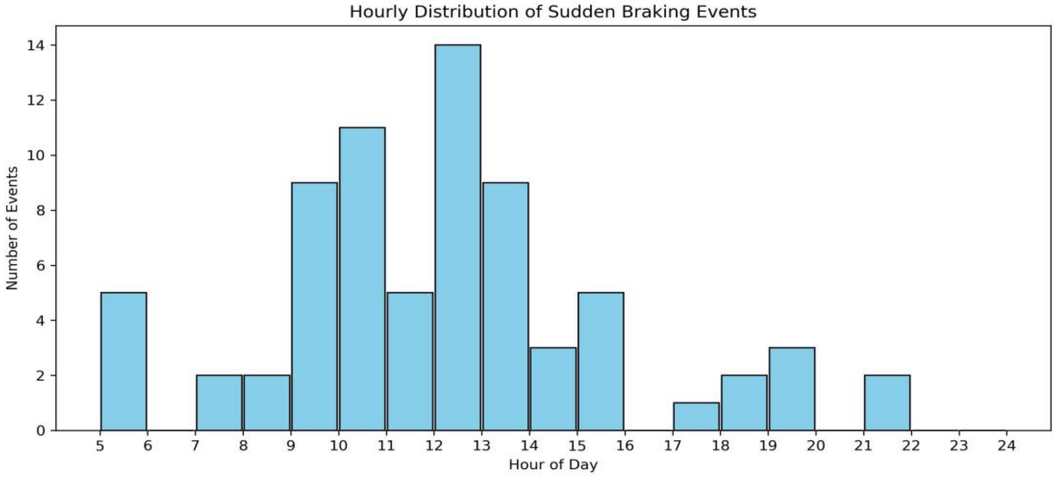


Figure 7. Hourly Distribution of Sudden Braking Events

The pie chart in Figure 8 presents a breakdown of sudden braking incidents across time slots during operational hours from 5:00 to 24:00. A clear concentration is observed during the 09:00 to 13:00 period, which is 53.4% of all recorded events. This increased percentage reflects the morning urban activity peak, when public buses travel regions that have high demand, increased pedestrians, cyclists, and vehicle traffic.

The 13:00-17:00 interval, including 23.3% of events, shows sustained traffic flow during the early afternoon hours. Braking events are reduced during the early morning hours of 05:00 to 9:00 and the late evening hours of 20:00 to 24:00.

The time distribution of sudden braking events is valuable for urban planning and identifying high-risk areas and intervals. Improving urban safety by adjusting speed restrictions, increasing visibility for pedestrians and cyclists, or altering driving behavior may prove more successful during high-risk intervals, such as mid-morning and early afternoon on weekdays.

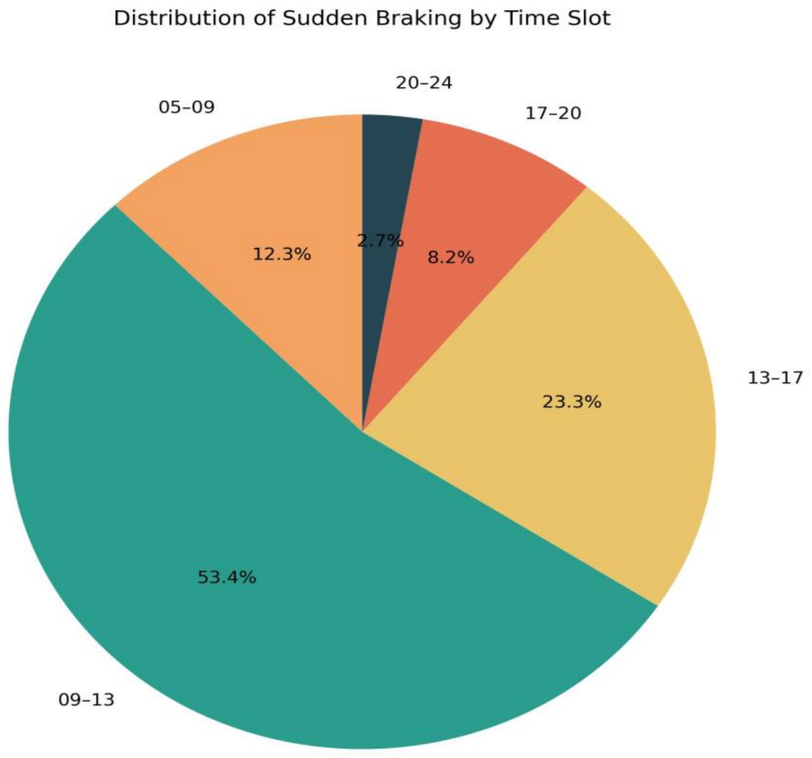


Figure 8. Hourly pattern analysis

5.2.2 Weekly Distribution

According to the weekday bar chart in Figure 9, the most sudden braking incidents occur from Monday to Friday, with Friday recording the highest frequency of occurrences. This graph explains the peak-hour weekdays in urban mobility. In reverse, weekend days (Saturday and Sunday) showed significantly fewer events, reflecting reduced service frequency and decreased pedestrian and bike traffic near crossings and intersections.

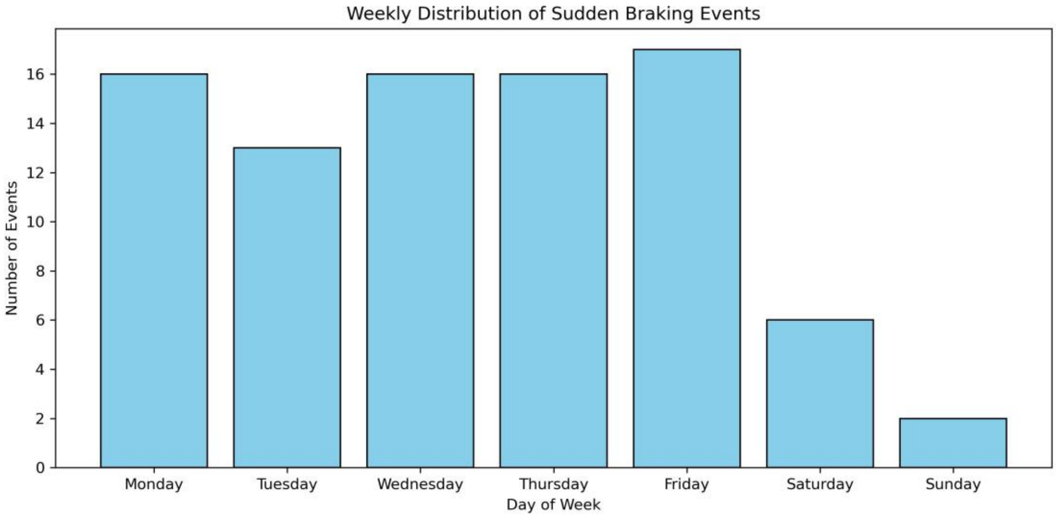


Figure 9. Weekly distribution of sudden braking incidents

5.2.3 Weekdays versus Weekend Distribution

As mentioned in the previous part, a separate bar chart shown in Figure 10 highlights that most events occur on weekdays. Nearly 90% of all sudden braking events took place on weekdays. This high percentage of weekday incidents validates the observation that urban safety risks linked to sudden deceleration are usually concentrated on weekdays. As during weekdays, pedestrians, cyclists, and vehicle traffic are denser.

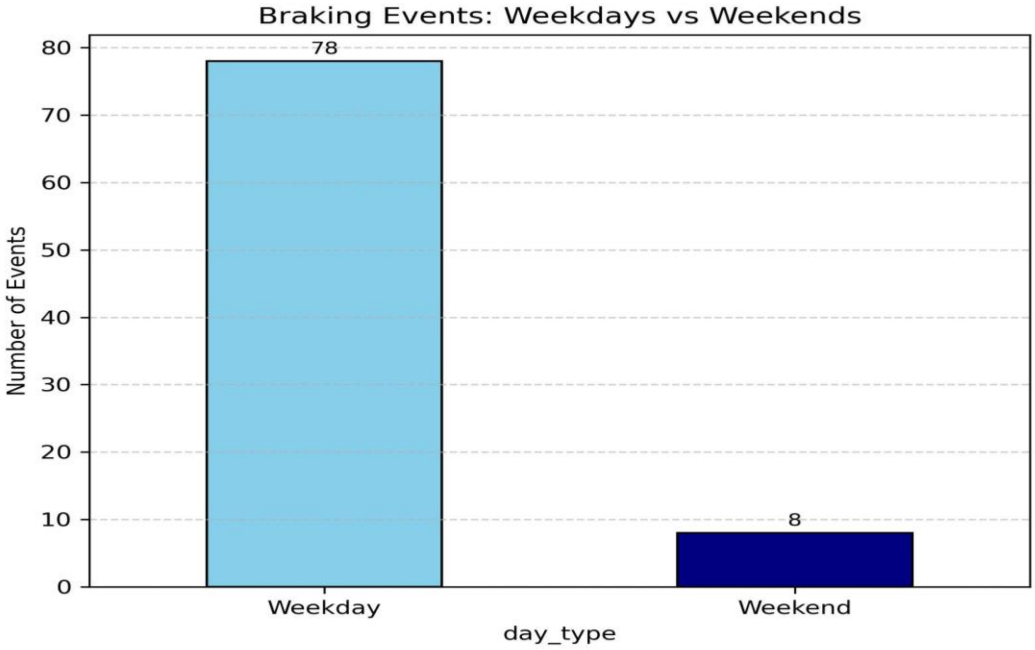


Figure 10. Weekdays versus Weekend Distribution

5.3 Spatial Distribution and Mapping

To better understand where sudden braking events are more likely to occur across the Helsinki-Espoo area, an interactive map was created using the Folium library in Python. The interactive map visualized the sudden braking events as a hotspot marker. Additionally, it was possible to explore where incidents most frequently occur. As shown in Figure 11, areas such as Taka-Töölö and Leppävaara were identified as having a higher density of sudden braking. Hotspots with a high density in all observations may reflect areas with complex traffic environments, pedestrian and cyclists crossing, or public facilities such as schools or shopping centers. Leppävaara serves as a central transit hub and extensive business district, featuring 19 events documented on this hotspot map.

The interactive map enables users to select each brake hotspot and retrieve information, including timestamp, speed, deceleration, and vehicle ID. This ability was enhanced to correlate event context with location.

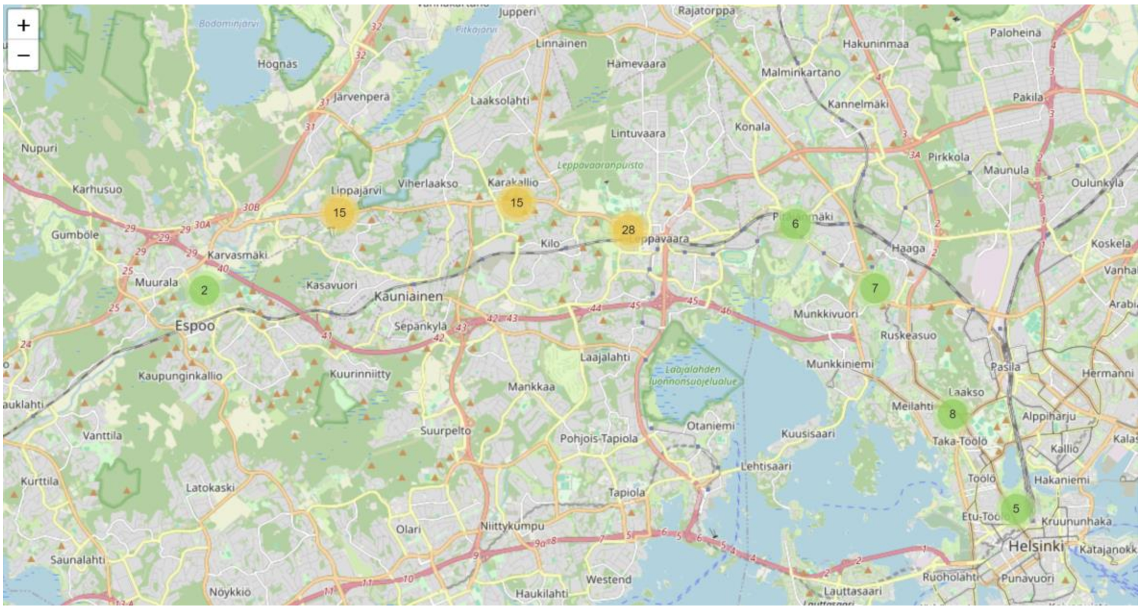


Figure 11. Interactive map with hotspot markers of sudden braking across the Helsinki-Espoo public bus.

The map has been saved as an HTML file and can be opened in any browser for dynamic exploration.

Also, to visualize braking hotspots, the interactive map was enhanced by overlaying the transit route in forward and return directions. This interactive map with a precise analysis of segments, such as repeating events near stops, and intersections, made it easier. This separate view of hotspot markers along the bus route may help to identify issues that are related to the directions.

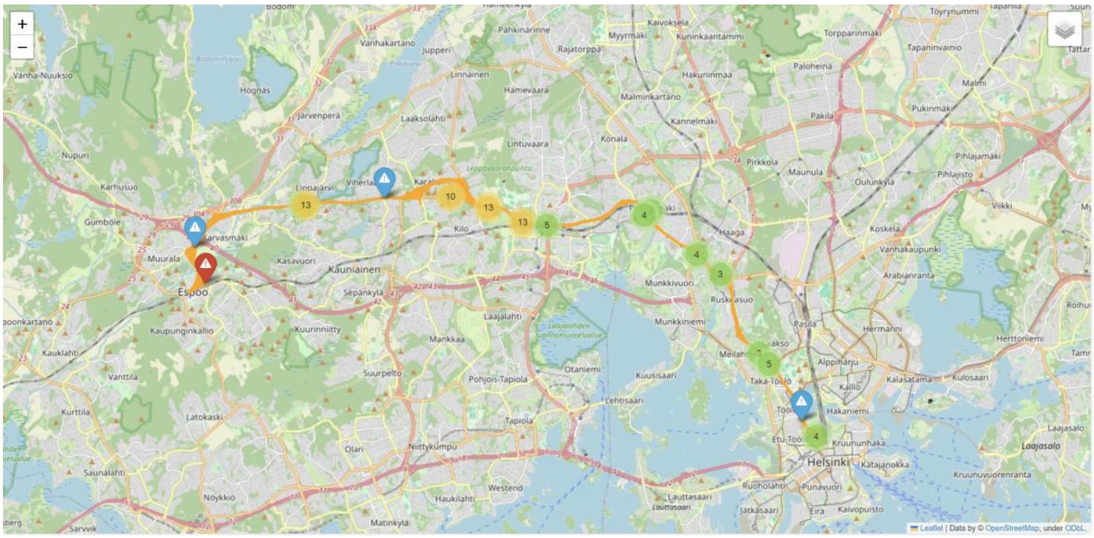


Figure 12. An Interactive map showing hotspot markers of Sudden Braking in both directions.

In addition, a heatmap visualization was implemented to show how sudden braking events are spread out along the transit route. Instead of displaying each event as a separate point, the heatmap used color, from red to orange to yellow, to show the braking intensity. The red corresponds to more braking events and the yellow to fewer events.

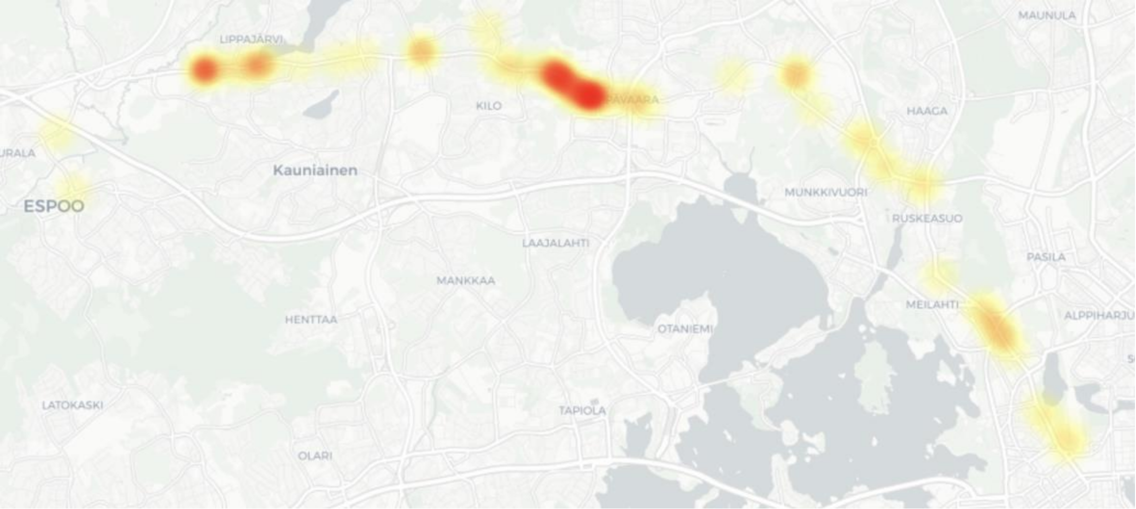
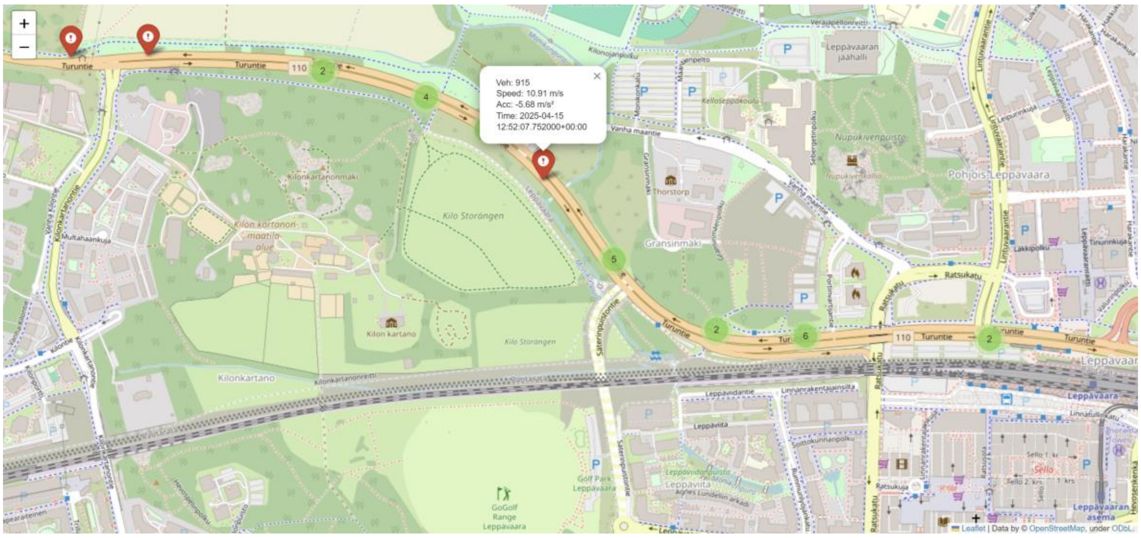


Figure 13. Heatmap of Sudden Braking showed high-intensity segments along the main transit route

The heatmap showed the high number of sudden braking events in the Leppävaara zones. The zoom-in interactive map was created to show more visible details, as visualized in Figure 14. These events are distributed along Turuntie road, which are close to the bus stops, intersections, and areas with many pedestrians and cyclists. Focusing on this area in the map made it easy to know the exact location, vehicle ID, date, speed, and acceleration. With a zoom-in map, it is possible to highlight some areas that need more attention in traffic design.



Veh: 915
Speed: 10.91 m/s
Acc: -5.68 m/s²
Time: 2025-04-15
12:52:07.752000+00:00

Figure 14. Zoomed-in heatmap showing Sudden Braking hotspot and braking details markers in the Leppävaara area

Interactive maps played a significant role in identifying where sudden braking occurred. Different visualizations explained more information to help urban planners and transit stakeholders. By grouping nearby incidents and showing them on the dynamic, browser-based map, it became easier to detect high-risk areas. These visual tools helped focus on specific street segments, such as those close to pedestrian crossings or central transit nodes.

The ability to zoom in and out, add more filters, and view braking event details caused flexibility for quick overview and in-depth analysis.

5.4 Clustering Methods

Clustering techniques were used to better understand sudden braking patterns. Based on the explanation in Chapter 4, DBSCAN and K-Means algorithms were applied to the cleaned datasets to identify data trends.

The DBSCAN algorithm was applied to the geographic coordinates of braking events. This algorithm does not require a specific number of clusters. The coordinates of braking events were extracted from the geometry column and converted into longitude and latitude pairs. The values for the algorithm’s parameters were adjusted through visual inspection. The `epsilon` parameter, which controls the maximum distance between two points of the same cluster, was set to 0.0015 degrees. The `min-samples` parameter was set to 5, defining the minimum number of points in a cluster. Figure 14 shown the three spatial distributions of braking events along the 2200 transit route. Points with label -1 represented noise and did not belong to any group.

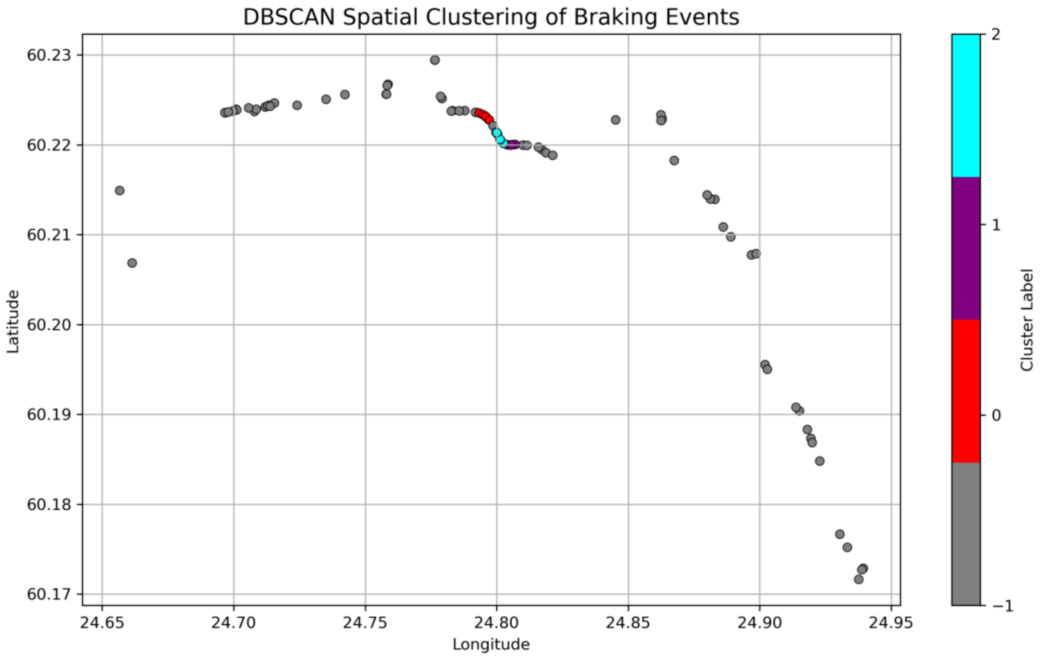


Figure 15. The DBSCAN clustering result shown hotspots of Sudden Braking events. Points with labels 0,1,2 were identified as a Sudden Braking group.

K-Means technique was used to detect braking patterns based on speed and acceleration. Before implementing the algorithm, the features were standardized by the `standardscaler` function from the scikit-learn library to ensure fair clustering. After normalization, visual inspection, and datasets, the number of clusters was chosen. The result of the K-Means algorithm was visualized in the scatter plot, in 4 clusters. Each dot in the scatter plot showed a sudden braking event. The x-axis was speed in m/s, and the y-axis was acceleration.

As shown in Figure 15, the method detected several behavior patterns. Cluster 0, the blue dots, showed moderate speed and deceleration. They were likely standard braking. Cluster 1, with the red dots, was events with sharp deceleration. These events were detected as near-miss scenarios. Cluster 2, the cyan dots represented braking events with higher speeds. Cluster 3, the pink dots, were high-speed braking incidents with moderate deceleration. These incidents were rare, but important to monitor for risk assessment.

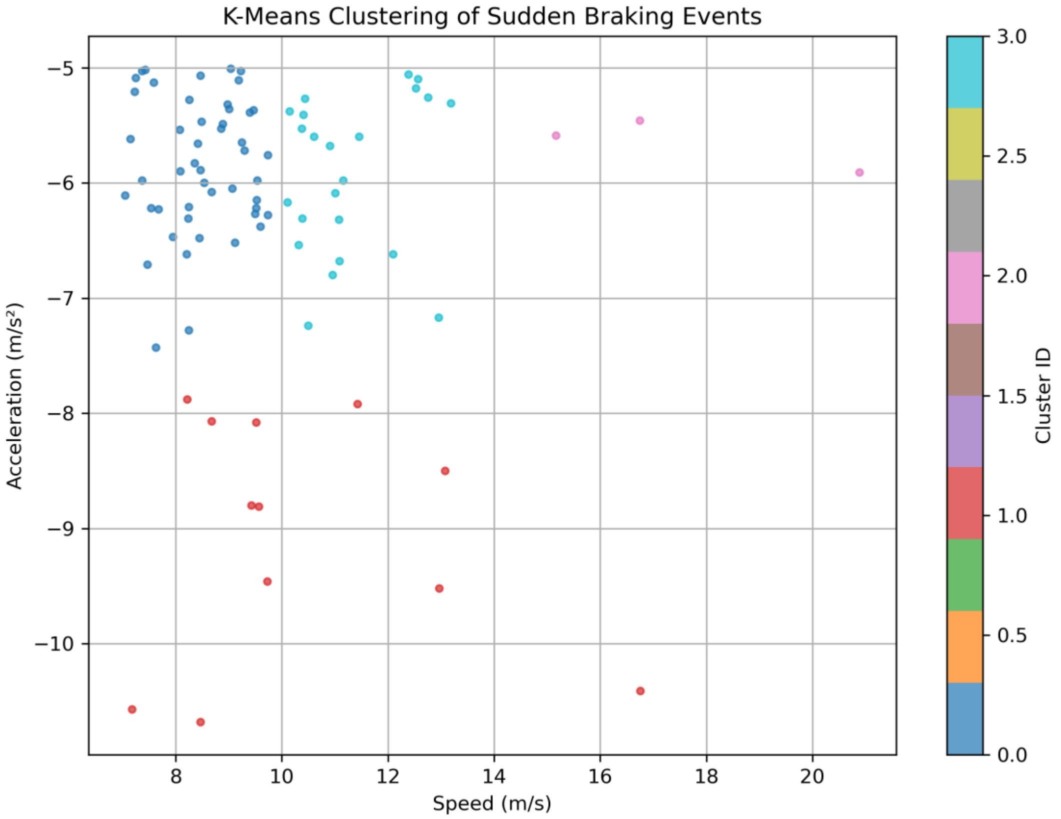


Figure 16. K-Means clustering of Sudden Braking events based on speed and acceleration.

This Chapter brought various forms of analysis to understand better when and where sudden braking events happen across Helsinki’s bus network. The findings showed that some locations repeatedly experienced more abrupt braking. This indicates that some bus route segments could be improved by optimizing traffic flow and increasing the route infrastructure, or adjusting operational practices to reduce risks for vulnerable road users such as pedestrians and cyclists.

6 Conclusion and Discussion

This study used real-time data from HSL to detect and analyse sudden braking events in Helsinki’s public transport system. The research combined data cleaning, Temporal and spatial filtering, and visual analysis to detect patterns of abrupt deceleration, which indicated risk for VRUs.

The findings showed that sudden braking events were not randomly distributed across time or location. Most events occurred during weekday morning and early afternoon, a temporal situation linked to peak travel hours and high pedestrian and cyclist traffic. Commercial centres, public transport hubs, and main intersections showed repeated braking patterns. These patterns were visible through the histograms, pie charts, and interactive maps.

Clustering methods such as DBSCAN and K-Means added a deeper layer to the analysis. DBSCAN highlighted hotspots of braking incidents, while K-Means identified patterns based on speed and deceleration features. K-Means method showed that all sudden braking events do not have the same safety risks. For example, high-speed, high-deceleration events were more critical than low-speed slowdowns. These tools helped to investigate the location of one-time and frequent braking events that may prioritize improving urban traffic safety.

These insights are valuable for stakeholders and urban planners. The spatial and temporal information of repeated sudden braking events can play a significant role in future urban planning. For instance, reviewing street infrastructure, optimizing bus stop placement, or modifying signal timings in areas where repeated braking occurs to detect high-risk areas and protect vulnerable road users.

Based on the findings in this study, there is a strong starting point for expanding the analysis of sudden braking events in the real-time data of the public transport system. Applying the exact approach and algorithms for other long route buses, such as lines 5520 and 4560, can bring more information and detect more high-risk areas. Also, analyzing other lines that have an overlap in some part of the route with line 2200 would allow researchers to investigate whether high-risk areas are unique for line 2200 or safety issues within the public transport network.

Future studies should apply the seasonal variations to compare data across summer and winter. The environmental variable may highlight the high-risk area or present more information on abrupt braking behavior. In addition, event times such as school hours, public holidays, and city festivals could be integrated to analyze the sudden braking event near-miss VRUs. Finally, combining spatial clustering with temporal variables and other factors mentioned above makes it possible to assess the risk. The multidimensional approach can provide a complete visual overview of where and when potential safety hazards increase. This approach would enable urban planners to implement static infrastructure improvement and dynamic operations such as seasonal speed rules. The data-driven research enhances modern city planning to create a safer, more reliable, and intelligent urban mobility systems that protect all road users, particularly near-miss vulnerable road users.

References

1 Helsinki Region Transport (HSL). Open data [Internet]. Helsinki: Helsinki Region Transport; [cited 2025 Apr 26]. Available from: https://www.hsl.fi/en/hsl/open-data

2 Saunders M, Lewis P, Thornhill A. Research methods for business students. Pearson education; 2009

3 Yin RK. Case study research: Design and methods. sage; 2009.

4 OECD. Frascati Manual 2015: Guidelines for collecting and reporting data on research and experimental development. Paris: OECD Publishing; 2015. Available from: https://doi.org/10.1787/9789264239012-en

5 Kananen J, Jyväskylän ammattikorkeakoulu. Design research (applied action research) as thesis research: a practical guide for thesis research. Jyväskylä: JAMK University of Applied Sciences; 2013.

6 Gettman D, Head L. Surrogate safety measures from traffic simulation models. Transportation research record. 2003;1840(1):104-15.

7 Mahmud S, Day C. Exploring Crowdsourced Hard—Acceleration and Braking Event Data for Evaluating Safety Performance of Low-Volume Rural Highways in Iowa.

8 Liu L, Racz D, Vaillancourt K, Michelman J, Barnes M, Mellem S, Eastham P, Green B, Armstrong C, Bal R, O'Banion S. Smartphone-based hard-braking event detection at scale for road safety services. Transportation research part C: emerging technologies. 2023 Jan 1;146:103949.

9 Helsinki Region Transport (HSL). HSL real-time open data [Internet]. Helsinki: HSL; [cited 2025 Apr 26]. Available from: https://api.digitransit.fi/hsl-rt/

10 Forum Virium Helsinki. Forum Virium Helsinki official website [Internet]. Helsinki: Forum Virium Helsinki; [cited 2025 Apr 26]. Available from: https://forumvirium.fi/

11 Li X, Dadashova B, Yu S, Zhang Z. Rethinking highway safety analysis by leveraging crowdsourced waze data. Sustainability. 2020 Dec 4;12(23):10127.

12 Forum Virium Helsinki. HSL real-time transit data index [Internet]. Helsinki: Forum Virium Helsinki; [cited 2025 Apr 26]. Available from: https://bri3.fvh.io/hsl-rt/

13 Helsinki Region Transport (HSL). HSL real-time MQTT broker [Internet]. Helsinki: HSL; [cited 2025 Apr 26]. Available from: https://mqtt.hsl.fi/

14 Digitransit. High-frequency positioning API [Internet]. Helsinki: Digi transit; [cited 2025 Apr 26]. Available from: https://digitransit.fi/en/developers/apis/5-realtime-api/vehicle-positions/high-frequency-positioning/

15 Helsinki Region Transport (HSL). Reittiloki: Real-time route log viewer [Internet]. Helsinki: HSL; [cited 2025 Apr 26]. Available from: https://reittiloki.hsl.fi/?date=2025-04-26&route.routeId=&route.direction=0&route.originStopId=

16 Lago A, Patel S, Singh A. Low-cost real-time aerial object detection and GPS location tracking pipeline. ISPRS Open Journal of Photogrammetry and Remote Sensing. 2024 Aug 1; 13:100069.

17 Jeong H, Park W, Lee J, Park S, Yun I. Influence of public bus driver's driving behaviors on passenger fall incidents: an analysis using digital tachograph data. Journal of Advanced Transportation. 2022;2022(1):2941327.

18 Boylan J, Chen WS, Meyer D. The Influence of Vehicle Characteristics on the Braking Behaviour of Young People as Measured Using Telematics. Journal of Road Safety. 2025 Feb 25;36(1):1-0.

19 Djahel S, Doolan R, Muntean GM, Murphy J. A communications-oriented perspective on traffic management systems for smart cities: Challenges and innovative approaches. IEEE Communications Surveys & Tutorials. 2014 Jul 17;17(1):125-51.

20 SEBATUNZI J. Proactive Detection of Dangerous Traffic Locations. Journal of Traffic and Transportation Management. 2023;4(1):29-33.

21 Katrakazas C, Quddus M, Chen WH, Deka L. Real-time motion planning methods for autonomous on-road driving: State-of-the-art and future research directions. Transportation Research Part C: Emerging Technologies. 2015 Nov 1;60:416-42.

22 Sohail A, Cheema MA, Ali ME, Toosi AN, Rakha HA. Data-driven approaches for road safety: A comprehensive systematic literature review. Safety science. 2023 Feb 1;158:105949.

23 Xu C, Gao J, Zuo F, Ozbay K. Estimating Urban Traffic Safety and Analyzing Spatial Patterns through the Integration of City-Wide Near-Miss Data: A New York City Case Study. Applied Sciences (2076-3417). 2024 Jul 15;14(14).

24 Hunter M, Saldivar-Carranza E, Desai J, Mathew JK, Li H, Bullock DM. A proactive approach to evaluating intersection safety using hard-braking data. Journal of big data analytics in transportation. 2021 Aug;3(2):81-94.

25 Elvik R. Risk of non-collision injuries to public transport passengers: Synthesis of evidence from eleven studies. Journal of Transport & Health. 2019 Jun 1;13:128-36.

26 Silvano AP, Ohlin M. Non-collision incidents on buses due to acceleration and braking manoeuvres leading to falling events among standing passengers. Journal of Transport & Health. 2019 Sep 1;14:100560.

27 Barnes J, Brown L, Morris A, Stuttard N. Bus passenger injury prevention: Learning from onboard incidents. Traffic injury prevention. 2023 Jan 2;24(1):98-102.

28 Mora R, Waintrub N, Figueroa-Martínez C. Bus drivers and their interactions with cyclists: an analysis of minor conflicts. Transportation research interdisciplinary perspectives. 2024 May 1;25:101074.

29 Desai J, Li H, Mathew JK, Cheng YT, Habib A, Bullock DM. Correlating hard-braking activity with crash occurrences on interstate construction projects in Indiana. Journal of big data analytics in transportation. 2021 Apr;3(1):27-41.

30 Desai J, Mathew JK, Li H, Mukai J, Sakhare RS, Bullock DM. Correlating Connected Vehicle Estimated Deceleration Events with Crash Incidents near Interstate Interchanges. Journal of Transportation Technologies. 2023 Jul 30;13(4):674-88.

31 Yarlagadda J, Jain P, Pawar DS. Assessing safety critical driving patterns of heavy passenger vehicle drivers using instrumented vehicle data–An unsupervised approach. Accident Analysis & Prevention. 2021 Dec 1;163:106464.

32 Vajpayee V, Saldivar-Carranza ED, Sakhare RS, Bullock DM. Large Scale Evaluation of Normalized Hard-Braking Events Derived from Connected Vehicle Trajectory Data at Signalized Intersections, Roundabouts, and All-Way Stops. Future Transportation. 2024 Aug 27;4(3):968-84.

33 Lagou G, Mantouka E, Barmpounakis E, Vlahogianni EI. Mapping risky driving behaviour in urban road networks. In: Proceedings of the 10th International Congress on Transportation Research (ICTR); 2021 Sep; Rhodes, Greece. Hellenic Institute of Transportation Engineers and Hellenic Institute of Transport; 2021. p. 1–10.

34 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python – Clustering [Internet]. Scikit-learn Documentation; 2024 [cited 2025 Apr 28]. Available from: https://scikit-learn.org/stable/modules/clustering.html#k-means

35 Holmgren J, Knapen L, Olsson V, Masud AP. On the use of clustering analysis for identification of unsafe places in an urban traffic network. Procedia Computer Science. 2020 Jan 1;170:187-94.

36 Lima V, Byrd V. Hotspot Prediction of Severe Traffic Accidents in the Federal District of Brazil. arXiv preprint arXiv:2312.17383. 2023 Dec 28.

37 Watson RB, Ryan PJ. Geospatial factors applied to road accidents: A review. Journal of Advances in Information Technology. 2024;15(3):451-7.

38 Elalouf A, Birfir S, Rosenbloom T. Developing machine-learning-based models to diminish the severity of injuries sustained by pedestrians in road traffic incidents. Heliyon. 2023 Nov 1;9(11).

39 Komol MM, Hasan MM, Elhenawy M, Yasmin S, Masoud M, Rakotonirainy A. Crash severity analysis of vulnerable road users using machine learning. PLoS one. 2021 Aug 5;16(8):e0255828.

40 Scikit-learn developers. Scikit-learn User Guide [Internet]. Scikit-learn; [cited 2025 Apr 28]. Available from: https://scikit learn.org/stable/user_guide.html

41 Idakwo PO, Adekanmbi O, Soronnadi A, David A. Geo-Parsing and Geo-Visualization of Road Traffic Crash Incident Locations from Print Media for Emergency Response and Planning. Available at SSRN 4866596.

42 GitHub - ForumViriumHelsinki/hsl-rt-analysis: HSL real time vehicle data analysis

43 Transportation Research Board. Transportation Research Thesaurus [Internet]. Washington, DC: TRB; [cited 2025 May 11]. Available from: https://trt.trb.org/

Appendix

Appendix 1: Detailed explanation of each column in HSL real-time data

Column	Description
desi	Vehicle internal line ID
dir	Direction of travel 1 or 2
oper	Operator ID
veh	Vehicle number of unique buses
spd	Speed in km/s
hdg	Heading (direction in degree)
acc	Acceleration in m/s2
dl	Delay in seconds. Maybe missing (Nan) Positive = Late, negative=early
odo	Odometer reading in meters
drst	Door status (True = door open, False = closed)
jrn	Journey ID
line	Line number
loc	Location source. Usually, GPS
stop	Stop ID
route	Route number
coccu	Occupancy estimate (likely 0= empty, higher = more people)
ts	Timestamp (UTC)
oday_start	Start of the operational day (when the service began)
geometry	Coordinates of the vehicle (longitude, latitude) Format point (x,y)

Appendix 2: Detection of Sudden Braking Events Script

```
import geopandas as gpd
# Step 1: Load geospatial bus movement data (after GPS cleaning and filtering)
gdf_all = gpd.read_parquet("cleaned_data.parquet")

# Step 2: Fill missing acceleration values with zero to avoid null issues
gdf_all["acc"] = gdf_all["acc"].fillna(0)

# Step 3: Initial Detection of Sudden Braking Events
# Criteria:
# - Strong deceleration (< -5.0 m/s²)
# - Not at an official stop
# - Reasonable speed (> 1.5 m/s)
# - Valid direction (1 or 2)
# - Within Helsinki-Espoo bounding box
braking_events = gdf_all[
    (gdf_all["acc"] < -5.0) &
    (gdf_all["stop"].isna()) &
    (gdf_all["spd"] > 1.5) &
    (gdf_all["dir"].isin([1, 2])) &
    (gdf_all["geometry"].x.between(24.4, 25.4)) &
    (gdf_all["geometry"].y.between(59.8, 60.5))
].copy()

print(f"Detected initial braking events: {len(braking_events)}")

# Step 4: Save initial braking events for record
braking_events.to_parquet("braking_events.geoparquet")
braking_events.to_csv("braking_events.csv", index=False)

# Step 5: Final Filtering — Remove suspicious or extreme values
braking_cleaned = braking_events[
    (braking_events["acc"] > -15.0) & # Remove physically unrealistic
deceleration
    (braking_events["spd"].between(1.5, 35)) &
    (braking_events["dir"].isin([1, 2])) &
    (braking_events["geometry"].x.between(24.4, 25.4)) &
    (braking_events["geometry"].y.between(59.8, 60.5))
].copy()

print(f"Final braking events after filtering: {len(braking_cleaned)}")

# Step 6: Save cleaned braking event dataset
braking_cleaned.to_parquet("braking_cleaned.geoparquet")
braking_cleaned.to_csv("braking_cleaned.csv", index=False)
```