Cartography M.Sc.

Master Thesis

An Exploratory Study of Solving Traffic Sensor Location Problem from the Perspective of Traffic Safety

Zihan Liu

2023

An Exploratory Study of Solving Traffic Sensor Location Problem from the Perspective of Traffic Safety

submitted for the academic degree of Master of Science (M.Sc.)

conducted at the Department of Aerospace and Geodesy,

Technical University of Munich

Author:	Zihan Liu
Study Course:	Cartography M.Sc.
Supervisors:	Dr.-Ing. Christian Murphy (TUM), Dr. Andreas Eich (LiangDao GmbH)
Reviewer:	Dr. Paulo Raposo (UT)
Cooperation:	LiangDao GmbH
Chair of the Thesis Assessment Board:	Prof. Dr.-Ing. Liqiu Meng
Date of Submission:	07.09.2023

Statement of Authorship

Herewith I declare that I am the sole author of the submitted Master's thesis entitled:

"An Exploratory Study of Solving Traffic Sensor Location Problem from the Perspective of Traffic Safety"

I have fully referenced the ideas and work of others, whether published or unpublished. Literal or analogous citations are clearly marked as such.

Munich, 07.09.2023 Liu, Zihan

Acknowledgements

First and foremost, I would like to express my deepest gratitude to the members of the Thesis Assessment Board. Thanks to Dr.-Ing. Christian Murphy and Dr. Andreas Eich for serving as my supervisors. Thanks to Dr. Paulo Raposo for serving as my reviewer. Thanks to Prof. Dr.-Ing. Liqiu Meng for serving as the board chair. Their precious comments are of great help to this thesis.

This thesis was completed within the framework of the cooperation between the Chair of Cartography and Visual Analysis and LiangDao GmbH. Therefore, I would like to express my sincere thanks to both sides for their firm support in many aspects. Thanks to Dr.-Ing. Yu Feng and M.Sc. Peng Luo for their valuable opinions during the experiment stage. Thanks to all the colleagues of the Chair of Cartography and Visual Analytics and LiangDao GmbH for their help during the thesis process.

This thesis was completed within the Cartography M.Sc. programme. I would like to show my extreme thanks to all the coordinators, professors, lecturers and classmates involved in this programme. It would not become such a success without any one of them. Particularly, thanks to M.Sc. Juliane Cron for her long-term massive commitment to this programme.

Last but not least, I would like to express my deepest thanks to my parents and all my family. Thanks to their unconditional love and unwavering support for me since my birth.

Abstract

With the maturity of supporting software and hardware, sensors with strong perception capabilities are being piloted as traffic sensors. Compared with current mainstream traffic sensors, these advanced traffic sensors can achieve the accurate detection and tracking of pedestrians and vehicles and obtain trajectory-level data. Therefore, they can achieve more than just traffic flow monitoring, such as traffic accident monitoring.

Before traffic sensors can function in the field, their location strategy is a critical decision to be made. The related discussion is generalized as the Traffic Sensor Location Problem (TSLP). The location strategy of traffic sensors is always determined by their functionality. Therefore, over time, TSLP has often been discussed from the perspective of traffic flow, while the discussion from the perspective of traffic safety has been rather rare. This thesis just aims at this gap to conduct an exploratory study on TSLP from the perspective of traffic safety.

This thesis proposes that from the perspective of traffic safety, traffic sensors should be located in traffic accident and near-accident hotspots. Based on this, a methodology for solving TSLP from the perspective of traffic safety is proposed. It consists of three parallel parts: network analysis, risk analysis and rule analysis. Network analysis is to use the network analysis methods to detect the historical traffic accident hotspots on the road network. Risk analysis is to explore the relationship between the accident risk of traffic intersections and their nearby geographic features and therefore predict the accident risk with the geographic features. Based on the prediction results, predicted high-risk traffic intersections are potential traffic accident or near-accident hotspots. Rule analysis is to discover the association rules between accident locations and their nearby geographic features, and accident-associated geographic features are expected to be found. Based on the association rules, places close to accident-associated geographic features are potential traffic accident hotspots. In this way, a methodology consisting of three hierarchical study objects, which are the road network, traffic intersections and accident locations, is proposed.

This thesis chooses the city of Wuppertal as the study area, and two open data sources, namely OpenStreetMap and the German Accident Atlas, are adopted. Data preparation is a key step before performing three analysis methods, data of three different study objects needs to be processed by road network modelling, traffic intersection detection, accident data filtering, and data enrichment. Network analysis is conducted on three levels of the road network: centrality measures on the node level, network kernel density estimation on the lixel level, and community detection on the community level. Risk analysis is conducted by training a random forest classification model with the data of four nearby cities, and predicting the accident risk of traffic intersections in Wuppertal. Rule analysis is conducted by implementing the association rule analysis between traffic accidents and their nearby geographic features, and it's conducted on the overall dataset, data clusters by their attributes and data clusters by community detection.

The results of the case study initially demonstrate the effectiveness of the methodology and three analysis methods proposed in this thesis. Network analysis methods can effectively identify the historical accident hotspots on the road network. The relationship between accident risk and geographic features can be effectively modelled and used for risk prediction, and association rules between accident risk and geographic features can

be effectively extracted. With their results, potential accident and near-accident hotspots can be predicted. What's more, this thesis proposes a general guideline on dealing with the TSLP from the perspective of traffic safety for other's reference.

This thesis is an explorative study. The discussion about solving TSLP from the perspective of traffic safety can be continued by going deeper with network analysis, combining risk analysis and rule analysis, iterative analysis and more. The related research about TSLP should continue so that spatial decision-making can support more for the deployment of advanced traffic sensors and therefore smart transportation.

Keywords: Traffic Sensor Location Problem, Traffic Safety, Traffic Accident Hotspot Detection, Spatial Decision-Making, Network Analysis, Machine Learning and Data Mining

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 General Overview

In order to achieve a balance between effectiveness and cost, the discussion about optimal traffic sensor locations has lasted for decades since traffic sensor networks were deployed. Owais (2022) summarized this type of problem as the Traffic Sensor Location Problem, which is abbreviated as TSLP.

Location strategies are always associated with sensor functionality. Currently, the main function of mainstream traffic sensors is to monitor traffic flow, which is a decisive element in transportation planning and traffic management (Owais, 2022). Therefore, TSLP always looks for optimal solutions from the perspective of traffic flow and answers two typical questions: how many sensors are needed, and what are the best locations for their deployment (Owais, 2022; Liu et al., 2023)?

However, thanks to enhanced computing power and real-time automatic data processing capabilities, emerging LiDAR sensors with high precision, all-day and all-weather working ability, and strong privacy protection ability compared with cameras, are being piloted as traffic sensors (J. Zhao et al., 2019; Z. Zhang et al., 2019; J. Zhang et al., 2020). These advanced sensor technologies, including HD cameras and LiDAR sensors, are helping the perception capabilities of traffic sensors to significantly improve. More advanced and refined functions can be achieved than just traffic flow monitoring. In particular, they have great application prospects in traffic safety, which is difficult to achieve with the current mainstream traffic sensors (Wu et al., 2018). They can detect the occurrence of traffic anomalies such as traffic accidents and near-accidents (that is, near misses), and monitor their full and detailed process. In this way, traffic anomalies can be reported to the police more quickly and early warning to follow-up road users via V2X (Vehicle to Everything). The full process monitoring can help with accident liability presumption and insurance claims, and more importantly, help us better understand the causes of traffic anomalies, which can help reduce traffic accidents and even help future autonomous vehicles better understand human driving behaviour so that they can better make decision while interacting with human drivers (Wu et al., 2018).

Before the new traffic sensors can play a role in improving traffic safety in the transportation network, the first step is location selection, that is, dealing with TSLP. However, the current discussion of TSLP is mainly from the perspective of traffic flow, the discussion from the perspective of traffic safety is rather rare. To the best of the author's knowledge, no relevant literature records were found.

Therefore, this thesis is just aiming at this gap to carry out an exploratory study of solving TSLP from the perspective of traffic safety.

1.2 Research Objectives

This thesis aims to extend the discussion on TSLP from only the perspective of traffic flow to the perspective of traffic safety. More specifically, this thesis wants to explore the data, methods and workflow to detect the optimal traffic sensor locations for monitoring traffic anomalies and improving traffic safety. The case study will be made to verify the

effectiveness of the proposed workflow, methods and data.

Traffic anomalies include traffic accidents and traffic near accidents. The former can be acquired from the existing historical traffic accident dataset, while there are few datasets about near accidents, so the latter can only be predicted. In this way, historical traffic accident hotspots and predicted near accident hotspots can be selected as the traffic sensor locations to monitor traffic anomalies.

Therefore, more practically, the objective of this thesis is to find the methods to detect historical traffic accident hotspots and predict near accident hotspots, design the workflow, and test the workflow with suitable data of a certain study area.

In this respect, this thesis proposes three methods to achieve the expected research objectives:
1) Network Analysis: historical traffic accident hotspot analysis on the road network,
2) Risk Analysis: traffic risk prediction of traffic intersections,
3) Rule Analysis: association rule analysis between geographic features and accident locations.

In this way, these three parts constitute a study that is progressive and oriented to three different research objects, that is, road networks, traffic intersections and accident locations.

1.3 Research Questions

Following the research objectives of this thesis, research questions are proposed below:

RQ 1: What are the status and the latest progress on the Traffic Sensor Location Problem? Has traffic safety-oriented TSLP already been discussed? And what is the difference between it and traffic flow-oriented TSLP?

RQ 2: How to deal with TSLP from the perspective of traffic safety? What methods can be beneficial?

RQ 3: How can the historical traffic accident hotspots on the road network be extracted so that the traffic sensor can be located in those high-risk locations?

RQ 4: Can the relationship between the accident risk of one traffic intersection and its surrounding environment be modelled? And can the model effectively predict the accident risk of traffic intersections in one city and be examined with the true value, so that the traffic sensor can be located in those predicted high-risk intersections, including potential near-accident hotspots?

RQ 5: Can some significant association rules between traffic accidents and geographic features be extracted, so that the traffic sensor can be located near those accident-associated geographic features, which are potential accident or near-accident hotspots?

1.4 Thesis Structure

This thesis consists of seven chapters:

Chapter 1 Introduction: This chapter introduces the topic of this thesis from its background and scientific relevance to the proposed research objectives and questions.

Chapter 2 Literature Review: This chapter reviews the progress on traffic sensors, their application in traffic safety, their location strategy, namely TSLP, and two categories of methods that will be used in this thesis.

Chapter 3 Methodology: This chapter introduces the general methodology proposed in this thesis.

Chapter 4 Case Study: This chapter introduces the detailed examination and application of the methodology proposed in Chapter 3 with the city of Wuppertal as the study area.

Chapter 5 Discussions: This chapter discusses the results of the case study, makes connections between analysis results and location strategy, answers the research questions, and points out the limitations and where to continue.

Chapter 6 Conclusions: This chapter summarises the whole thesis, proposes a general guideline on solving TSLP from the perspective of traffic safety and includes an outlook.

2 Literature Review

This chapter includes five sections. It starts with a review of **2.1 Progress on Traffic Sensors**, and emerging **2.1.1 Traffic Sensors with Strong Perception Abilities** are introduced. Advanced traffic sensors can bring new functionalities, and traffic safety improvement could be one of them. So a review of **2.2 Research and Application of Advanced Traffic Sensors in Traffic Safety** is followed, especially the **2.2.1 Research on Traffic Accident and Near-Accident**. Regardless of their functions, the location strategy of traffic sensors is always an important decision to make. So **2.3 Traffic Sensor Location Problem in Transportation Science** is introduced, the current status of TSLP is summarised as **2.3.1 Traffic Flow-Oriented TSLP**, and **2.3.2 Traffic Safety-Oriented TSLP** is to be discussed in this thesis. **2.4 Network Analysis Methods** and **2.5 Machine Learning and Data Mining Methods** could be beneficial to the solution of traffic safety-oriented TSLP. The exact methods and algorithms, including **2.4.1 Centrality Measures**, **2.4.2 Network KDE**, **2.4.3 Community Detection**, **2.5.1 Random Forest Model** and **2.5.3 Association Rule Learning**, and their selection reasons are introduced.

The structure of this chapter is also illustrated in Figure 2.1.

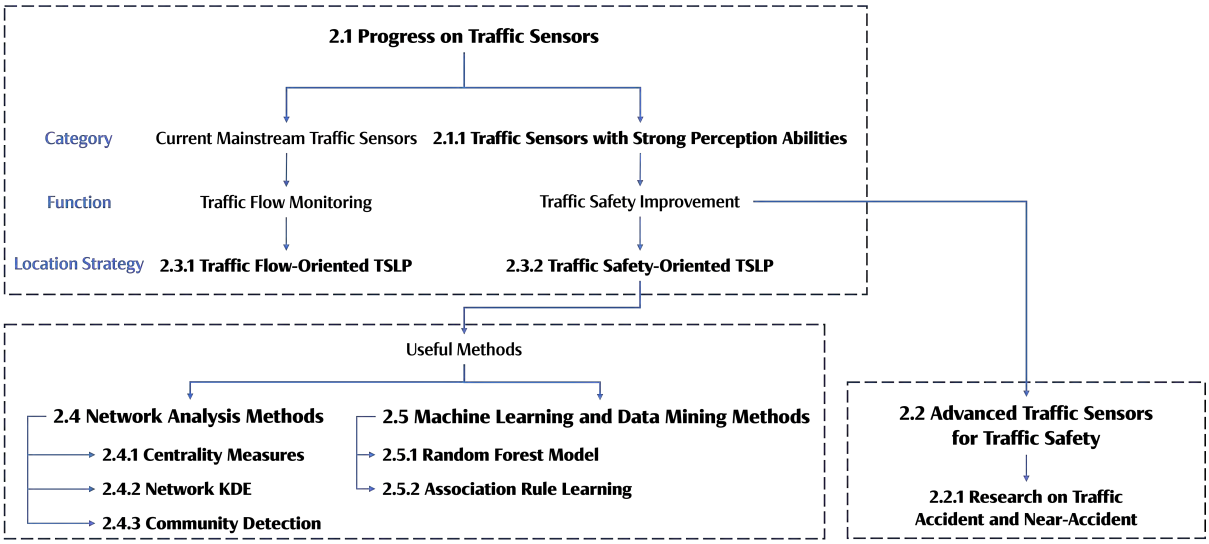

Figure 2.1 Overview of Literature Review

2.1 Progress on Traffic Sensors

Over time, sensor technology has become ubiquitous. In road traffic, sensor technology supports its surveillance, control, and management, and is crucial for Intelligent Transportation Systems (ITS) (Guerrero-Ibáñez et al., 2018; J. Zhao et al., 2019). Based on their installation locations, traffic sensors can be classified into two categories: intrusive (or in-roadway) sensors and non-intrusive (or over-roadway) sensors (Tewolde, 2012; Guerrero-Ibáñez et al., 2018).

Intrusive traffic sensors (Tewolde, 2012; Guerrero-Ibáñez et al., 2018) are installed on pavement surfaces and can be divided into four types by their principles: (1) passive

magnetic sensors, (2) pneumatic road tube sensors, (3) inductive loop detectors (ILDs) and (4) piezoelectric. Thanks to their technology maturity and high accuracy in detecting vehicles, intrusive sensors are used most in traffic control systems. While the drawback is the high costs of installation, maintenance and repair. Table 2.1 shows the applications of these four types of intrusive sensors that are currently in use for road traffic.

Non-intrusive traffic sensors (Tewolde, 2012; Guerrero-Ibáñez et al., 2018) are installed at different places on the roads and can be divided into three groups by their installation locations: (1) roadside mast-mounted, (2) bridge-mounted and (3) across-roadside. By their principles, they can be divided into seven types as shown in Table 2.2. They can detect vehicle's transit and other parameters such as vehicle speed and lane coverage. They provide many of the intrusive traffic sensors' functions with fewer difficulties, but they are expensive and highly affected by environmental conditions.

Table 2.1 Intrusive Sensors Currently Used for Traffic Control (Tewolde, 2012; Guerrero-Ibáñez et al., 2018)

Sensor Type	Application and Use
Pneumatic road tube	Short-term traffic counting, traffic classification by axel count and spacing, for the purpose of planning and research studies
Inductive Loop Detector (ILD)	Vehicle passage, presence, count, occupancy and speed
Magnetic sensors	Identify stopped and moving vehicles
Piezoelectric	Classify vehicles by axel count and axel spacing and to measure vehicle weight and speed

Table 2.2 Non-Intrusive Sensors Currently Used for Traffic Control (Tewolde, 2012; Guerrero-Ibáñez et al., 2018)

Sensor Type	Application and Use
Video cameras	Detection of vehicles across several lanes and can classify vehicles by their length, and report vehicle presence, flow rate, occupancy, and speed for each class
Radar sensors	Object detection and measurements of distance and speed, calculate the vehicle data such as volume, speed, occupancy, length, etc.
Infrared	Provide data on vehicle presence at traffic signals, volume, speed, length, and queue measurement
Ultrasonic	Provides vehicle count, presence, occupancy information and vehicle speed.
Acoustic array sensors	Measure vehicle passage, presence, and speed
Road surface condition sensors	Collect information on weather conditions, such as surface temperature, dew point, water film height, road conditions and grip
RFID (Radio-frequency identification)	Track vehicles mainly for toll management

2.1.1 Traffic Sensors with Strong Perception Abilities

Although each sensor technology has its inherent strengths and weaknesses, a common problem of current mainstream sensor technologies lies in their inability to get trajectory-level data and low performance in the accurate detection and tracking of pedestrians and vehicles (J. Zhao et al., 2019). In some papers, the data in need is summarised as High-Resolution Micro Traffic Data (HRMTD), which includes speed, location, direction and timestamp (Lv et al., 2019). Besides, there should be no doubt that autonomous vehicles

will take over the driving task in the future with a compact system of accurate onboard sensors, sophisticated algorithms, and powerful computing. It raises the question of how future infrastructure-based traffic sensor systems should be developed in alignment with autonomous driving technology to make the roads and all road users a seamless and cooperative system (J. Zhao et al., 2019; Z. Zhang et al., 2019; J. Zhang et al., 2020).

In the short run, getting the HRMTD of all road users will become a big leap in traffic monitoring, it will greatly help to improve traffic safety, traffic operation and control, traffic management, and performance measurement (J. Zhao et al., 2019). In the long run, a real-time vehicle-road cooperative system will be needed and built (J. Zhao et al., 2019). However, the initial step is always to apply more advanced and reliable traffic-sensing technologies.

In the past decade, researchers have used LiDAR and HD cameras for this purpose (J. Zhao et al., 2019; Z. Zhang et al., 2019; J. Zhang et al., 2020). LiDAR sensors and HD cameras are different in data quality, data coverage and expected performance aspects: (1) Data Quality: Data from the former is point cloud data, which has high accuracy but relatively lower density, while data from the latter are high-resolution images. (2) Data Coverage: Compared with the latter, the former can cover a wider detection range at all angles simultaneously, while its unit price is more expensive. (3) Expected Performance: The former can work 24/7, and its data analysis requires much less processing and computing power (J. Zhao et al., 2019; Z. Zhang et al., 2019; J. Zhang et al., 2020). What's more, LiDAR can better protect road users' privacy such as their biometric information, which makes LiDAR sensors more popular than HD cameras in those countries that focus on privacy protection like Germany.

2.2 Research and Application of Advanced Traffic Sensors in Traffic Safety

According to the Global Status Report on Road Safety 2018 of the World Health Organization (WHO) (WHO, 2018), every year the lives of approximately 1.3 million people are cut short as a result of road traffic accidents, and between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury. Road traffic injuries cause considerable economic losses to individuals, their families, and nations as a whole. It is estimated that road traffic accidents cost most countries 3% of their gross domestic product (WHO, 2018). Therefore, the United Nations General Assembly has set an ambitious target of halving the global number of deaths and injuries from road traffic accidents by 2030 (Resolution A/RES/74/299 Adopted by the General Assembly).

In the context of such urgent social needs, relevant research on traffic safety is necessary. Using traffic sensors with strong perception abilities could help to monitor the full process of traffic accidents and acquire detailed data on them, which can be very helpful for research on traffic safety and traffic accidents.

Several studies have been carried out to explore the application prospects of traffic sensors in traffic safety. Goldhammer et al. (2012) presented a novel multi-sensor network to perceive the intersection environment for traffic safety applications. This network consists of laser scanners, cameras, signal phase tapping and an I2V (Infrastructure to

Vehicle) communication unit. H. Zhao et al. (2012) presented a network of horizontal laser scanners to monitor an intersection.

2.2.1 Research on Traffic Accident and Near-Accident

A near-accident or near-miss is an event that has the potential to develop into a collision situation between two vehicles or between a vehicle and a pedestrian or bicyclist (Huang et al., 2020). These events are important to monitor and analyze for preventing collisions in the future (Huang et al., 2020).

The research on using LiDAR and HD cameras to identify traffic near-accident is still very novel. Wu et al. (2018) present a method for traffic near-accident identification based on the trajectories of road users extracted from roadside LiDAR data. Three parameters, including time difference to the point of intersection (TDPI), distance between stop position and pedestrian (DSPP) and vehicle-pedestrian speed-distance profile, were developed for vehicle-pedestrian near-crash identification. Huang et al. (2020) propose an integrated two-stream convolutional network architecture that performs real-time traffic near-accident detection of road users in traffic video data and builds a traffic near-accident dataset.

However, the current research is rather on a small scale like an intersection and hasn't discussed the location strategy of traffic sensors on a larger scale like a city.

2.3 Traffic Sensor Location Problem in Transportation Science

For effectiveness and cost purposes, location always needs to be considered when a traffic sensor is to be deployed in the traffic system. The discussion about the location strategy of traffic sensors has been a long-standing topic in transportation science for over 30 years Owais, 2022. Owais (2022) summarized this topic as the Traffic Sensor Location Problem, abbreviated as TSLP. It aims to answer two typical questions: how many sensors are needed, and where are the best locations for their deployment?

2.3.1 Current Status of TSLP: Traffic Flow-Oriented TSLP

The location strategy of traffic sensors is always closely related to their functions. Therefore, the current status of TSLP is determined by the current functions of traffic sensors, or more precisely, by the data these sensors acquire.

Traffic flow data is central to transportation planning, operations and management, it's also crucial for Intelligent Transportation Systems (ITS), where complete traffic flow data is required (Owais, 2022). Over time, traffic sensors have been recognized as sources of such data (Owais, 2022). Therefore, currently, the discussion on TSLP is focused on the acquisition of expected traffic flow data. It's called Traffic Flow-Oriented TSLP in this thesis.

Owais (2022) classified the past TSLP studies into six categories:

(1) O/D estimation/updating (ODE): An O/D trip matrix is formed by a two-dimensional matrix, where the value of each cell denotes the number of trips between the row number

(origin zone) and column number (destination zone). O/D matrix estimating/updating is one of the earliest applications of traffic flow data. It usually seeks the most accurate O/D matrix estimation using statistical methods.

(2) Flow observability (OBS): The flow observability problem addresses all network flow elements (i.e., O/D, link, and path flow), aiming at calculating all flow values by directly measuring a subset.

(3) Link flow inference (LFI): The link flow inference problem generally searches for the minimum locations of sensors to infer all other link flows through the conservation equation of the basic nodes (i.e., inflow = outflow).

(4) Path reconstruction (PRC): The path reconstruction approach investigates how the captured flow information can be used to draw a complete picture of the paths used among O/D pairs.

(5) Screen line/traffic surveillance (SLP): It simply seeks the lowest number of sensors in a network in which no vehicle can move from the origin (O) to a destination (D) without passing at least one sensor. This process ensures complete traffic coverage of a network.

(6) Travel time estimation (TME): AVI (Automatic Vehicle Identification) sensors or other similar technologies are essential because they can precisely record the vehicle-passing time of two consecutive AVI sensors. Consequently, they can be used to estimate the travel time for all traffic with careful network placement.

2.3.2 Traffic Safety-Oriented TSLP

As introduced in 2.1, with the development of vehicle-road cooperative systems and ITSs, sensors with strong perception abilities, like LiDAR sensors and HD cameras, are piloted as traffic sensors. As introduced in 2.2, these traffic sensors can achieve more than just acquiring the traffic flow data, they can monitor the micro traffic behaviours in detail, like near accidents, thanks to their perception ability. Therefore, these sensors can play a big role in monitoring traffic accidents and near accidents so that traffic safety can be improved by monitoring, early warning and response to these traffic anomaly events.

As these sensors with strong perception abilities are just piloted as traffic sensors, and it's still at a very early stage, their location strategy hasn't been well discussed yet. This thesis aims for this gap to explore the possible solution to traffic safety-oriented TSLP.

2.3.3 Relevance of TSLP to Site Selection Problem

In GIScience, there is a relevant hot topic called the site selection problem (Rikalovic et al., 2014). But how is the relevance of TSLP to the site selection problem and are there some methods that can be borrowed from it to deal with TSLP?

As shown in the table below, according to actual conditions, the site selection problem can be classified into three classes by their different scales: (1) macro scale, (2) meso scale, (3) micro scale.

The site selection problem in GIScience mainly focuses on the macro scale, where Multi-Criteria Decision Analysis (MCDA) (Rikalovic et al., 2014) and viewshed analysis methods are often used. TSLP focuses on the meso scale, which means the site selection of

small features in a large space. If the traffic sensor locations in a specific intersection need to be determined, then it should be classified as the micro scale, which is more dependent on human experience.

Table 2.3 Comparison Table of Site Selection at Different Scales

Scale	Description	Example
Macro Scale	large features in a large space	power station site selection in a city (often select one from several candidates)
Meso Scale	small features in a large space	traffic sensor location selection in a city
Micro Scale	small features in a small space	traffic sensor location selection in an intersection

2.4 Network Analysis Methods

Network theory (National Research Council, 2005) is a part of graph theory. It defines networks as graphs where the nodes or edges possess attributes. Network analysis (National Research Council, 2005)is a method of studying the relationships between nodes in a network. It involves analyzing the edges between nodes, as well as the characteristics of the nodes themselves. Network analysis can be used to study a wide range of systems, including social networks, transportation networks, biological networks, etc.

There are many types of network analysis (National Research Council, 2005), including centrality analysis, community detection, topological analysis, influence analysis, flow analysis and more.

In the following, several network analysis methods used in this thesis will be introduced.

2.4.1 Centrality Measures

Centrality measures (Bonacich, 1987; Borgatti, 2005) are answers to the question "What characterizes an important node?" In network analysis, centrality measures assign numbers or rankings to nodes within a network corresponding to their network position, which identifies the most important nodes. The word "important" has a wide number of meanings, leading to many different definitions of centrality. The common centrality measures include betweenness centrality, degree centrality, closeness centrality and more. Suitable centrality measures can help to identify the key nodes in the network, such as the traffic accident hotspots on the road network.

Applications of centrality measures (Bonacich, 1987; Borgatti, 2005) include identifying the most influential person(s) in a social network, key infrastructure nodes in the Internet or urban networks, super-spreaders of disease, and brain networks.

2.4.2 Network Kernel Density Estimation

A classical Kernel Density Estimate (KDE) estimates the continuous density of a set of events in a two-dimensional space (Y.-C. Chen, 2017; Gelb, 2021). The density is estimated at sampling points, traditionally the centres of pixels dividing the study area into equal zones. While classical KDE is not adapted to analyze the density of events occurring on a network, like accidents and crimes in streets, or leaks on a network of water pipes. Indeed, calculating density values for locations outside the network is meaningless

and the Euclidean distance underestimates the real distance between two objects on the network. Moreover, networks are not isotropic spaces (uniformity in all orientations). In other words, it is not possible to move in every direction, but only along the edges of the network.

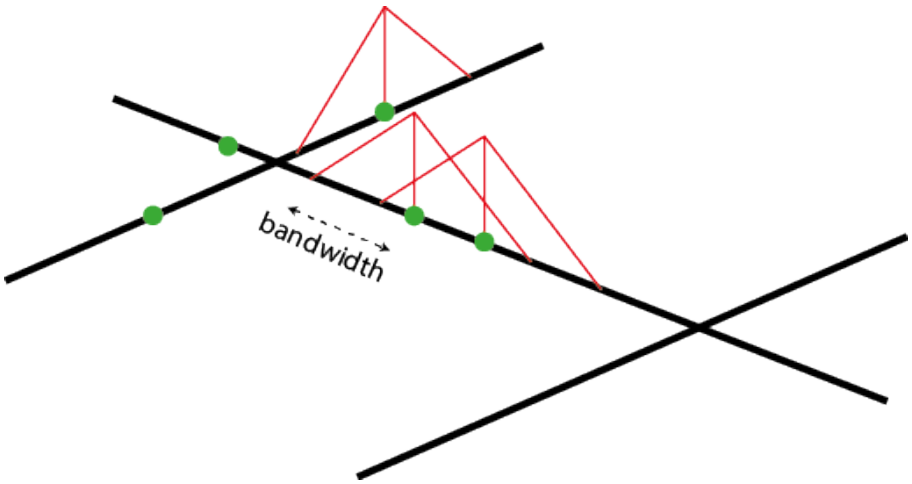

Figure 2.2 Schematic Diagram of Kernel Function (Gelb, 2021)

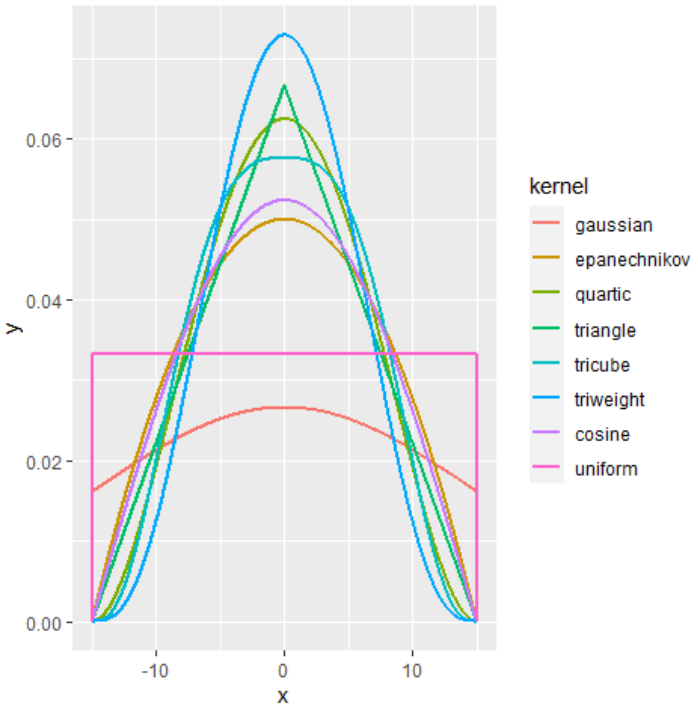

Figure 2.3 Different Kernel Functions (Gelb, 2021)

To calculate a Network Kernel Density Estimate (NKDE) (Gelb, 2021), it is possible to use lixels instead of pixels. A lixel is a linear equivalent of a pixel on a network. The lines of the network are split into lixels according to a chosen resolution. The centres of the lixels are sampling points for which the density will be estimated. The network distances between objects instead of Euclidean distances are calculated. The kernel function is adjusted to deal with the anisotropic space.

Figure 2.2 shows a schematic diagram of how the kernel function works in a network space. Black lines represent the network space, and each green point represents one event that occurred along the network space. Red lines represent the kernel functions. The kernel function is used to assign the "influence" of each point to the network space (Gelb, 2021). This "influence" is limited within the bandwidth (as shown in Figure 2.2) and it decreases when getting away from the event. Therefore, the kernel density of each sampling point is the accumulated "influence" of nearby events. The mathematical expression of the kernel function is as follows:

$$K(x) >= 0,\ if\ x < bandwidth$$

$$K(x) = 0,\ if\ x >= bandwidth$$

$$\int K(x) = 1$$

Figure 2.3 shows multiple different kernel functions. Most of the kernels look alike.

2.4.3 Community Detection

In the study of complex networks, a network is said to have community structures if the nodes of the network can be easily grouped into (potentially overlapping) sets of nodes such that each set of nodes is densely connected internally (Fortunato, 2010).

Community detection refers to the procedure of identifying community structures in networks according to their structural properties. The common community detection methods include (Fortunato, 2010):

(1) Minimum-cut method: This is one of the oldest algorithms used for dividing networks into parts. The network is divided into a predetermined number of parts, chosen such that the number of edges between groups is minimized. However, this method will find communities regardless of whether they are implicit in the structure, and it will find only a fixed number of them.

(2) Hierarchical clustering: A similarity measure quantifying some (usually topological) type of similarity between node pairs is defined, and similar nodes are grouped into communities according to this measure. This method will be used in this thesis.

(3) Girvan-Newman algorithm: It identifies edges in a network that lie between communities and then removes them, leaving behind just the communities themselves. The identification is performed by employing the betweenness centrality, which assigns a number to each edge which is large if the edge lies "between" many pairs of nodes (Girvan and Newman, 2002).

(4) Modularity maximization: This is one of the most widely used methods for community detection. Modularity is a benefit function that measures the quality of a particular division of a network into communities. The modularity maximization method detects communities by searching over possible divisions of a network for one or more that have particularly high modularity. A popular approach to modularity maximization is the Louvain method, which iteratively optimizes local communities until global modularity can no longer be improved given perturbations to the current community state (Blondel et al., 2008). Louvain method will be used in this thesis.

(5) Statistical inference: This method attempts to fit a generative model to the network data, which encodes the community structure.

(6) Clique-based method: Cliques are subgraphs in which every node is connected to every other node in the clique. As nodes can not be more tightly connected than this, many approaches to community detection are based on the detection of cliques in a graph and the analysis of how these overlap.

2.5 Machine Learning and Data Mining Methods

Machine learning and data mining often employ the same methods and overlap significantly, but while machine learning focuses on prediction, based on known properties learned from the training data, data mining focuses on the discovery of (previously) unknown properties in the data.

Machine learning approaches are traditionally divided into three broad categories (El Naqa and Murphy, 2015; Jordan and Mitchell, 2015), which correspond to learning paradigms, depending on the nature of the "signal" or "feedback" available to the learning system:

(1) Supervised learning (El Naqa and Murphy, 2015; Jordan and Mitchell, 2015): The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.

(2) Unsupervised learning (El Naqa and Murphy, 2015; Jordan and Mitchell, 2015): No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

(3) Reinforcement learning (El Naqa and Murphy, 2015; Jordan and Mitchell, 2015): A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle or playing a game against an opponent). As it navigates its problem space, the program is provided with feedback that's analogous to rewards, which it tries to maximize. Although each algorithm has advantages and limitations, no single algorithm works for all problems.

Performing machine learning can involve creating a model, which is trained on some training data and then can process additional data to make predictions. Various types of models have been used and researched for machine learning systems. Common machine learning models include (El Naqa and Murphy, 2015; Jordan and Mitchell, 2015) (1) artificial neural networks, (2) decision trees, (3) support vector machines, (4) regression analysis, (5) Bayesian networks, (6) Gaussian processes, (7) genetic algorithms (GA), (8) belief functions.

2.5.1 Random Forest Model

In machine learning, ensemble learning uses multiple learning algorithms to achieve better predictive performance than any of the constituent learning algorithms alone (Opitz and Maclin, 1999).

Random forest or random decision forest is an ensemble learning method for classification,

regression and other tasks that operates by constructing a multitude of decision trees at training time (Opitz and Maclin, 1999). For a classification task, the output of a random forest is the class selected by most trees. For a regression task, the mean or average prediction of the individual trees is returned (Ho, 1995). Random forests correct for decision trees' habit of over-fitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient-boosted trees (Hastie et al., 2001).

In addition, decision tree learning is a supervised learning approach commonly used in machine learning and data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. It includes two main types: classification tree and regression tree. The former is when the predicted outcome is the discrete class to which the data belongs, while the latter is when the predicted outcome can be considered as a real number, that is, a continuous number (Maimon and Rokach, 2014).

In this thesis, the random forest model for classification tasks will be used to predict the traffic accident risk of traffic intersections.

2.5.2 Model Boosting

Model boosting is an ensemble meta-algorithm for primarily reducing bias and variance in supervised learning, and a family of machine learning algorithms that convert weak learners to strong ones (Zhou, 2012). Specifically, a single machine learning model might make prediction errors depending on the accuracy of the training dataset, while model boosting tries to overcome this issue by training multiple models sequentially to improve the accuracy of the overall system.

XGboost (eXtreme Gradient Boosting) (T. Chen and Guestrin, 2016) is a popular example of model boosting. It is an open-source software library that implements optimized distributed gradient boosting machine learning algorithms under the Gradient Boosting framework. XGboost can be used to improve the prediction performance of random forest model.

2.5.3 Association Rule Learning

Association rule learning (Agrawal et al., 1993) is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.

Agrawal et al. (1993) originally defined the problem of association rule learning as:

Let $I = i_1, i_2, ..., i_n$ be a set of n binary attributes called items. Let $D = t_1, t_2, ..., t_n$ be s set of transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of the items in I. Then a rule is defined as an implication of the form:

$$X \Rightarrow Y, \ where \ X, Y \subseteq I$$

Every rule is composed by two different sets of items, also known as itemsets, X and Y, where X is called antecedent or left-hand-side (LHS) and Y is called consequent or

right-hand-side (RHS). The antecedent is that itemset that can be found in the data while the consequent is the itemset found when combined with the antecedent. The statement $X \Rightarrow Y$ is often read as if X then Y, where the antecedent X is the if and the consequent Y is the then. This simply implies that, in theory, whenever X occurs in a dataset, then Y will occur as well.

Two common indicators of association rule analysis are support and confidence. Support is an indicator of how frequently the itemset appears in the dataset, while confidence is the percentage of all transactions satisfying X that also satisfy Y. Their mathematical formulas are as follows:

$$support = P(X \bigcup Y) = \frac{(number\ of\ transactions\ containing\ X\ and\ Y)}{(total\ number\ of\ transactions)}$$

$$confidence(X \Rightarrow Y) = P(Y|X) = \frac{support(X \bigcup Y)}{support(X)}$$

2.5.3.1 Apriori Algorithm

A well-known algorithm of association rule learning is the Apriori algorithm (Agrawal et al., 1994). It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often. The name of the algorithm is Apriori because it uses prior knowledge of frequent itemset properties. The algorithm process is described below:

The frequent item set is generated by counting the number of occurrences of each item and calculating the support values. Then the item set is pruned by picking a minimum support threshold. This process will be repeated by counting pairs, triplets, and quadruplets... of items in the item set I.

$$support = P(A) = \frac{(number\ of\ transactions\ containing\ A)}{(total\ number\ of\ transactions)}$$

In this thesis, association rule learning has two applications. the aim is to detect the association rules between geographic features and accident locations, In this thesis, the aim is to detect the frequent geographic features that are near traffic accident locations. Therefore, the Apriori algorithm will be used to detect the frequent item set of the nearby geographic features of traffic accident locations. The support indicator has a practical implication, namely, the percentage of each geographic feature near the accident locations, which are recorded in the case city.

3 Methodology

3.1 Overview

This thesis aims to explore the solution of TSLP from the perspective of traffic safety. As introduced previously, the location strategy of traffic sensors is determined by their functions. To achieve the function of monitoring traffic anomalies to improve traffic safety, the location of traffic sensors needs to be determined accordingly. The proposed solution in this thesis is to put sensors where traffic accidents or near accidents tend to happen. Therefore, traffic safety-oriented TSLP has been transformed into the detection of traffic anomaly hotspots, including historical accident hotspots and potential near accident hotspots.

In order to achieve the above purpose, the methodology as shown in Figure 3.1 is designed. From left to right, from top to bottom, this methodology is composed of five parts: (1) Part 0 - Data Preparation, (2) Part 1 - Network Analysis, (3) Part 2 - Risk Analysis, (4) Part 3 - Rule Analysis, (5) Part 4 - Traffic Sensor Location Problem.

Part 0 is to process and prepare the data from data sources so that they are suitable for further analysis. Part 1, Part 2 and Part 3 are three parallel parts using different analysis methods (network analysis, risk analysis and rule analysis respectively), focusing on different study objects (road network, traffic intersection, and accident location respectively), but having the same purpose, namely traffic anomaly hotspot detection. Part 4 is to detect the historical traffic accident hotspots or potential traffic near-accident hotspots from the analysis results. The detailed steps within each part will be explained in 3.2, 3.3, 3.4 and 3.5 respectively.

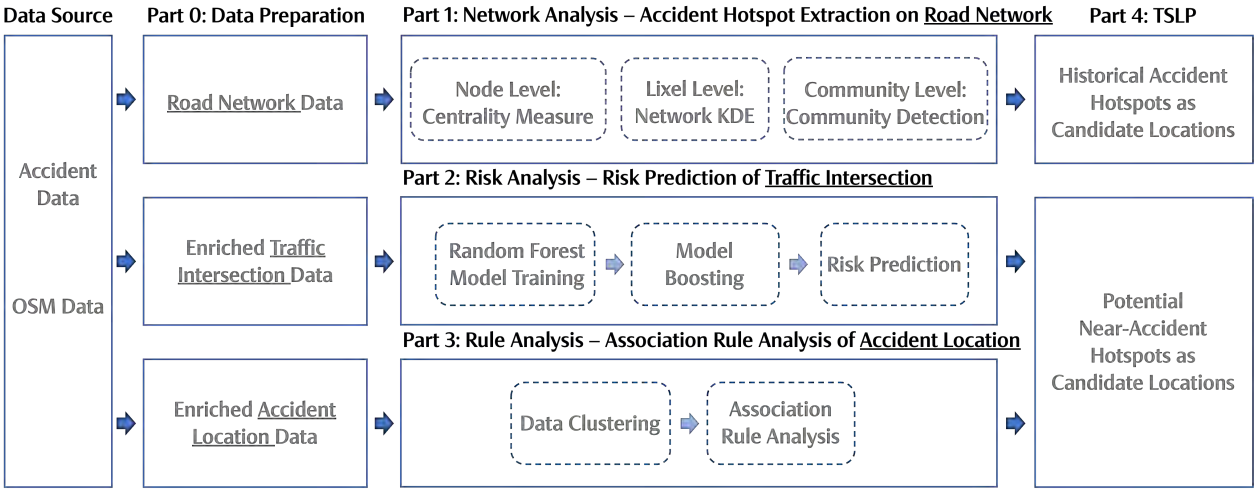

Figure 3.1 Overview of the Methodology

3.2 Part 0: Data Preparation

The first part is Part 0 - Data Preparation. The data to be used is cleaned, processed and enriched so that it can be used for further analysis.

In this thesis, the process of enriching the data of study objects (respectively, road networks, traffic intersections, and accident locations) by querying their nearby geographic features (including roads, amenities, facilities and land use types) from OSM is called data enrichment. In this way, the data of study objects is enriched and beneficial for further analysis.

For the following parts, the data of each different study object needs to be prepared.

For Part 1 - Network Analysis, road network data needs to be acquired and modelled as a directed graph, and the historical accident data (point data representing the accident locations) should be processed to be added to the road network data. The method proposed in this thesis is to calculate the accident linear density of each road segment (edge on the graph) and add the linear density as an attribute of the corresponding road segment. Lixels of the road network need to be created as introduced in 2.4.2 for network kernel density estimation. In this way, a road network with accident information is created, and the historical accident hotspots can be identified from the road network using the analysis methods proposed in this thesis.

For Part 2 - Risk Analysis, the study objects are traffic intersections. Therefore, the data of traffic intersections (point data representing the location of traffic intersections) need to be acquired, and they need to be enriched with nearby geographic feature data (queried from OSM). In this way, the model representing the relationship between the accident risk and the nearby geographic features of the traffic intersections can be trained and used for the following prediction.

For Part 3 - Rule Analysis, the study objects are accident locations (point data). These data need to be enriched with nearby geographic feature data (queried from OSM). In this way, the association rules between accident locations and nearby geographic features can be explored.

3.3 Part 1: Network Analysis

Part 1 - Network Analysis aims to extract the historical accident hotspots on the road networks by multiple levels: node level, lixel level, and community level. Figure 3.2 shows the methodology of Part 1 - Network Analysis.

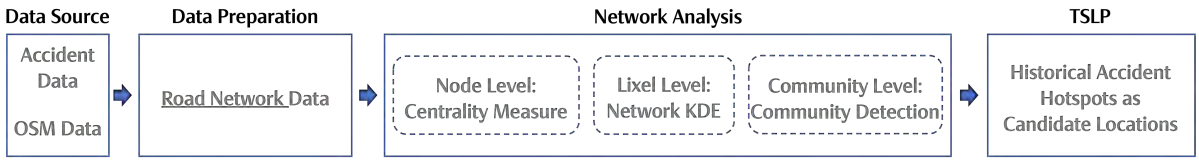

Figure 3.2 Methodology of Part 1 - Network Analysis

With the prepared road network data, network analysis on the different levels of the road network can be made in a parallel way as follows. The extracted historical accident hotspots can be regarded as candidate locations for traffic sensors.

3.3.1 Node Level: Centrality Measure

In a practical sense, nodes in the road network represent the intersections of roads, which are just the places where traffic sensors tend to be placed. So the work of detecting the potential installation places can be replaced with detecting the suitable nodes in the road network. In this case, the suitable nodes are those nodes with high traffic accident risk.

As introduced in 2.4.1, in network analysis, centrality measures assign numbers or rankings to nodes within a network corresponding to their network position. If the assigned ranking can somehow reflect the traffic accident risk of each node, the potential installation places of traffic sensors can be correspondingly got. Then the work is to find or design a suitable centrality measure.

Among all the centrality measures, betweenness centrality and PageRank centrality could be the candidates.

Betweenness centrality of a node v is the sum of the fraction of all-pairs shortest paths that pass through v, and its mathematical formula is as below:

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

where V is the set of nodes, $\sigma(s,t)$ is the number of shortest (s,t)-paths, and $\sigma(s,t|v)$ is the number of those paths passing through some node v other than s, t. If $s = t$, $\sigma(s,t) = 1$, and if $v \in s,t$, $\sigma(s,t|v) = 0$ (Brandes, 2001).

By assigning the weight of each edge (namely each road segment) of the road network with its inversed traffic accident linear density, an updated road network is obtained. The assigned weight is regarded as the passing cost of each edge. Thus, those high accident-risk nodes will be connected to edges with lower passing costs, and they will lie on more shortest paths. Therefore, these nodes will get higher betweenness centrality and can be filtered out.

PageRank centrality computes a ranking of the nodes in the graph based on the structure of the incoming edges. Both the number and the weight of incoming edges are considered. It was originally designed as an algorithm to rank web pages.

By assigning the weight of each edge of the road network with its traffic accident linear density, another updated road network is obtained. Thus, those nodes that are connected to more edges with higher traffic accident linear density will have higher PageRank centrality and can be filtered out.

It's worth mentioning that in both cases, the topology of the road network has an important impact on the centrality results, besides the traffic accident distribution. The locations of nodes on the road network (central or marginal) greatly influence its betweenness centrality, those marginal nodes can hardly get high betweenness centrality even if there are a lot of historical traffic accidents. This will obscure some of the potential hot nodes. For PageRank centrality, the impact is more obvious, nodes connecting more edges get higher PageRank centrality, but it can be easily dealt with by considering both the number of connecting edges and their accident linear density.

The solution to eliminating the excessive impact of road topology is to design a more reasonable centrality measure to be better suited for this application case. The relevant discussion will continue in 5.5 (1).

3.3.2 Lixel Level: Network Kernel Density Estimation

As introduced in 2.4.2, Network Kernel Density Estimation (Network KDE) is the updated version of classical Kernel Density Estimation (KDE), and it deals with the density of events occurring on a network space (Gelb, 2021).

Obviously, traffic accidents always happen along the road network, which is just a network space. Network KDE is the ideal way of assessing the accident density along the road network and detecting historical accident hotspots. Therefore, compared to other lixels, the lixels with higher kernel density are candidate locations for traffic sensors.

Several key parameters need to be determined while applying Network KDE: bandwidth, lixel, and kernel function. For the bandwidth, the experimental value of 300 meters given by Gelb (2021) for accident data can be used. Gelb (2021) also gives two methods of obtaining the optimal bandwidths. And even an adaptive bandwidth can be used. For the lixel, a value equal to the perception range of traffic sensors (about tens of meters) can be adopted. For the kernel function, as shown in Figure 2.3, most of the kernel functions look alike, and different kernel function does not make a significant difference to the results. The quartic kernel function is recommended to be used (Gelb, 2021).

3.3.3 Community Level: Community Detection

As introduced in 2.4.3, Community Detection is to identify groups of nodes that exhibit strong interconnectivity or share similar characteristics. It detects the communities of a network with its topology, node attributes and edge attributes including their directions and weights.

By assigning the weight of each edge of the road network, that is, each road segment, with its traffic accident linear density, an updated road network is obtained. In this way, on this updated road network, the edge with higher accident linear density will get a higher weight, and a group of nodes connected to each other by edges with high weights can be detected as a community. The detected community consists of a group of edges with high accident linear density, and they can be recognized as high accident-risk communities. Therefore, the detected high-accident-risk communities are candidate locations for traffic sensors.

How to choose a suitable community detection algorithm is the key to successfully detecting high accident-risk communities. In this case, the Louvain Algorithm can be used to achieve the task.

3.4 Part 2: Risk Analysis

Part 2 - Risk Analysis aims to predict the accident risk of traffic intersections based on historical accident data. Figure 3.3 shows the methodology of Part 2 - Risk Analysis.

3.4.1 Random Forest Model Training and Model Boosting

After data preparation, the enriched traffic intersection data can be obtained. It includes the historical accident number in each traffic intersection and the nearby geographic

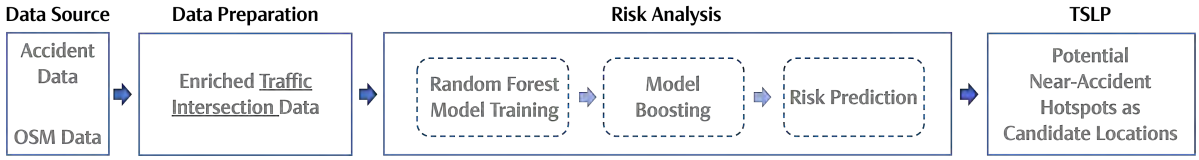

Figure 3.3 Methodology of Part 2 - Risk Analysis

feature data (in the form of the number of each geographic feature). Therefore, the model representing the potential relationship between the historical accident number and nearby geographic features can be trained. Then it's a regression task.

Considering that the relationship between them should be more like an association relationship rather than a causation relationship. So compared with the regression task, the classification task will be more suitable to discover the potential relationship. The historical accident number can be transferred into two classes: high accident risk (accident number is larger than a threshold) or low accident risk (accident number is smaller than a threshold) according to their annual accident number. In this way, the model representing the potential relationship between the historical accident risk and nearby geographic features can be trained.

In this case, the Random Forest Classifier is the ideal model to be trained. The task for the model would be assigning the accident risk (high-risk or low-risk) to each traffic intersection according to its nearby geographic features. So nearby geographic features will work as features, and accident risk will work as labels for the model training and testing.

But before the model training, a key task is to prepare the training set and testing set. It's usually done by randomly splitting the dataset into two parts. In this case, using the data of one place to predict the accident risk in another place could be a better choice. But of course, these two places should share geographical proximity and similar environments. In this way, the accident risk in one place can be obtained completely rather than partially, and relatively larger training and testing sets can be obtained to get better model performance.

Furthermore, as introduced in 2.5.2, the model can be boosted with XGBoost methods to achieve a better model performance.

3.4.2 Risk Prediction

Once the model is built and boosted, the accident risk of certain traffic intersections could be predicted with the nearby geographic feature data. Therefore, the predicted high-accident-risk intersections are just the candidate locations for traffic sensors, including those actual low-accident-risk intersections, which could be potential near-accident hotspots.

The prediction results of the model can be presented in the form of a confusion matrix. And the prediction scores of the model can be gotten in the form of precision and recall. The precision score wants to answer "What proportion of positive identifications was actually correct?" and the recall score wants to answer "What proportion of actual positives was identified correctly?"

Moreover, the importance of each geographic feature can be obtained from the trained model, it represents the importance of each geographic feature in training the model.

3.5 Part 3: Rule Analysis

Part 3 - Rule Analysis aims to make the association rule analysis between accident locations and nearby geographic features. Figure 3.4 shows the methodology of Part 3 - Rule Analysis.

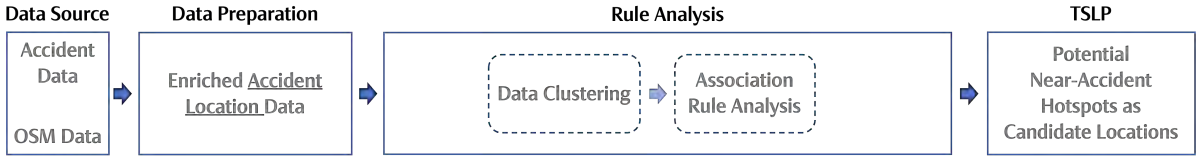

Figure 3.4 Methodology of Part 3 - Rule Analysis

The idea is to cluster the accident location data into different groups and within each of them, the accidents are somehow similar. Then the association rule analysis is made to explore the association rules with nearby geographic features. In other words, the association rules between one group of accidents and the nearby geographic features of the accident locations are explored. Therefore, those locations near accident-associated geographic features are just the candidate locations for traffic sensors.

What's more, the association rules of the whole traffic accident dataset can be analyzed. Some holistic rules could be found out. A comparison analysis between the results from the whole dataset and the results from the clusters can be made.

3.5.1 Data Clustering

The aim of data clustering is to divide the complex traffic accident dataset into several internally similar clusters. The data clustering method to be used is the community detection method, which is good at detecting internally similar community structures on the network. But there should be a network of accident data first.

To create a network, each accident is regarded as one case (one node in the network). If two accidents are somehow similar (like they share a certain number of identical attributes, which is more than a threshold), then one link between the two accidents is built, and the weight of the link is set based on the degree of similarity between the two accidents. The threshold here is a key parameter to set, a suitable threshold should make the edge number of the built network neither too large nor too small so that the community detection method can be effectively applied to the network. In this way, a network showing the similarity relation between each pair of accidents is built, and the community detection method can be used to detect the community within the built network. The detected communities are just the expected accident data clusters.

3.5.2 Association Rule Analysis

Association rule analysis is used here to find the association relationship within each group of accident data. Certain geographic features associated with a group of accident data are expected to be extracted.

Association rule analysis is often used to extract those rules in the form of "If... Then...". In this case, the rule would be "If there is a traffic accident, then there are (usually) some geographic features nearby". As introduced in 2.5, there are two common indicators of association rule analysis, namely the support indicator and the confidence indicator.

Considering the data characteristics of traffic accident location data, there are only negative values (location with accident) and no positive values (location without accident). In more detail, each piece of data shows an accident location, while there is no non-accident location. Only rules like "if an accident, then geographic feature Y" can be discovered, and rules like "if geographic feature Y, then an accident" can not be discovered, because the situations where there is "geographic feature Y" but there is no accident can not be obtained. Therefore, the confidence indicator can not be calculated, which is just used to judge whether the rule "if some geographic features are nearby, then there is (more likely) a traffic accident" is true.

However, traffic intersection data can have both negative and positive values (some traffic intersections have accident records while some do not), both the rules like "if an accident, then geographic feature Y" and "if geographic feature Y, then an accident" can be discovered. Therefore both the support indicator and confidence indicator can be calculated. This part can continue with the traffic intersection data to better discover potential association rules.

Considering that the analysis of the accident location has practical meaning and can make up a hierarchical methodology from the road network to traffic intersections and accident locations, this part can continue to use traffic location data.

4 Case Study

4.1 Study Area

In this work, the city of Wuppertal is selected as the study area.

Wuppertal is a city in North Rhine-Westphalia, the most populous state of Germany. With a population of approximately 355,000, Wuppertal is the seventh-largest city in North Rhine-Westphalia as well as the 17th-largest city in Germany (Wikipedia, 2023). Figure 4.1 shows the location of the city of Wuppertal (in dark blue colour) in North Rhine-Westphalia (in light blue colour) and in the administrative district of Düsseldorf (in medium blue colour).

Figure 4.1 Location of the City of Wuppertal

Choosing a medium-sized city like Wuppertal as the study area can provide an appropriate amount of data for the case study so that the amount of data is not too large or too small. If the data amount is too large, the corresponding processing time will get too long, while if the data amount is too small, the analysis results may not be representative.

Meanwhile, the city of Wuppertal is promoting its smart city construction strategy and has begun to lay out roadside LiDAR sensors on a relatively large scale. The analysis results of this thesis can provide a reference for the city of Wuppertal and hopefully make up a continuous and complete study from the location selection to implementation and data application.

What's more, in Part 2 Risk Analysis, the data of four nearby cities and counties of Wuppertal, which are Ennepe-Ruhr-Kreis, Kreis Mettmann, Solingen, and Remscheid, are used for risk prediction model building to predict the risk of traffic intersections in Wuppertal. Their geographical proximity and similar geographical and social environ-

ments make the prediction reliable. Figure 4.2 shows the location of these four cities and counties (in medium blue colour). The data of another nearby county Oberbergischer Kreis (southeast of Wuppertal) hasn't been used for the model training, because its area is too large and its main part is far away from Wuppertal, which may influence the model performance.

Figure 4.2 Locations of the Four Cities and Counties around Wuppertal

4.2 Data Source

The data used in the case study are all open data. There are two main data sources, one is the German Accident Atlas, and the other is OpenStreetMap.

4.2.1 German Accident Atlas

The German Accident Atlas (https://unfallatlas.statistikportal.de/) is published by the Federal Statistical Office and is updated annually. It records all the road traffic accidents that happened in Germany due to vehicular traffic on public roads or places, with persons killed or injured or involving material damage. Accidents involving only material damage are not contained. The data is based on reports from police stations, so accidents, where the police were not called, are not contained.

For each road traffic accident record, 24 attributes (as shown in Table 4.1) are included. The attributes include the unique accident ID, location information (geographic coordinates and administrative divisions), time information (year, month, weekday and hour), environment information (light and road conditions) and accident attribute information (accident type and kind).

Table 4.1 The Attribute List of Road Traffic Accident Record

Column Name	Content
ID	Serial Number
ULAND	Land
UREGBEZ	Administrative Region
UKREIS	Administrative District
UGEMEINDE	Municipality
UJAHR	Year of Accident
UMONAT	Month of Accident
USTUNDE	Hour of Accident
UWOCHENTAG	Day of the Week
UKATEGORIE	Road Accidents Involving Personal Injury
UART	Kinds of Accidents
UTYP1	Type of Accidents
ULICHTVERH	Light Conditions
IstRad	Accident with Bicycle
IstPKW	Accident with Passenger Car
IstFuss	Accident with Passenger
IstKrad	Accident with Motorcycle
IstGkfz	Accident with Goods Road Vehicle
IstSonstige	Accident with Other
USTRZUSTAND	Road Surface Conditions
LINREFX	Coordinates of the Place of Accident
LINREFY	(ETRS 89 / UTM Zone 32N)
XGCSWGS84	Coordinates of the Place of Accident
YGCSWGS84	(WGS84)

Table 4.2 13 Road Classes Extracted from OSM

OSM Tag	**Description**
highway = motorway	A restricted access major divided highway.
highway = motorway_link	
highway = trunk	The most important roads in a country's system that aren't motorways.
highway = trunk_link	
highway = primary	The next most important roads in a country's system. (Often link larger towns.)
highway = primary_link	
highway = secondary	The next most important roads in a country's system. (Often link towns.)
highway = secondary_link	
highway = tertiary	The next most important roads in a country's system. (Often link smaller towns and villages)
highway = tertiary_link	
highway = unclassified	The least important roads in a country's system. (Often link villages and hamlets.)
highway = residential	Roads that serve as access to housing.
highway = living_street	The residential streets where pedestrians have legal priority over cars.

Table 4.3 8 Classes of Traffic Facilities Extracted from OSM

OSM Tag	**Description**
highway = bus_stop	A small bus stop.
highway = crossing	A.k.a. crosswalk.
highway = give_way	A give way sign.
highway = motorway_junction	Indicates a junction or exit.
highway = speed_camera	A fixed roadside or overhead speed camera.
highway = stop	A stop sign.
highway = traffic_signals	Lights that control the traffic.
highway = turning_circle	A rounded, widened area to facilitate easier vehicle turning.

Table 4.4 8 Classes and 6 Kinds of Amenities Extracted from OSM

OSM Tag	Description
(amenity = sustenance)	bar, biergarten, cafe, fast_food, food_court, ice_cream, pub, restaurant
(amenity = education)	college, driving_school, kindergarten, language_school, library, toy_library, research_institute, training, music_school, school, traffic_park, university
(amenity = transportation)	bicycle_parking, bicycle_repair_station, bicycle_rental, boat_rental, boat_sharing, bus_station, car_rental, car_sharing, car_wash, compressed_air, vehicle_inspection, charging_station, driver_training, ferry_terminal, fuel, grit_bin, motorcycle_parking, parking, parking_entrance, parking_space, taxi
(amenity = financial)	bank, bureau_de_change
(amenity = healthcare)	clinic, dentist, doctors, hospital, nursing_home, pharmacy, social_facility, veterinary
(amenity = entertainment)	arts_centre, brothel, casino, cinema, community_centre, conference_centre, events_venue, exhibition_centre, fountain, gambling, love_hotel, music_venue, nightclub, planetarium, public_bookcase, social_centre, stripclub, studio, swingerclub, theatre
(amenity = public_service)	courthouse, fire_station, police, post_box, post_depot, post_office, prison, ranger_station, townhall
(amenity = facilities)	bbq, bench, dog_toilet, dressing_room, drinking_water, give_box, mailroom, parcel_locker, shelter, shower, telephone, toilets, water_point, watering_place
amenity = kindergarten	For children too young for a regular school.
amenity = school	Primary, middle and secondary schools.
amenity = doctors	A doctor's practice or surgery.
amenity = hospital	A hospital providing in-patient medical treatment.
amenity = social$_{f}acility$	A facility that provides social services.
(amenity = parking)	motorcycle_parking, parking, parking_entrance, parking_space

Table 4.5 26 Kinds of Land Use Types Extracted from OSM

OSM Tag	Description
landuse = allotments	A piece of land given over to local residents for growing vegetables and flowers.
landuse = animal_keeping	An area of land that is used to keep animals, particularly horses and livestock.
landuse = basin	An area artificially graded to hold water.
landuse = brownfield	Describes land scheduled for new development where old buildings have been demolished and cleared.
landuse = cemetery	Place for burials.
landuse = commercial	Predominantly commercial businesses and their offices.
landuse = construction	A site under active development and construction of a building or structure.
landuse = education	An area predominately used for educational purposes.
landuse = farmland	An area of farmland used for tillage.
landuse = farmyard	An area of land with farm buildings.
landuse = flowerbed	An area designated for flowers.
landuse = forest	Managed forest or woodland plantation.
landuse = garages	One-level buildings with boxes commonly for cars.
landuse = grass	An area of mown and managed grass.
landuse = greenfield	Describes land scheduled for new development where there have been no buildings before.
landuse = greenhouse_horticulture	Area used for growing plants in greenhouses
landuse = industrial	Predominantly industrial land uses such as workshops, factories, or warehouses.
landuse = meadow	Land primarily vegetated by grass and non-woody plants.
landuse = orchard	Intentional planting of trees or shrubs maintained for food production.
landuse = plant$_{n}ursery$	Intentional planting of plants maintaining for the production of new plants.
landuse = railway	Area for railway use.
landuse = recreation_ground	An open green space for general recreation.
landuse = religious	An area used for religious purposes.
landuse = residential	Land where people reside.
landuse = retail	Predominantly retail businesses such as shops.
landuse = village_green	A distinctive area of grassy public land in a village centre.

4.2.2 OpenStreetMap

OpenStreetMap (https://www.openstreetmap.org/), OSM for short, is a free, open geographic database updated and maintained by a community of volunteers via open collaboration.

In the case study, road network data (OSM tag: *highway*), POI data (OSM tag: *amenity*), land use data (OSM tag: *landuse*) and other data of the study area are extracted from the OSM database for further analysis. The data used are introduced in detail as follows.

OSM tag *highway* is used to describe roads and footpaths. Table 4.2 shows 13 road classes used in the case study. They are all drivable roads rather than paths. Actually, there are 14 more road and path classes in OSM, like "service" and "pedestrian". Considering that more than 99% of the recorded traffic accidents happened along the first 13 road classes, and introducing the last 14 road classes will double the data volume, only the first 13 road classes are extracted and used for further analysis.

Also, OSM tag *highway* is used to describe traffic facilities, such as traffic signals and crossings. Table 4.3 shows 8 kinds of traffic facilities used in the case study.

OSM tag *amenity* is used to describe all kinds of facilities used by visitors and residents. Table 4.4 shows 8 classes of amenities and 6 kinds of amenities used in the case study. There is no specific tag for each class of amenity, so multiple kinds of amenities are aggregated into one class, according to the classification in OSM. 6 kinds of amenities with vulnerable road users gathering are specifically selected for further analysis.

OSM tag *landuse* is used to describe the purpose for which an area of land is being used. Table 4.5 shows 26 kinds of land use types used in the case study.

4.3 Part 0: Data Preparation

As shown in Figure 4.3, several data preparation works need to be done before the analysis, including 4.3.1 road network modelling, 4.3.2 road intersection detection, 4.3.3 accident data filtering, and 4.4.4 data enrichment.

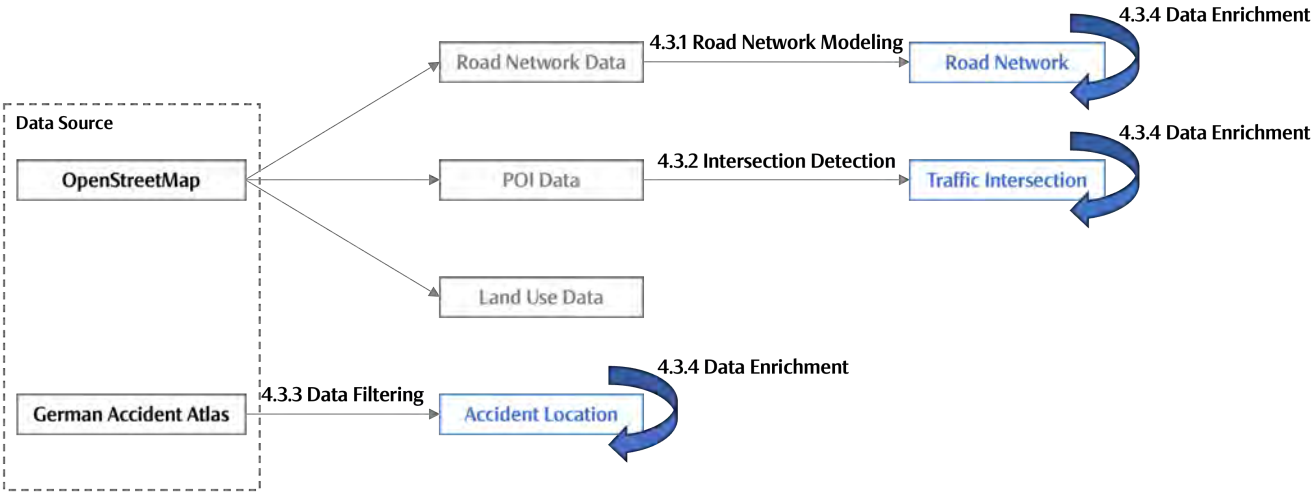

Figure 4.3 The Workflow of Data Preparation

After these data preparation works, the analysis objects of each part are obtained, which are road network of network analysis, traffic intersections of risk analysis and accident locations of rule analysis respectively.

The details of each step are described below.

4.3.1 Road Network Modelling

Figure 4.4 Road Network of Wuppertal after Modelling

The road data directly acquired from OSM is rather detailed and not neat, so road network modelling is needed to get a qualified road network.

Python package OSMnx (https://osmnx.readthedocs.io/) is used to achieve the aim. All the drivable public streets (but not service roads, like internal roads in factories and parking lots) within the boundary of the study area are acquired as the initial road network. The road network's topology is simplified by removing interstitial nodes, that is, all nodes that are not intersections or dead-ends. But the course of the road is preserved.

Figure 4.4 shows the road network modelling results of Wuppertal. Nodes, namely road intersections, are in dark blue colour and edges, namely road segments, are in light blue colour. It's worth mentioning that the bottom right part of the network is not connected to the main part of the network (actually the small part connects the main part through the road network outside the city of Wuppertal). Therefore, this modelled road network consists of two disconnected sub-networks. This will influence the latter network analysis results. But considering that it's a rather small part, the influence is rather limited.

The further network analysis will be based on this modelled road network.

4.3.2 Traffic Intersection Detection

Figure 4.5 An Example Showing the Relationship between a Traffic Intersection in the Practical Sense (Source: OSM) and in the Mathematical Sense

Figure 4.6 The Satellite Image of the Example Traffic Intersection (Source: Apple Maps)

A traffic intersection is defined as a junction where two or more roads converge, diverge, meet or cross at the same height. There is no such data directly representing traffic intersections in OSM. So traffic intersections need to be extracted from other data in OSM. In this thesis, such a process is called Traffic Intersection Detection.

In this case, traffic intersections in the practical sense and in the mathematical sense need to be distinguished. Taking Figure 4.5 as an example, the left graph (a screenshot on OSM) shows one traffic intersection in Wuppertal, it can be modelled as shown in the right graph. In the mathematical sense, there are four intersections, because roads with a certain width are treated as lines without widths and their meet creates four intersections. But actually, as shown in Figure 4.6, there is only one traffic intersection in the practical sense, because the four intersections have practical widths and they merge together as one traffic intersection.

In other words, if traffic intersections are to be detected from the modelled road network, those intersections within a limited distance from each other are only intersections in the mathematical sense. They are caused by ignoring the width of roads during road network modelling. In the practical sense, there is only one traffic intersection.

Based on this principle, the common method for traffic intersection detection is to aggregate nearby "intersections" of the road network, namely the road network nodes as shown in Figure 4.4. For example, the function *consolidate_intersections* of the Python package OSMnx just follows this idea to consolidate nodes as traffic intersections. The common parameter of maximum distance between nearby nodes should be considered with the design specifications of the local road system, 10 meters is an empirical value.

However, the number of results of such a process could be relatively large. One way to reduce the number and keep the ones of higher importance is to select only those intersections of higher-class roads. In this thesis, another method is proposed to achieve a similar aim even more effectively. In OSM data, there is one class of data called traffic signals. These traffic signals can help to identify those important traffic intersections in a city.

In this case study, a buffer of 22.5m (empirical value, and the width of a two-way six-lane road with a lane width of 3.75m) is selected to aggregate nearby traffic signals as one traffic intersection, and the centroid of the dissolved buffers is set as the location of the detected traffic intersection. The selection aim of the buffer width is to merge the intersections belonging to one traffic intersection and to separate the intersections belonging to different traffic intersections.

The detected 345 traffic intersections in Wuppertal are shown in Figure 4.7.

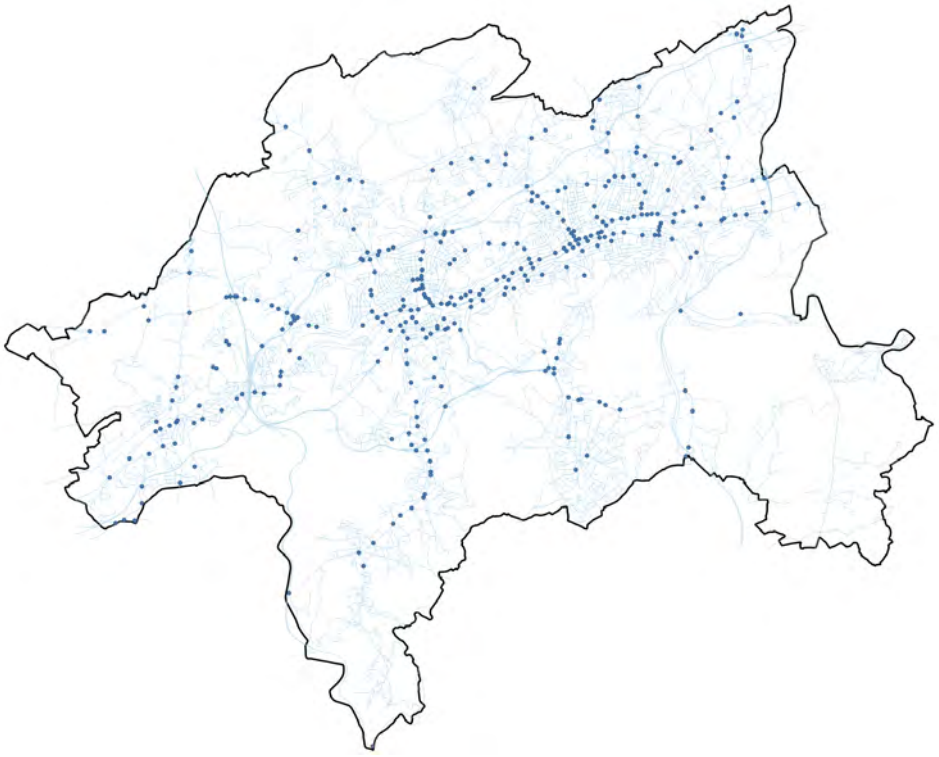

Figure 4.7 Traffic Intersections Detected in Wuppertal

4.3.3 Accident Data Filtering

As shown in Table 4.6, the road traffic accident data is filtered using its attributes ULAND, UREGBEZ, UKREIS, and UGEMEINDE, which together make up the Official Municipality Key (AGS), so that only the accident data happened in the study area is kept. Among them, $ULAND = 05$ refers to North Rhine-Westphalia, $UREGBEZ = 1$ and $UREGBEZ = 9$ refer to Regierungsbezirk Düsseldorf and Regierungsbezirk Arnsberg respectively.

Table 4.6 Official Municipality Key of the Study Area

City/County	Car Sign	ULAND	UREGBEZ	UKREIS	UGEMEINDE	AGS
Wuppertal	W	05	1	24	000	05124000
Solingen	SG	05	1	22	000	05122000
Remscheid	RS	05	1	20	000	05120000
Kreis Mettmann	ME	05	1	58	000	05158000
Ennepe-Ruhr-Kreis	EN	05	9	54	000	05954000

The German Accident Atlas includes the accident data of North Rhine-Westphalia from the year 2019 to the year 2022. The four years of data will be used in the following parts. Table 4.7 shows the number of traffic accidents that occur each year in Wuppertal.

Table 4.7 Accident Number of Each Year in Wuppertal

	Year 2019	**Year 2020**	**Year 2021**	**Year 2022**	**Total**
Wuppertal	910	770	814	902	3396

4.3.4 Data Enrichment

After the previous data preparation works as introduced in 4.3.1, 4.3.2 and 4.3.3, the study objects of the following three parts are obtained, that is, road network, traffic intersections, and accident locations. The next step is the so-called data enrichment, which means using the two data sources to enrich each study object for further analysis.

For the road network, the nearby accident data can enrich it. In this case study, the accident linear density (accident number per unit length of road) of each edge (road segment) of the road network is added to the road network. The data quality of the German Accident Atlas is rather high, the accident location is registered in the middle of the corresponding road. Buffers of 10 meters are created for each road segment to count the accident number for each road segment.

For traffic intersections, both the traffic accident data and OSM data can enrich them. In this case study, the accident number and the surrounding geographic features, like the road class, road number, nearby amenity and land use type, are added to each traffic intersection. Table 4.8 shows all the attributes added to each traffic intersection. Buffers of 200m are created for each traffic intersection and the corresponding geographic features and traffic accidents intersecting with the buffers are extracted. 200m is an empirical value based on the test results. Among the geographic features of 100m, 200m, 400m and 1000m, geographic features of 200m show the highest importance in the model training,

Table 4.8 Attribute Table of Traffic Intersections after Data Enrichment

Attribute Name	Content	Category
nearbyNodes	number of nearby road network nodes	1
nearbyEdges	number of nearby road network edges	1
nearbyEdgeClass	number of nearby road network edge classes (i.e. road classes)	1
nearMotorway	whether near motorway or motorway_link road	1
nearTrunk	whether near trunk or trunk_link road	1
nearPrimary	whether near primary or primary_link road	1
nearSecondary	whether near secondary or secondary_link road	1
nearTertiary	whether near tertiary or tertiary_link road	1
nearUnclassified	whether near unclassified road	1
nearResidential	whether near residential road	1
nearLivingStreet	whether near living_street road	1
nearbyBusStop	number of nearby bus stops	2
nearbyCrossing	number of nearby crossings	2
nearbyGiveWay	number of nearby give way signs	2
nearbyMotorwayJunction	number of nearby motorway junctions	2
nearbySpeedCamera	number of nearby speed cameras	2
nearbyStop	number of nearby stop signs	2
nearbyTrafficSignals	number of nearby traffic signals	2
nearbyTurningCircle	number of nearby turning circles	2
nearbySustenance	number of nearby amenities of sustenance	3
nearbyEducation	number of nearby amenities of education	3
nearbyTransportation	number of nearby amenities of transportation	3
nearbyFinancial	number of nearby amenities of finance	3
nearbyHealthcare	number of nearby amenities of healthcare	3
nearbyEntertainment	number of nearby amenities of entertainment, arts & culture	3
nearbyPublicservice	number of nearby amenities of public service	3
nearbyFacilities	number of nearby amenities of facilities	3
nearbyKindergarten	number of nearby kindergartens	3
nearbySchool	number of nearby schools	3
nearbyDoctors	number of nearby doctors	3
nearbyHospital	number of nearby hospitals	3
nearbySocialfacility	number of nearby social facilities	3
nearbyParking	number of nearby car parks	3
nearAllotments	whether near land use type of allotments	4
nearAnimalKeeping	whether near land use type of animal keeping	4
nearBasin	whether near land use type of basin	4
nearBrownfield	whether near land use type of brownfield	4
nearCemetery	whether near land use type of cemetery	4
nearCommercial	whether near land use type of commercial	4
nearConstruction	whether near land use type of construction	4
nearEducation	whether near land use type of education	4
nearFarmland	whether near land use type of farmland	4
nearFarmyard	whether near land use type of farmyard	4
nearFlowerbed	whether near land use type of flowerbed	4
nearForest	whether near land use type of forest	4
nearGarages	whether near land use type of garages	4
nearGrass	whether near land use type of grass	4
nearGreenfield	whether near land use type of greenfield	4
nearGreenhouseHorticulture	whether near land use type of greenhouse horticulture	4
nearIndustrial	whether near land use type of industrial	4
nearMeadow	whether near land use type of meadow	4
nearOrchard	whether near land use type of orchard	4
nearPlantNursery	whether near land use type of plant nursery	4
nearRailway	whether near land use type of railway	4
nearRecreationGround	whether near land use type of recreation ground	4
nearReligious	whether near land use type of religious	4
nearResidentialLU	whether near land use type of residential	4
nearRetail	whether near land use type of retail	4
nearTrafficIsland	whether near land use type of traffic island	4
nearVillageGreen	whether near land use type of village green	4
isAccident	whether there is accident	5

so 200m is chosen. Actually, multiple buffers of different sizes can be created to get multiple levels of geographic feature data used for model training and prediction. For the sake of simplicity, this thesis only uses the 200m buffer.

For accident locations, OSM data can also enrich them with nearby geographic features. Buffers of 50m are created for each accident location and the corresponding geographic features intersecting with the buffers are extracted.

4.4 Part 1: Accident Hotspot Identification on Road Network

From three levels, accident hotspots of the road network of Wuppertal are identified.

4.4.1 Node Level: Hot Node Identification

As introduced in 3.3, betweenness centrality and PageRank centrality are to be used, and two updated road networks are to be built. For the first, the inversed accident linear density ld_i shown in the formula below is treated as the weight w_1 of each edge (each road segment) of the road network. For the second, the accident linear density ld shown in the formula below is treated as the weight w_2 of each edge (each road segment) of the road network.

In the first formula, parameter n is the number of nearby accidents of each edge within a buffer of 5 meters, parameter l is the length of each edge in meters, and parameter ld is the linear density of traffic accident of each edge, that is, how many accidents per hundred meters. In the second formula, the inversed accident linear density ld_i is calculated and scaled to suit further analysis.

The betweenness centrality and PageRank centrality of two updated road networks are calculated respectively. For comparison purposes, the centrality results of the original road network are also calculated. The results are shown in Figure 4.8 and Figure 4.9.

4.4.2 Lixel Level: Hot Lixel Identification

The network kernel density of the nearby traffic accidents of each lixel, which is set as 50 meters long (which is roughly the perception range of a LiDAR sensor or an HD camera), is calculated. According to the recommendations by Gelb (2021), the quartic kernel function is used, and the bandwidth is set as 300 meters.

The kernel density of the cumulative four years is calculated. The kernel density of each year is calculated respectively. The results are shown in Figure 4.10 and 4.11.

4.4.3 Community Level: Hot Community Identification

Based on the road network of Wuppertal, the community detection method is carried out. The improved accident linear density shown in the formula below is regarded as the weight w_3 of each edge of the road network. The design of the formula is up to a reasonable analysis results.

In this way, the higher the accident linear density of one edge, the larger the weight of the edge, thus the closer the relationship between two nodes of the edge. And community detection method can detect those communities of nodes with close relationships. In this case, the Louvain Algorithm is used for community detection.

$$w_3 = ld_{im} = \frac{n}{\sqrt{\frac{l}{1000}}} + 1$$

4.4.4 Results

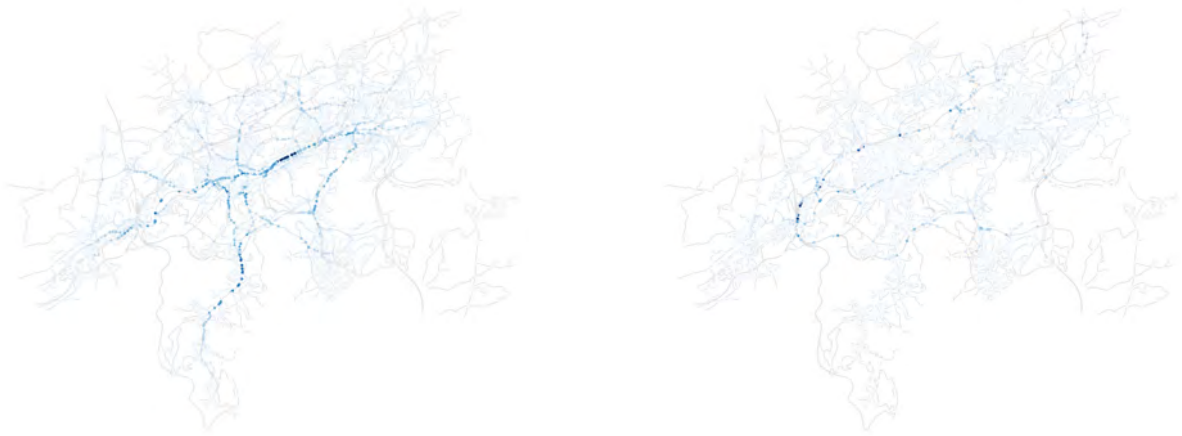

Figure 4.8 Betweenness Centrality Result (Left: Original Road Network, Right: Updated Road Network)

Figure 4.9 PageRank Centrality Result (Left: Original Road Network, Right: Updated Road Network)

Figure 4.8 shows the betweenness centrality results of the original road network (road length as the weight of each edge) and updated road network (inversed road traffic accident linear density as the weight of each edge). The darker the blue colour of the node, the higher its centrality.

The left graph represents the degree of each node lying on more shortest paths (the passing cost is the sum distance - road length, namely the sum weight of passing edges)

between any other two nodes on the network. The darker the blue colour, the node will lie on more shortest paths. On the right graph, with the inversed accident linear density as the weight, the passing costs of the high accident-risk nodes will decrease, and these nodes will lie on more shortest paths, which helps them get higher centrality. The dark blue nodes are just those high-centrality nodes.

Figure 4.9 shows the PageRank centrality results of the original road network (the weight of each edge is the same) and updated road network (road traffic accident linear density as the weight of each edge). The darker the blue colour of the node, the higher its centrality.

The left graph represents the degree of each node connecting to more edges. On the right graph, with the accident linear density as the weight, those nodes connecting to more edges with higher weight will get higher centrality. In the physical meaning, the dark blue nodes on the right graph are those connecting riskier and more edges.

Figure 4.10 shows the network kernel density estimation results of cumulative four years. The darker the blue colour of the lixel, the higher the accident density of the lixel, and therefore more accidents happened near the lixel in the past four years.

Figure 4.11 shows the estimation results of each year. There are small variations between years, but the general distribution of the hotspots is similar.

Figure 4.12 shows the community detection results. Those communities with high accident risk are successfully detected and highlighted in the figure. Their distribution is somehow consistent with the density estimation results in Figure 4.10.

Figure 4.10 NKDE Results (Years 2019-2022)

4.5 Part 2: Risk Prediction of Traffic Intersection

As shown in Figure 4.13, in this part, the random forest classifier model is trained with the data of four nearby cities/counties, and the accident risk of traffic intersections in

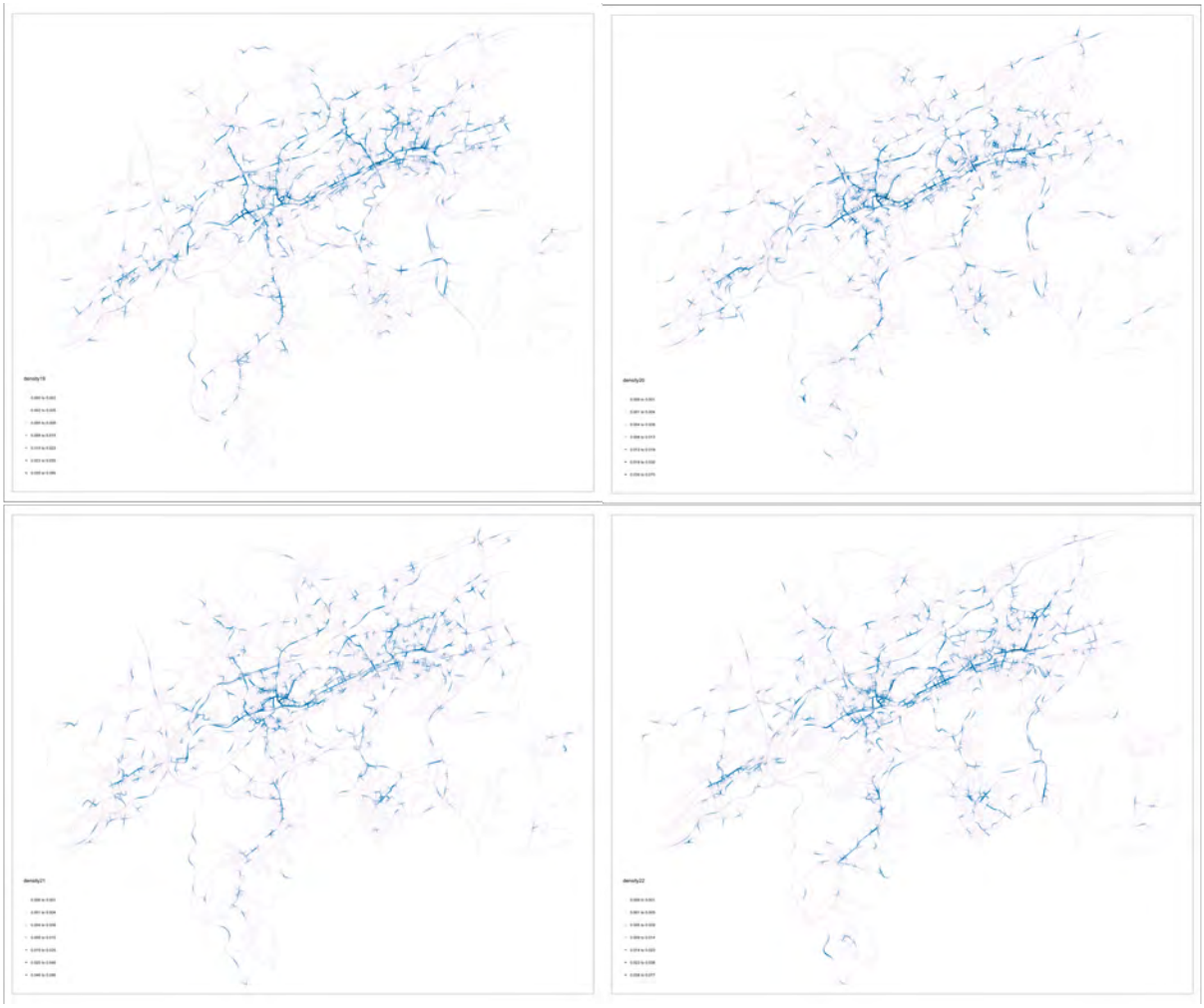

Figure 4.11 NKDE Results (Year 2019, Year 2020, Year 2021, Year 2022)

Figure 4.12 Community Detection Results

Wuppertal is predicted. These four nearby cities have similar geographic, cultural, economic and social environments to Wuppertal, so the model built with the data of these four cities should be applicable in Wuppertal.

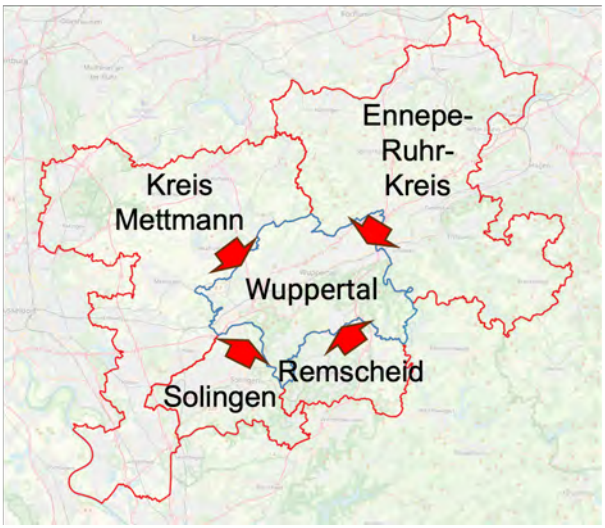

Figure 4.13 Schematic Diagram of Model Training

4.5.1 Random Forest Model Building

After the data preparation work in 4.3.2, 946 traffic intersections in the four nearby cities or counties are extracted, and 345 traffic intersections in Wuppertal are extracted. They are used for the model training and testing.

As shown in Table 4.8, 4 classes and 60 items of nearby geographic features are enriched into each traffic intersection, therefore the geographic environment of each traffic intersection is somehow represented. These items form the training and testing features of the model.

The accident risk of each traffic intersection is regarded as the training and testing labels of the model. In this case study, if the number of traffic accidents occurring in the traffic intersection (within the buffer of 100 meters of the centre of the traffic intersection) is less than 4, that is, less than or equal to 1 time per year, the traffic intersection will be regarded as low-risk, otherwise will be regarded as high-risk.

Therefore, the training and testing sets shown in Table 4.9 are prepared.

Table 4.9 Division of Training and Testing Sets

	Description	Number	Features	Labels
Training Set	four nearby cities/counties	946	946 * 60	946 * 1
Testing Set	Wuppertal	345	345 * 60	345 * 1

4.5.2 Model Enhancement and Risk Prediction

With the trained model, the accident risk of traffic intersections in Wuppertal is predicted. The predicted results are compared with truth values to verify the effectiveness of the

trained model. Figure 4.14 and Table 4.10 show the prediction results in the form of a confusion matrix and prediction scores respectively. Besides, the importance of each feature can be got out of the model as shown in Figure 4.17 and Table 4.11 so that the features highly related to traffic accidents can be identified.

In 3.4, it's designed with an intermediate step of model enhancement to improve the behaviours of the model. Because, in most tasks, correctness is the top aim of model training, like the pattern recognition task. However, given that the aim of the model is to reveal the relationship between traffic accident risk and geographic features rather than blindly pursuing the correct rate, an excessive pursuit could be counterproductive by covering up some key information. Therefore, in this case study, model enhancement is deprecated.

4.5.3 Results

		Prediction	
		Low-Risk	High-Risk
Fact	Low-Risk	155	22
	High-Risk	73	95

Figure 4.14 Prediction Results of the Model (Confusion Matrix)

Table 4.10 Prediction Scores of the Model

	Precision	Recall
Low-Risk	0.68	0.88
High-Risk	0.81	0.57

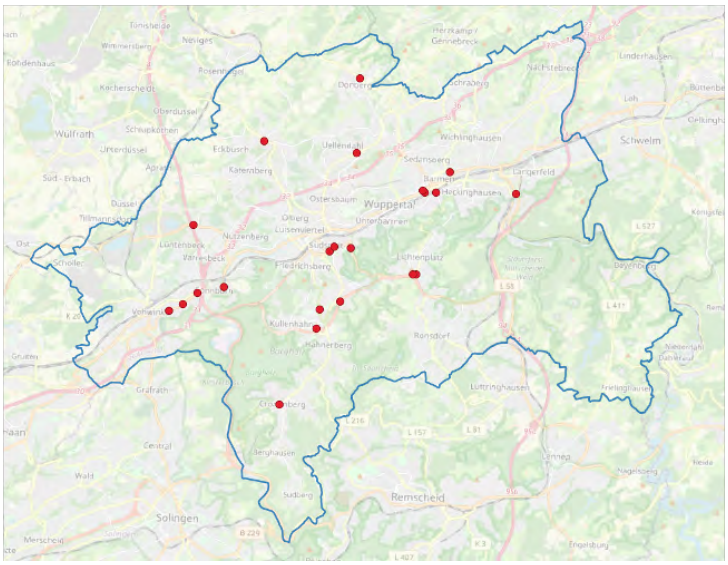

Figure 4.15 Geographic Distribution of Potential Near-Accident Hotspots

Figure 4.14 shows the prediction results of the model in the form of a confusion matrix. The columns show the predicted accident risk, and the rows show the actual accident risk. For example, 155 + 22 traffic intersections are actually low-risk, but 155 are predicted as low-risk, and 22 are predicted as high-risk. Based on the prediction results, the prediction scores are gotten as shown in Table 4.10. Precision is a score representing "among all items predicted as X, what percentage are actually X", while Recall is a score representing "among all items that are actually X, what percentage are predicted as X".

As shown in Figure 4.14 and Table 4.10, for those actual low-risk traffic intersections, the model behaves very well with an accuracy of 88%. And those predicted high-risk traffic intersections can be recognized as potential near-accident hotspots. Their geographic distribution is shown in Figure 4.15.

For those actual high-risk traffic intersections, the model behaves badly with an accuracy of 57%. However, as shown in Figure 4.16, the predicted low-risk ones concentrate on the slightly high-risk part. And in 70% of the traffic intersections, less than 2 accidents per year happened there. For those intersections with a high number of historical accidents, there could be some other causes than geographic features.

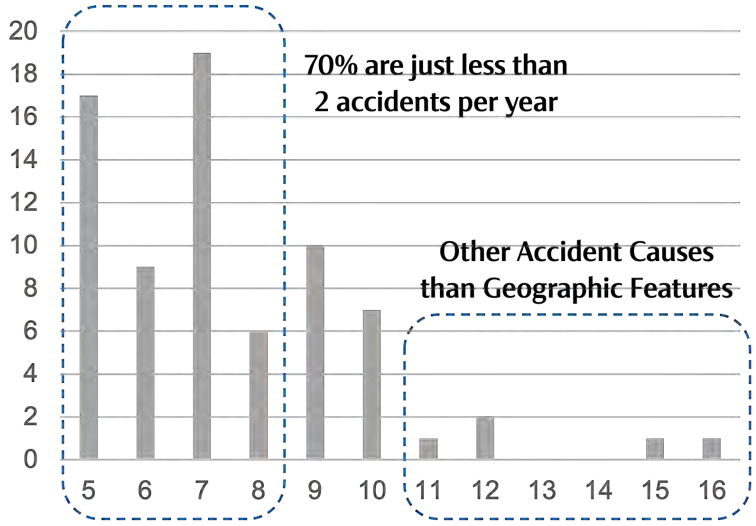

Figure 4.16 Accident Number of Actual High-Risk but Predicted Low-Risk Intersections

Table 4.11 Top Important Geographic Features of the Trained Model

Feature	Importance	Description
nearbyCrossing	0.10	number of crossing
nearbyEdges	0.08	number of road segment
nearbyNodes	0.06	number of intersection
nearbyTransportation	0.06	number of transportation amenities
nearbyEdgeClass	0.05	number of road class
nearbyParking	0.05	number of parking lots
nearbyBusStop	0.04	number of bus stops

The importance of each geographic feature is estimated by the trained model, as shown in Figure 4.17. And top 7 important geographic features are shown in 4.11. They are all transportation-relevant features without exception. Among them, the number of crossings

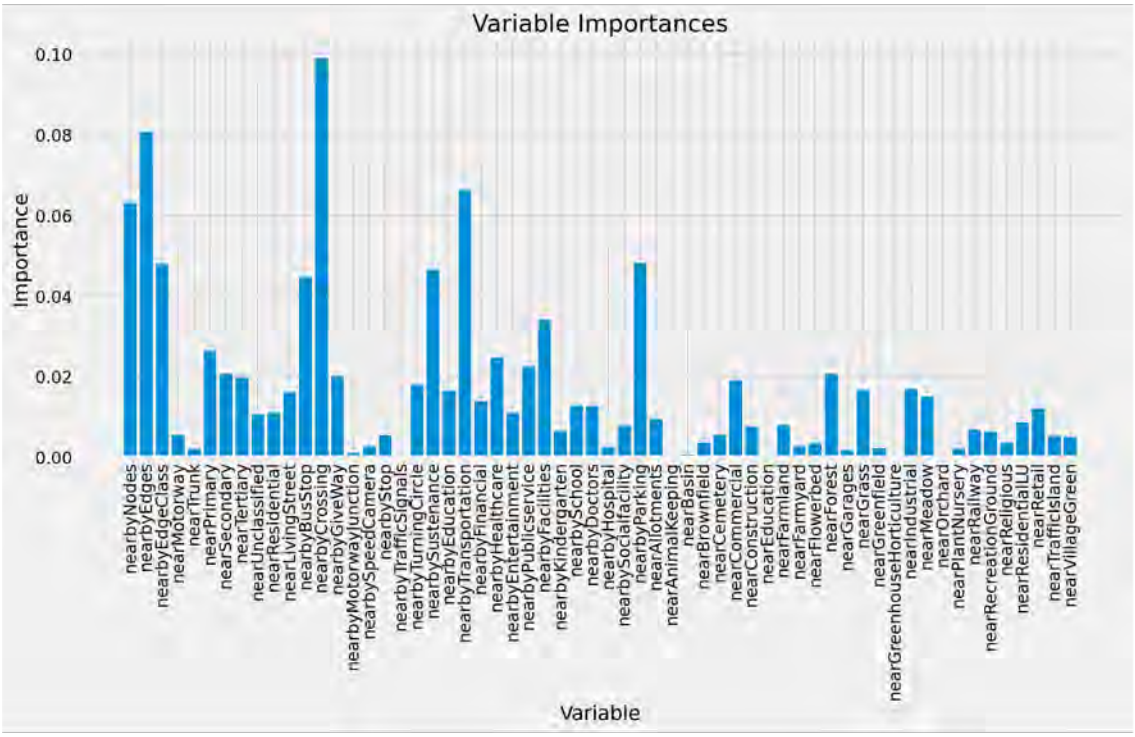

Figure 4.17 Histogram of the Importance of each Geographic Feature

is the most important feature, followed by the complexity of nearby road networks, that is, the number of nearby intersections, and the number and class number of nearby road segments.

Among road classes, the primary, secondary and tertiary roads are of the most importance in turn. Among POIs, besides the transportation class, the sustenance class of POIs is of the most importance. Among traffic facilities, crossings, bus stops and give-way signs are of high importance, while the others are of little importance, especially traffic signals are of no importance. Compared with the others, land use type is of lesser importance. And the forest land use type is surprisingly of more importance than any other land use type.

4.6 Part 3: Association Rule Analysis of Accident Location

The idea of association rule analysis is to explore the associated geographic features of accident locations in Wuppertal. The association rule analysis will be made on the whole accident dataset first to discover the overall rules. Meanwhile, considering the complexity and diversity of the accident data of the whole city, data clustering is used to group similar accidents, and then association rule analysis is made on the different groups again.

4.6.1 Data Clustering

Obviously, the easiest way of data clustering is to cluster data directly by their attributes as shown in Table 4.1, like the accident types, accident time, and accident conditions.

And community detection method is also used to make the data clustering.

Table 4.12 Part of the Attribute List of Road Traffic Accident Record

Column Name	Content
UJAHR	Year of Accident
UMONAT	Month of Accident
USTUNDE	Hour of Accident
UWOCHENTAG	Day of the Week
UKATEGORIE	Road Accidents Involving Personal Injury
UART	Kinds of Accidents
UTYP1	Type of Accidents
ULICHTVERH	Light Conditions
IstRad	Accident with Bicycle
IstPKW	Accident with Passenger Car
IstFuss	Accident with Passenger
IstKrad	Accident with Motorcycle
IstGkfz	Accident with Goods Road Vehicle
IstSonstige	Accident with Other
USTRZUSTAND	Road Surface Conditions

Over four years, Wuppertal has seen 3,396 accidents. As shown in Table 4.1, each accident record has 24 attributes. Besides ID and other location attributes, 15 attributes shown in Table 4.12 are to be used to tell the similarity between two accidents.

Figure 4.18 Community Detection Results (Visualized with Gephi)

For community detection, the first step is to build a network used for the community detection method. Each of the 3396 accidents is regarded as a node of the network. The number of identical attributes between any pair of accidents is checked for network building. If the number of identical attributes between one pair of accidents is larger than the threshold, then one edge is built between this pair of accidents. The threshold is set to 14 out of 15 so that the built edges between accidents will not be too many or

too few, and therefore the detected communities will not be too many or too few. Thus, the network showing the similarity between accidents is built.

With the built network, the community detection method is applied to it. The community detection method used here is the Louvain Algorithm. As introduced in 2.4.3, this algorithm is a heuristic algorithm, so the results change from time to time, but the degree of change is acceptable, and the results tend to be similar. Figure 4.18 shows the community detection results, where 6 communities are detected. Except for the community in green colour, other communities intertwine with each other, which shows that the community detection results are not ideal. The current method cannot effectively cluster accident data.

4.6.2 Association Rule Analysis between Accident Locations and Geographic Features

The Apriori algorithm introduced in 2.5.3 is used for association rule analysis. It aims to extract frequent item sets. In this case, it shows whether, for a certain group of accidents, there are some features in common, that is, whether they are commonly near certain geographic features.

Association rule analysis is used for the whole dataset, certain types of accidents, and the resulting clusters of community detection. The results for the whole dataset are shown in 4.13.

Table 4.13 ARA Results of the Whole Dataset

No.	Support	ItemSets	Length
1	0.804406	(nearResidentialLU)	1
2	0.656445	(nearResidential)	1
3	0.570408	(nearResidentialLU, nearResidential)	2
4	0.477821	(nearTransportation)	1
5	0.417684	(nearCrossing)	1
6	0.392676	(nearParking)	1
7	0.392676	(nearTransportation, nearParking)	2
8	0.379280	(nearSecondary)	1
9	0.376600	(nearTransportation, nearResidentialLU)	2
10	0.342662	(nearCrossing, nearResidentialLU)	2
11	0.335814	(nearTransportation, nearResidential)	2
12	0.333135	(nearForest)	1
13	0.317059	(nearSecondary, nearResidentialLU)	2
14	0.313188	(nearBusStop)	1
15	0.302173	(nearTransportation, nearResidentialLU, nearParking)	3
16	0.302173	(nearResidentialLU, nearParking)	2
17	0.301876	(nearCommercial)	1
18	0.292647	(nearCrossing, nearResidential)	2
19	0.292647	(nearPrimary)	1
20	0.292349	(nearGiveWay)	1
21	0.281929	(nearResidential, nearParking)	2
22	0.281929	(nearTransportation, nearResidential, nearParking)	3
23	0.277464	(nearTertiary)	1
24	0.270914	(nearTransportation, nearResidentialLU, nearResidential)	3
25	0.265853	(nearResidentialLU, nearBusStop)	2

4.6.3 Results

Table 4.14 ARA Results of Certain Types of Accidents

No.	Support	Itemsets
1	0.615385	(nearResidentialLU)
2	0.538462	(nearForest)
3	0.407692	(nearResidential)
4	0.400000	(nearMotorway)
5	0.392308	(nearPrimary)
6	0.376923	(nearCrossing)
7	0.369231	(nearTransportation)
8	0.276923	(nearParking)
9	0.269231	(nearCommercial)

(a) Accidents Involving Trucks

No.	Support	Itemsets
1	0.866460	(nearResidentialLU)
2	0.810559	(nearResidential)
3	0.566770	(nearTransportation)
4	0.472050	(nearParking)
5	0.428571	(nearCrossing)
6	0.372671	(nearBusStop)
7	0.368012	(nearSecondary)
8	0.315217	(nearSustenance)
9	0.315217	(nearGiveWay)

(b) Accidents Involving Passengers

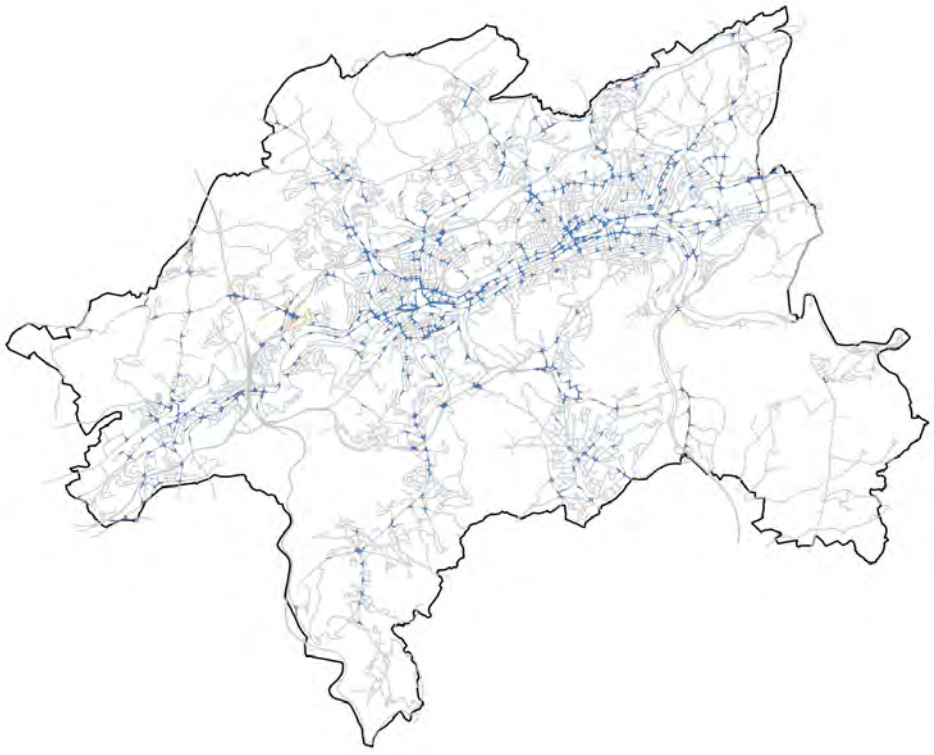

Figure 4.19 Parts of the Road Network that is Close to Crossings

Table 4.13 shows the association rules between the whole traffic accident dataset and nearby geographic features. It's ranked with the support measure, which means for a certain proportion (*Support*) of accidents, they are near certain geographic features (*Itemsets*). The *Length* means the length of *Itemsets*, which means how many items are in the *Itemsets*. According to the results, 80.4% accidents occurred near residential land use type, 65.6% accidents occurred near residential roads, and 47.8%, 41.8%, and 39.3% accidents occurred near transportation amenities, crossings and parking lots, respectively.

Table 4.14 shows the association rules between the accidents involving (a) trucks and (b) passengers and nearby geographic features. Compared with the results of the whole dataset in 4.13, they have shown significant differences. Accidents involving trucks significantly less occurred near the residential land use type and residential roads, while they

significantly more occurred near the forest land use type, motorway roads and primary roads. Accidents involving passengers significantly more occurred near the residential land use type, residential roads and transportation amenities. These are just two examples, other clusters of accidents with certain attributes also show some significant differences compared to the whole dataset, like different accident times and accident types. Due to space limitations, no further details are provided here.

Taking crossing as an example, for each crossing, a buffer of 50 meters is built, and the part of the road network that intersects buffers is shown in blue colour. Therefore, the potential accident hotspots associated with crossings are shown in Figure 4.19.

However, the association rule analysis of the accident clusters from community detection doesn't show enough significant differences compared with the results of the whole dataset. In a sense, this is like a random sampling of the whole dataset.

All in all, there are surely some accident-associated geographic features and different accident data clusters can show different association rules relevant to geographic features. It is worth pointing out that clustering data by their single attribute is more effective in revealing association rules than community detection, which takes an overall consideration of the similarity between accidents.

5 Discussions

In this chapter, the discussions about the results of the case study are made, and the connections between the case study, methodology and research gap are built.

Specifically speaking, it first connects the analysis process and decision-making, namely sensor location selection, and discusses the candidate sensor locations obtained from the analysis results. Secondly, it answers the research questions proposed in 1.3. Thirdly, it discusses the relationship between traffic accidents and geographic features and its prospective application. Fourthly, it discusses the difference between different analysis and their results. Fifthly, it discusses the limitations of this thesis and lists where to continue in the future work.

5.1 From Analysis Results to Candidate Sensor Locations

In the case study, network analysis, risk analysis and rule analysis are made respectively in 4.4, 4.5 and 4.6.

4.4 uses network analysis methods of three different levels to identify accident hotspots on the road network. Given that these identified hotspot results are all based on historical data, they are called historical accident hotspots in the following. These historical accident hotspots can be regarded as candidate sensor locations.

4.5 uses the random forest classification model representing the relationship between traffic accident risk and nearby geographic features to predict the accident risk of traffic intersections in another city. Given that these high-risk traffic intersections are based on the prediction, they are called potential accident hotspots. Among them, those predicted high-risk but actual low-risk traffic intersections are called potential near-accident hotspots. Because even if there were few historical traffic accidents, there are still potential near-accident hotspots. These potential accident and near-accident hotspots can be regarded as candidate sensor locations.

4.6 uses the association rule analysis method to obtain the accident-relevant geographic features, like residential roads, transportation amenities, crossings and parking lots. Then the parts of the road network that are close to these accident-relevant geographic features have potential accident risks. They are called potential accident hotspots. These potential accident hotspots can be regarded as candidate sensor locations.

Therefore, candidate sensor locations can be divided into two categories: historical accident hotspots and potential accident and near-accident hotspots. The former is obtained in 4.4 and discussed in 5.1.1, and the latter is obtained in 4.5 and 4.6, and discussed in 5.1.2.

5.1.1 Historical Accident Hotspots as Candidates

On three levels of road networks, that is, node, lixel and community, historical accident hotspots are identified.

As shown in Figure 4.8 (Right) and Figure 4.9 (Right), the historical hotspots on the node level are extracted using centrality measures. The nodes with a darker blue colour

are the high centrality nodes, which are the extracted historical hotspots and therefore the candidate sensor locations.

As shown in Figure 4.10 and Figure 4.11, the historical hotspots on the lixel level are extracted using network kernel density estimation. The lixels with a darker blue colour are the high kernel density lixels, which are the extracted historical hotspots and therefore the candidate sensor locations.

As shown in Figure 4.12, the historical hotspots on the community level are extracted using the community detection method. The highlighted communities (multiple small-scale road networks) are the high-risk communities, which are the candidate sensor locations.

On these three different levels, besides the geographic distribution of historical traffic accidents, the topology of the road network, namely the structure of the road network, always more or less influences the analysis results.

5.1.2 Potential Accident and Near-Accident Hotspots as Candidates

4.5 uses the random forest classification model to predict the potential accident hotspots. As shown in Figure 4.14, 117 traffic intersections are predicted as high-risk, and they are regarded as potential accident hotspots. Among them, 22 traffic intersections are actually low-risk as shown in Figure 4.15, and they are regarded as potential near-accident hotspots.

4.6 uses the Apriori algorithm to detect the frequent geographic features that are near accident locations. As shown in Table 4.13, for the whole accident dataset, some accident-associated geographic features are detected. As shown in Table 4.14, for certain types of traffic accidents, some accident-associated geographic features are also detected. Taking crossing as an example, Figure 4.19 shows the potential near-accident hotspots associated with crossings in the road network.

5.2 Answers to Research Questions

For the five research questions proposed in 1.3, according to the results of the case study made in this thesis, five answers are made respectively as follows:

Answer 1: Research Question 1 is answered by Chapter 2 - Literature Review. As introduced in 2.3 and 2.1, the functions of traffic sensors determine their location strategy, and they focus on traffic flow monitoring. Therefore, the current TSLP can be summarised as the traffic flow-oriented TSLP. As introduced in 2.2 and 2.1, research on using traffic sensors to monitor traffic safety is still rare, so the discussion about the corresponding location strategy, that is traffic safety-oriented TSLP is also very rare. The difference between these two types of TSLP is mainly their different monitor aims and objects, which source from their difference in functions. Traffic flow-oriented TSLP tends to monitor as much traffic flow as possible with the least traffic sensors, while the traffic safety-oriented tends to monitor as many traffic accident or near-accident events as possible with the least traffic sensors.

Answer 2: Research Question 2 is answered by Chapter 4 - Case Study. Traffic safety-oriented TSLP can be solved by detecting traffic accident hotspots and near-accident

hotspots. According to the results of the case study, network analysis methods, machine learning and data mining methods are beneficial.

Answer 3: Research Question 3 is answered in 4.4. According to the results of the case study, historical traffic accident hotspots can be extracted by multiple network analysis methods. The hotspots on multiple levels including node, lixel and community can be successfully extracted with centrality measures, network kernel density estimation and community detection, respectively. Therefore, those extracted historical traffic accident hotspots can be regarded as candidate locations for traffic sensors.

Answer 4: Research Question 4 is answered in 4.5. According to the model performance, the relationship between the accident risk of one traffic intersection and its surrounding environment can be modelled with the random forest classification model, and the accident risk of traffic intersections in one city can be predicted. Therefore, those predicted high-risk intersections, can be regarded as candidate locations for traffic sensors.

Answer 5: Research Question 5 is answered in 4.6. The association rules between traffic accidents and geographic features can be extracted. Therefore, those locations near accident-associated geographic features can be regarded as candidate locations for traffic sensors.

5.3 Relationship between Traffic Accidents and Geographic Features

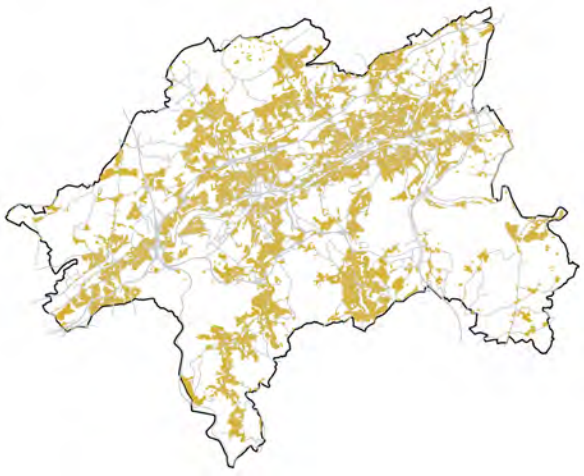

Figure 5.1 Distribution of Residential Land Use Types in Wuppertal

Both 4.5 Risk Analysis and 4.6 Rule Analysis show some accident-associated geographic features. In 4.5, Figure 4.17 and Table 4.11 show the importance of geographic features to the trained model, which reflects which geographic features are of high association to the accident risks. In 4.6, Table 4.13 shows the association rule analysis results of the whole dataset, which reflects which geographic features are accident-associated. These two results have shown strong consistency, where crossings, parking lots, transportation amenities and bus stops commonly rank high. However, there are also some inconsistencies between the two results. Table 4.13 shows that residential roads and residential land use types are the strongest associated geographic features, while Figure 4.17 shows that residential roads and residential land use types are of small importance in the model.

This is because the residential roads and residential land use types already exist in large numbers in the city as shown in 5.1, and therefore most accidents are inevitably near certain residential roads and residential land use types.

It's worth pointing out that the relationship found in this thesis between traffic accidents and nearby geographic features (risk high or low, accident or not) is an association relationship rather than a causation relationship. These geographic features are not the causes of traffic accidents, but they can serve as the symptoms of accident risk. The mechanism works in such a way that the accident risk gets higher if near certain geographic features and gets lower if near others.

This relationship can be used to locate potential accident or near-accident hotspots, and therefore traffic sensors can be located there. However, like what the examples of residential roads and residential land use types have revealed, the analysis of the association relationship is still incomplete and can be continued with further work. Not only the association rules from traffic accidents to geographic features, the association rules from geographic features to traffic accidents also need to be analyzed. The discussion about this is continued in 5.5 (2).

5.4 Differences between Three Analysis Methods

The three analysis methods proposed in this thesis, namely network analysis, risk analysis and rule analysis, have some differences, which leads to the differences between their results.

These three methods focus on different study objects. Network analysis focuses on the different levels of the road network: node level, lixel level and community level, which represent intersections, road segments, and a small group of roads. Risk analysis focuses on the traffic intersections which are artificially defined. Rule analysis focuses on all accident locations along the road network, and obtains all the road segments close to the accident-associated geographic features.

Moreover, these three methods obtain different actual or predictive results. Network analysis obtains historical accident hotspots, while risk analysis and rule analysis obtain predicted potential accident and near-accident hotspots.

In this way, the results of three different methods have different characteristics, they can not be directly compared. They each have advantages and disadvantages, and the choice depends on the specific situation. The discussion will continue in 6.2.

5.5 Limitations and Where to Continue

Based on the current work, a lot of reflections are made. This thesis has some limitations and can be continued with the following ideas, and these ideas can provide references for future work.

(1) Current network analysis is still limited, go deeper with designing customized centralities and algorithms.

The network is an efficient way of modelling the real road network. The elements in the road network, including node, edge and community, all have practical meanings. The

analysis of the network can help to detect the target elements, like the high-risk node or community studied in this thesis.

In the network analysis part of the case study, the key work is to determine useful centrality measures with practical meaning to detect high-risk nodes and efficient community detection algorithms to detect high-risk communities. Although the chosen ready-made centrality measures and community detection algorithms can work, the results are not ideal enough. The current chosen centrality measures are either too global or too local, they are too much influenced by the road network topology and do not focus on the accident risk. The current chosen community detection algorithm also places too much focus on road topology rather than accident risk.

Therefore, the network analysis can be continued with designing customized centrality measures and community detection algorithms to get more suitable for the current task: to detect high-risk nodes and communities. An ideal centrality measure for this task should consider both the local and global conditions and focus on the accident risk rather than road network topology. An ideal community detection method for this task should focus on detecting high-risk communities and reduce the impact of road network topology.

What's more, other attributes of the road network, such as road class. can be introduced and used for network analysis. Different road classes represent the different importance of roads. The introduction of road class can help to reveal the key roads and nodes.

(2) Limited by data, the rule analysis is incomplete. Combine risk analysis and rule analysis, and apply new data to them.

As introduced in 3.5.2, association rule analysis can not be fully employed in the case study, because the data used in the case study, namely accident locations, only has negative values (there is an accident), and positive values (there is no accident) are unable to be defined and sampled (there is no such data like a no-accident location record).

But if the rule analysis is continued with risk analysis, and the study object is transferred to traffic intersections, then association rule analysis can be fully employed. Because there will be both positive values (low-risk) and negative values (high-risk) in the data. Therefore, not only the support indicators ("if traffic accident risk high/low, then geographic feature X") can be calculated, but confidence indicators ("if geographic feature X, then traffic accident risk high/low") can be calculated.

In this way, the relationship between traffic accidents and geographic features can be better investigated. Risk analysis and rule analysis can be compared with each other and form a more complete study.

(3) Analysis can be iterated based on results, which have shown the potential of crossings.

Both 4.5 Risk Analysis and 4.6 Rule Analysis show that crossings have a strong association relationship with accident risks, especially compared with traffic signals. In this thesis, as introduced in 4.3.2, traffic intersections are identified with traffic signal data, while they can also be identified with crossing data.

Therefore, risk analysis can be done with the identified traffic intersections, and as introduced in the last point, rule analysis can also be done with the identified traffic intersections. Although limited by time, it is not carried out in this thesis, it's still a very meaningful iterative work to continue.

6 Conclusions

6.1 Summary

This thesis is an exploratory study. It aims to explore the location strategy of traffic sensors in the application scenario of traffic safety. Specifically, this thesis aims to explore the possible solutions for detecting or predicting traffic accident and near-accident hotspots on the road network, thereby dealing with the traffic sensor location problem from the perspective of traffic safety.

This thesis proposed three possible solutions in its methodology, which are accident hotspot identification on road networks, risk prediction of traffic intersections, and association rule analysis of accident locations and geographic features. These three possible solutions all use open data sources, they are based on different analysis methods, aiming at different study objects, and acquiring different types of results. Therefore, this thesis is more about the extensive exploration of possible solutions, rather than going very deep into a certain one.

With the case study, the effectiveness of the methodology as well as three possible solutions proposed in this thesis is initially verified. Network analysis methods, machine learning and data mining methods are beneficial for solving TSLP from the perspective of traffic safety. Network analysis methods can effectively identify the historical accident hotspots on the road network. The relationship between accident risk and geographic features can be effectively modelled and used for risk prediction, and association rules between accident risk and geographic features can be effectively extracted. With their results, potential accident and near-accident hotspots can be predicted. What's more, a general guideline on how to deal with TSLP from the perspective of traffic safety is proposed.

This thesis is just an initial exploration. Many aspects of the work can be continued, either the methods being used or the problem being discussed. Therefore, an outlook of the possible future work is concluded.

6.2 General Guideline

Hereby, a general and initial guideline on dealing with TSLP from the perspective of traffic safety is proposed. Under which conditions, a certain method may be more suitable is concluded. The advantages and disadvantages of each method are also concluded.

(1) Network analysis aims at the road network by multiple levels and directly acquires the historical accident hotspots on the road network. This method can directly and clearly acquire the accident hotspots and can be used for multiple types of locations on the road network. However, it requires multi-year historical accident data in the study area, otherwise, the results can not be obtained or are not accurate.

In network analysis, those three different levels represent different practical meanings. The node represents the intersection on the road network, the pixel represents each tiny road segment, and the community represents a connected part of the road network. Therefore, when carrying out practical sensor location selection tasks, the analysis level can be decided based on the desired installation locations. If an intersection is to be

covered, then choose node level. If any place on the road network can be selected, then choose the lixel level. And if a high-risk community is to be covered, then choose the community level.

(2) Risk analysis aims at the traffic intersections and predicts the potential accident and near-accident hotspots. This method can be used in a study area without historical accident data if a model representing the relationship between accident risk and nearby geographic features has been trained in other similar areas. However, this method can only reflect the accident risk associated with geographic features, the accident risk caused by other factors can not be predicted.

(3) Rule analysis aims at the accident locations and finds out the accident-associated geographic features. This method can also be used in a study area without historical accident data, as long as the association rules in other similar areas have been obtained.

All in all, these three analysis methods aim at different study objects on the road network, some obtain historical hotspots and others predict potential hotspots. The selection of an exact analysis method should consider the actual project needs (like only choosing from traffic intersections) and data situation.

6.3 Outlook

There is still a long way to go before transportation becomes smart, safe and sustainable enough in one day. To achieve that, the process of applying more advanced traffic sensors in more scenarios will not stop. Therefore, the study and discussion about the location strategy of traffic sensors, that is Traffic Sensor Location Problem, have to continue, extend, and adapt to the new scenarios in future work.

In the scenario of traffic safety, traffic sensors have many prospects in detecting, alarming, recording and understanding traffic accidents to make transportation safer and more efficient. But "where to put them" and "how many in need" always need to be answered to achieve the key balance between cost and effectiveness. At the current pilot stage, the initial exploration of this problem has great pioneering significance. It will help to reveal the application prospect of advanced traffic sensors in improving traffic safety and help the implementation of related research and applications.

The future work of solving TSLP from the perspective of traffic safety can continue in many ways. More effective network analysis methods including accident centrality measures and community detection methods can be developed to analyze the accident hotspots on the road network, and more attributes of the road network can be used in the process. The association relationship between accident risk and geographic features can be further and completely analyzed, and it can be implemented on a larger geographic scale. With this relationship, the accident hotspots can be better predicted. More work should be carried out to explore other methods used for solving TSLP from the perspective of traffic safety, like optimization methods and heuristic methods.

Bibliography

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 207–216. https://doi.org/10.1145/170035.170072

Agrawal, R., Srikant, R., Road, H., & Jose, S. (1994). Fast Algorithms for Mining Association Rules.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Bonacich, P. (1987). Power and Centrality: A Family of Measures. *American Journal of Sociology*, *92*(5), 1170–1182. https://doi.org/10.1086/228631

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, *27*(1), 55–71. https://doi.org/10.1016/j.socnet.2004.11.008

Brandes, U. (2001). A faster algorithm for betweenness centrality. *The Journal of Mathematical Sociology*, *25*(2), 163–177. https://doi.org/10.1080/0022250X.2001.9990249

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Y.-C. (2017). A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, *1*(1), 161–187. https://doi.org/10.1080/24709360.2017.1396742

Council, N. R. (2005, December 15). *Network Science*. National Academies Press. https://doi.org/10.17226/11516

El Naqa, I., & Murphy, M. J. (2015). What Is Machine Learning? In I. El Naqa, R. Li, & M. J. Murphy (Eds.), *Machine Learning in Radiation Oncology: Theory and Applications* (pp. 3–11). Springer International Publishing. https://doi.org/10.1007/978-3-319-18305-3_1

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

Gelb, J. (2021). spNetwork: A Package for Network Kernel Density Estimation. *The R Journal*, *13*(2), 460. https://doi.org/10.32614/RJ-2021-102

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821–7826. https://doi.org/10.1073/pnas.122653799

Goldhammer, M., Strigel, E., Meissner, D., Brunsmann, U., Doll, K., & Dietmayer, K. (2012). Cooperative multi sensor network for traffic safety applications at intersections. *2012 15th International IEEE Conference on Intelligent Transportation Systems*, 1178–1183. https://doi.org/10.1109/ITSC.2012.6338672

Guerrero-Ibáñez, J., Zeadally, S., & Contreras-Castillo, J. (2018). Sensor Technologies for Intelligent Transportation Systems. *Sensors*, *18*(4), 1212. https://doi.org/10.3390/s18041212

Hastie, T., Friedman, J., & Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer. https://doi.org/10.1007/978-0-387-21606-5

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282 vol.1. https://doi.org/10.1109/ICDAR.1995.598994

Huang, X., He, P., Rangarajan, A., & Ranka, S. (2020). Intelligent Intersection: Two-stream Convolutional Networks for Real-time Near-accident Detection in Traffic Video. *ACM Transactions on Spatial Algorithms and Systems*, *6*(2), 1–28. https://doi.org/10.1145/3373647

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Liu, Z., Luo, P., Murphy, C. E., & Meng, L. (2023). The Potential of VGI and Traffic Flow Generation Model for Solving the Traffic Sensor Location Problem. *Abstracts of the ICA*, *6*, 1–2. https://doi.org/10.5194/ica-abs-6-145-2023

Lv, B., Xu, H., Wu, J., Tian, Y., Zhang, Y., Zheng, Y., Yuan, C., & Tian, S. (2019). LiDAR-Enhanced Connected Infrastructures Sensing and Broadcasting High-Resolution Traffic Information Serving Smart Cities. *IEEE Access*, *7*, 79895–79907. https://doi.org/10.1109/ACCESS.2019.2923421

Maimon, O. Z., & Rokach, L. (2014, September 3). *Data Mining With Decision Trees: Theory And Applications (2nd Edition)*. World Scientific.

Opitz, D., & Maclin, R. (1999). Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research*, *11*, 169–198. https://doi.org/10.1613/jair.614

Owais, M. (2022). Traffic sensor location problem: Three decades of research. *Expert Systems with Applications*, *208*, 118134. https://doi.org/10.1016/j.eswa.2022.118134

Rikalovic, A., Cosic, I., & Lazarevic, D. (2014). GIS Based Multi-criteria Analysis for Industrial Site Selection. *Procedia Engineering*, *69*, 1054–1063. https://doi.org/10.1016/j.proeng.2014.03.090

Tewolde, G. S. (2012). Sensor and network technology for intelligent transportation systems. *2012 IEEE International Conference on Electro/Information Technology*, 1–7. https://doi.org/10.1109/EIT.2012.6220735

WHO. (2018). *Global status report on road safety 2018*. World Health Organization.

Wikipedia. (2023, August 20). Wuppertal. In *Wikipedia*. Retrieved September 6, 2023, from https://en.wikipedia.org/w/index.php?title=Wuppertal&oldid=1171328584 Page Version ID: 1171328584

Wu, J., Xu, H., Zheng, Y., & Tian, Z. (2018). A novel method of vehicle-pedestrian near-crash identification with roadside LiDAR data. *Accident Analysis & Prevention*, *121*, 238–249. https://doi.org/10.1016/j.aap.2018.09.001

Zhang, J., Xiao, W., Coifman, B., & Mills, J. P. (2020). Vehicle Tracking and Speed Estimation From Roadside Lidar. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 5597–5608. https://doi.org/10.1109/JSTARS.2020.3024921

Zhang, Z., Zheng, J., Xu, H., & Wang, X. (2019). Vehicle Detection and Tracking in Complex Traffic Circumstances with Roadside LiDAR. *Transportation Research Record*, *2673*(9), 62–71. https://doi.org/10.1177/0361198119844457

Zhao, H., Sha, J., Zhao, Y., Xi, J., Cui, J., Zha, H., & Shibasaki, R. (2012). Detection and Tracking of Moving Objects at Intersections Using a Network of Laser Scanners. *IEEE Transactions on Intelligent Transportation Systems*, *13*(2), 655–670. https://doi.org/10.1109/TITS.2011.2175218

Zhao, J., Xu, H., Liu, H., Wu, J., Zheng, Y., & Wu, D. (2019). Detection and tracking of pedestrians and vehicles using roadside LiDAR sensors. *Transportation Research Part C: Emerging Technologies*, *100*, 68–87. https://doi.org/10.1016/j.trc.2019.01.007

Zhou, Z.-H. (2012, June 6). *Ensemble Methods: Foundations and Algorithms*. CRC Press.